UNIVERSITY OF CALIFORNIA

Los Angeles

Relating Optical Speech to Speech Acoustics and Visual Speech Perception

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Electrical Engineering

by

Jintao Jiang

2003

The dissertation of Jintao Jiang is approved.

Kung Yao

Lieven Vandenberghe

Patricia A. Keating

Lynne E. Bernstein

Abeer Alwan, Committee Chair

University of California, Los Angeles

2003

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGEMENTS

First and foremost, I am deeply indebted to my advisor, Prof. Abeer Alwan, for giving the opportunity to work with her. Without her guidance, encouragement, support, or patience, I would not have been able to complete this dissertation. I have also learned a lot from her technical and editorial advice and insights.

I am very grateful to Dr. Lynne E. Bernstein, Prof. Patricia A. Keating, and Dr. Edward T. Auer for their continuous help and insightful suggestions during the course of the project. I would also like to thank Brian Chaney, Sumiko Takayanagi, Jennifer Yarbrough, Patrick Barjam, Sven Mattys, Taehong Cho, Marco Baroni, and Paula E. Tucker.

My thanks also go to other members of my committee, previous and current, Profs. Kung Yao, Lieven Vandenberghe, and Sun-Ah Jun for being on my committee and taking time to review my dissertation and offer valuable comments and suggestions.

Students at UCLA Speech Processing and Auditory Perception Laboratory have given many suggestions to my work, and I have had fruitful and interesting discussions with them. I thank Katherine Rogowski for countless help that she has given me.

I would like to acknowledge the contributions of Abeer Alwan, Lynne E. Bernstein, Patricia A. Keating, and Edward T. Auer to Chapters 1-5.

I would also like to acknowledge the contributions of Abeer Alwan, Harriet Nock, Giridharan Iyengar, Chalapathy Neti, and Gerasimos Potamianos to Chapter 6.

This work was supported in part by NSF (KDI-9996088). Any opinions or conclusions expressed in this dissertation do not necessarily reflect the views of NSF, and

no official endorsement should be inferred.

Last but not least, I would like to thank my wife Xiaojin and my parents for their patience and support.

VITA

November 21, 1972	Born, Hubei Province, P.R.China
1995	B.S., Electronic Engineering Tsinghua University Beijing, P.R.China
1995-1998	Graduate Student Researcher Tsinghua University Beijing, P.R.China
1998	M.S., Electronic Engineering Tsinghua University Beijing, P.R.China
1998-2003	Graduate Student Researcher University of California, Los Angeles Los Angeles, California
2000-2001	Teaching Assistant Department of Electrical Engineering University of California, Los Angeles
2002	Summer Intern Bell Labs, Lucent Technologies Murray Hill, New Jersey
2003	Summer Intern IBM T.J. Watson Research Center Yorktown Heights, New York

PUBLICATIONS AND PRESENTATIONS

Bernstein, L.E., Jiang, J., Alwan, A., and Auer, E.T. (2001). "Visual phonetic perception and optical phonetics," *Proc. AVSP*, Scheelsminde, Denmark, 104-109.

Jiang, J., Alwan, A., Auer, E.T., and Bernstein, L.E. (2001). "Predicting visual consonant perception from physical measures," *Proc. EUROSPEECH*, Aalborg, Denmark, 179-182.

Jiang, J., Alwan, A., Bernstein, L.E., Auer, E.T., and Keating, P.A. (2002). “Similarity structure in perceptual and physical measures for visual consonants across talkers,” *Proc. ICASSP*, Orlando, Florida, 441-444.

Jiang, J., Alwan, A., Bernstein, L.E., Auer, E.T., and Keating, P.A. (2002). “Predicting face movements from speech acoustics using spectral dynamics,” *Proc. Int’l Conf. on Multimedia and Expo (ICME)*, Lausanne, Switzerland, 181-184.

Jiang, J., Alwan, A., Bernstein, L.E., Keating, P.A., and Auer, E.T. (2000). “On the Correlation between facial movements, tongue movements, and speech acoustics.” *Proc. ICSLP*, Beijing, China, 1: 42-45.

Jiang, J., Alwan, A., Keating, P.A., Auer, E.T., and Bernstein, L.E. (2002). “On the relationship between face movements, tongue movements, and speech acoustics,” *EURASIP J. on Applied Signal Proc. on Joint Audio-Visual Speech Proc.*, Nov. 2002, 1174-1188.

Jiang, J., Alwan, A., Keating, P.A., and Bernstein, L.E. (2000). “On the correlation between orofacial movements, tongue movements, and speech acoustics,” *J. Acoust. Soc. Am.*, 107(5): 2904.

Jiang, J., Alwan, A., Keating, P.A., Bernstein, L.E., Auer, E.T. (2000). “On the correlation between articulatory and acoustic data,” *J. Acoust. Soc. Am.*, 108(5): 2508.

ABSTRACT OF THE DISSERTATION

Relating Optical Speech to Speech Acoustics and Visual Speech Perception

by

Jintao Jiang

Doctor of Philosophy in Electrical Engineering
University of California, Los Angeles, 2003
Professor Abeer Alwan, Chair

Since relatively few studies have examined optical phonetics and the relationship between speech acoustics and optical movements, visual speech synthesis and audio-visual speech recognition is currently based on limited optical phonetic knowledge. This dissertation aims at quantifying the relationship between speech acoustics and optical movements and finding physical cues to visual speech perception. A database of 69 consonant-vowel (CV) syllables and three sentences (each repeated four times) spoken by two males and two females was analyzed. A three-dimensional optical motion capture system and an electromagnetic midsaggital articulography system were used to capture face and tongue movements, respectively. The relationships between the three data streams were examined using multilinear regression. Results showed that the relationship between speech acoustics and optical movements was not uniform across talkers, vowel context, place of articulation, and individual optical, articulatory, or acoustic channels. For CV syllables, the correlations between face movements and tongue movements were high, (r = 0.70 - 0.88) and articulatory data could be well predicted from speech acoustics (r = 0.74 - 0.82). Based on the non-uniformity of the relationships for CV syllables, a

dynamical model was proposed to enhance the relationship between speech acoustics and optical movements, and an improvement of about 17% was reported. These correlations were further confirmed using mutual-information-face analysis.

To examine the relationship between visual consonant perception and optical measures, another database of 69 CV syllables (repeated twice) was recorded (without tongue movements). Each talker's syllable productions were presented for identification in a visual-only condition to 6 hearing participants. Physical and perceptual measures were analyzed with multilinear regression, multidimensional scaling, and phoneme equivalence class analyses. Results showed that physical measures accounted for about 66% of the variance of visual consonant perception. Among the different facial regions, the lip area (55%) was the most informative, although the cheeks (36%) and the chin (32%) also contribute significantly to visual perceptual intelligibility. The correlations were not uniform across vowel context and talkers. Implications for visual speech synthesis are discussed.

CHAPTER 1. INTRODUCTION

1.1. Audio-Visual Speech Processing

Research in audio-visual speech perception and recognition has confirmed the multi-modality of human speech perception (Binnie et al., 1974; Cox et al., 1997; Erber, 1975; Luettin and Dupont, 1998; MacLeod and Summerfield, 1987; Potamianos et al., 2001; Smith, 1989; Sumby and Pollack, 1954): In face-to-face communication, perceivers can obtain and process information from the faces and voices of talkers simultaneously. Visual speech can also generate illusions such as in the McGurk effect (McGurk and MacDonald, 1976): When an acoustic /ba/ and a video tape of a person saying /ga/ are presented simultaneously, viewers frequently report perception of /da/. Face movements are basically by-products of speech production through vocal tract activities. Visual speech is much less intelligible than speech acoustics; hence, its role in speech perception is complementary (Green and Kuhl, 1991), while the auditory modality is dominant. However, information exploited by perceivers from faces becomes more important in the case of background acoustic noise, non-native listeners or speakers, and complex or unfamiliar content. Furthermore, visual information can generate user-friendly interfaces and attract more attention because humans can perceive more emotion and expression from visual speech than from audio (Henton and Litwinowics, 1994). Nevertheless, presenting natural faces requires large bandwidth and data storage, and this method lacks flexibility in generating new speech materials. These limitations of natural faces have

prompted researchers to develop realistic synthetic talking faces. The effort to create talking machines began several hundred years ago (Lindsay, 1997), and over the years most speech synthesis efforts have focused mainly on speech acoustics. The desire to create talking faces along with synthetic voices has been inspired by many potential applications.

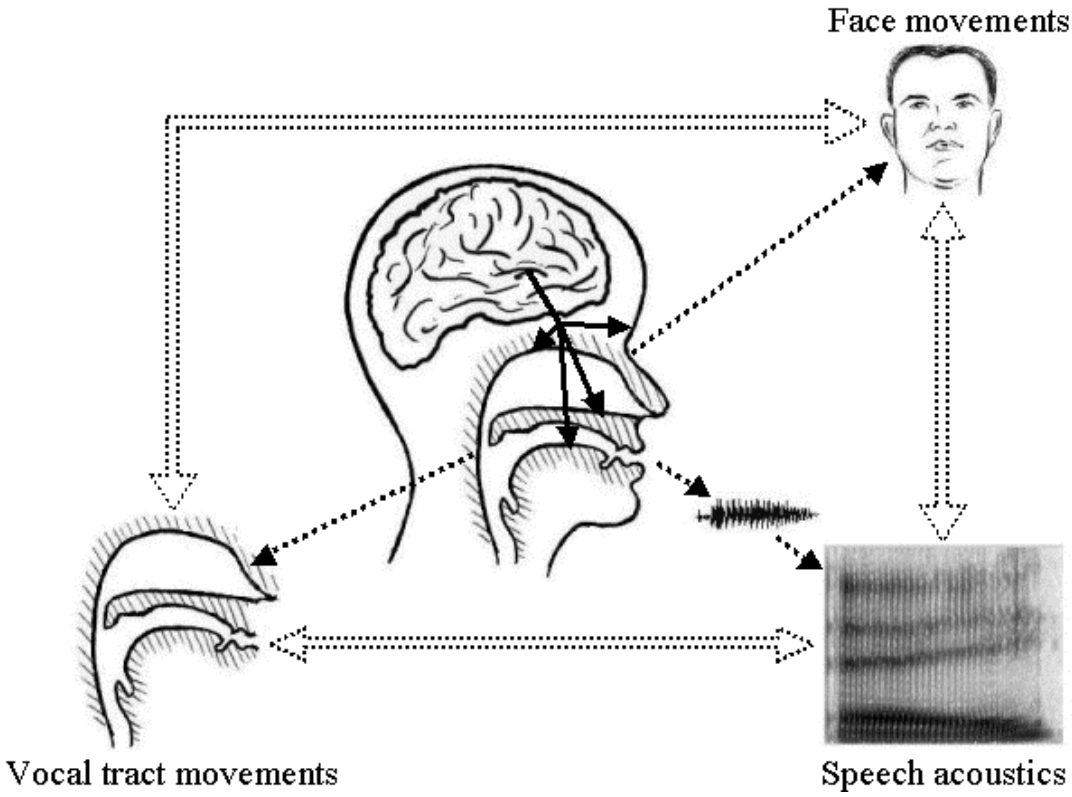


Figure 1.1 The relationships between face movements, vocal tract movements, and speech acoustics.

Many years of research on acoustic phonetics, recognition, and synthesis have led to tremendous achievements. However, relatively few studies have quantitatively examined visual speech signals. Speech production involves control of various speech articulators to produce acoustic speech signals. Under the central control of the human

brain, the various speech articulators (tongue, jaw, velum, larynx, and lips) work together to produce speech movements, and thus acoustic speech sounds result from those movements. Predictable relationships between speech movements and external face movements and thus also between external face movements and speech acoustics are expected (Figure 1.1). Towards this end, this dissertation aims at quantifying optical signals and examining their relationships with acoustic signals. A better understanding of the relationships between speech acoustics and optical (face and tongue) movements would be helpful to develop better synthetic talking faces and for other applications as well (Rubin and Vatikiotis-Bateson, 1998). For example, in audio-visual speech recognition, optical (facial) information (both the correlated and un-correlated information with speech acoustics) could be used to compensate for noisy speech waveforms (Luettin and Dupont, 1998; Potamianos et al., 2001); the correlations between optical information and speech acoustics could also be used to enhance auditory comprehension of speech in noisy situations (Girin et al., 2001).

More than five decades of research on auditory perception has yielded a large body of information on acoustic phonetics and phonology. Acoustic phonetics has benefited from numerous careful measurements and perceptual experiments on natural or synthetic acoustic speech. The relationship between acoustic physical measures and auditory speech perception has been very helpful in developing acoustic phonetics and thus has been used in speech communication, synthesis, and recognition systems. Similarly, examining the relationship between optical measures and visual speech perception is an indispensable step towards creating knowledge about optical phonetics.

Only a few papers, however, have examined optical (facial) phonetics, and thus optical speech synthesis is currently based on limited optical phonetic knowledge. For example, what makes a talker visually intelligible is not well quantified, and how physical gestures and acoustic properties are related remains unknown. Visual speech perception, which is a viewer's subjective judgment of optical stimuli, should have a basis in physical stimulus characteristics. Finding physical cues to visual speech perception can help us understand and model visual speech perception and ultimately improve visual speech synthesis (Benoît et al., 1996). Towards this end, this dissertation also aims at examining the relationship between visual speech perception and optical measures.

1.2. How to Examine the Relationship between Data Sets

The objectives of this dissertation are to examine the relationship between articulatory (face and tongue) movements and speech acoustics and the relationship between visual perceptual data and optical measures. If data $\boldsymbol{Y}$ (e.g., face movements) can be predicted from data $\boldsymbol{X}$ (e.g., speech acoustics), the relationship can be defined as $\boldsymbol{Y} = f(\boldsymbol{X})$, where f is an unknown function. Function f can indicate the strength of the relationship: strong for one-to-one mapping, moderate for many-to-one mapping, and weak for one-to-many mapping. Function f can also indicate which components are important for predicting $\boldsymbol{Y}$ from $\boldsymbol{X}$. Unfortunately, little is known about the function f, and examining the relationship directly based on function f is only beginning to be studied using methods from cognitive neuroscience. A completely behavioral method is to design a mapping

from ***X*** to ***Y*** and then examine the correlation between ***Y*** and its prediction from ***X***. Montgomery and Jackson (1983) used linear regression for the mapping between visual speech perception and physical measures. Studies that have examined the relationship between face movements and speech acoustics have used linear regression as well as neural networks (Barker and Berthommier, 1999; Massaro et al., 1999; Yehia et al., 1998; Yehia et al., 1999). In these studies, neural networks have been shown to account for more of the variance in the mapping. However in our view, multilinear regression provides a better opportunity to probe the underlying mapping function. Therefore, in this dissertation multilinear regression is used for exploring the relationship between ***Y*** and ***X*** (Sen and Srivastava, 1990). As a result, the function f is approximated by multilinear regression, which is a one-to-one mapping. To assess the strength of the relationships between original data (***Y***) and its prediction from ***X***, Pearson correlations are computed. With a linear mapping, if a strong relationship exists, then this strong relationship can help build a parsimonious explanation of how ***Y*** and ***X*** are related.

1.3. The Relationship between Articulatory Movements and Speech Acoustics

Many researchers have attempted to determine the relationship between speech acoustics and vocal tract shapes (e.g., Alwan et al., 1997; Atal et al., 1978; Badin et al., 1995; Bangayan et al., 1996; Fant, 1960; Flanagan, 1965; Ladefoged et al., 1978; Lieberman and Blumstein, 1988; Narayanan et al., 1997; Narayanan and Alwan, 2000; Schroeter and

Sondhi, 1992, 1994; Sondhi and Schroeter, 1987; Stevens and House, 1955; Stevens, 1972). The inverse problem (predicting vocal tract shapes from speech acoustics) has a long history of difficulty, and neural network methods have been used to quantitatively examine the relationship between speech acoustics and vocal tract shapes (Atal and Rioul, 1989; Rahim and Goodyear, 1990; Shirai and Kobayashi, 1991). Levinson and Schmidt (1983) used adaptive computation for the mapping. The codebook method has also been explored (Dusan and Deng, 1998; Ouni and Laprie, 2000).

Although considerable research has been conducted on the relationship between speech acoustics and vocal tract shapes, direct examination of the relationship between speech acoustics and face movements has only recently been undertaken (Agelfors et al., 1999; Barker and Berthommier, 1999; Massaro et al., 1999; Mori and Sonoda, 1998; Vignoli, 2000; Yehia et al., 1998). Yehia et al. (1998) used linear regression to examine the relationships between tongue movements, external face movements (lips, jaw, cheeks), and speech acoustics for two English sentences. The authors reported that face movements could be well predicted from tongue movements (with 83% of the variance accounted for). Furthermore, acoustic line spectral pairs (LSPs; Sugamura and Itakura, 1986) were moderately predicted from face movements and tongue movements (with about 50% of the variance accounted for). Barker and Berthommier (1999) examined the relationship between face movements and the LSPs of 54 French nonsense words. Using multilinear regression, the authors reported that face movements predicted from LSPs and RMS energy accounted for 56% of the variance of obtained measurements, while predicted acoustic features from face movements accounted for only 30% of the variance.

These studies have established that lawful relationships can be demonstrated between different speech data streams. However, the previous studies were based on limited data. In (Yehia et al., 1998), only two English sentences spoken by one English talker were studied. Barker and Berthommier (1999) examined the relationship using nonsense French phrases, and no tongue movements were recorded. In order to have confidence about the generalization of the relationships those studies reported, additional research with more varied speech materials and larger databases is needed.

Speech production is a dynamical process, and it is expected that correlations between articulatory movements and speech acoustics change along time and thus context. However, using sentences (Yehia et al., 1998) or phrases (Barker and Berthommier, 1999) as basic analysis units may not be appropriate. Recently, Lee et al. (2001) used a Kalman filtering technique to examine these relationships for phonemes, and the authors reported high correlations (more than 0.95) between predicted (from speech acoustics) and measured vocal tract trajectories. Barbosa and Yehia (2001) reported that linear analysis on segments of duration 0.5 second could yield high correlations. Therefore, it is reasonable and desirable to use consonant-vowel (CV) syllables (or even phonemes) for the mapping. Towards this end, CV nonsense syllables were studied in this dissertation. A database of sentences was also recorded, and correlation results were compared with CV syllables. In addition, analyses were also performed to examine possible effects associated with a talker' gender and visual intelligibility.

The relationships between face movements, tongue movements, and speech

acoustics are most likely globally nonlinear. Indeed, nonlinear techniques (neural networks, codebooks, and Hidden Markov Models) have been used in other studies (Agelfors et al., 1999; Barker and Berthommier, 1999; Massaro et al., 1999; Yamamoto et al., 1998; Yehia et al., 1999). Yehia et al. (1998) also stated that various aspects of the speech production system are not related in a strictly linear fashion, and nonlinear methods may yield better results. However, from an implementation viewpoint, linear correlation is easy to implement and mathematically tractable. Therefore, in (Barbosa and Yehia, 2001; Lee et al., 2001), linear prediction (or Kalman filtering) was applied to small speech segments. In other words, these local linear functions can be used to approximate global nonlinear functions. Therefore, the relationships (between face movements, tongue movements, and speech acoustics) for CV syllables, which span a short time interval (locally), can be approximated by linear functions. A popular linear mapping technique for this purpose is multilinear regression (Sen and Srivastava, 1990). For sentences, nonlinear techniques should probably be applied.

In this dissertation, a dynamical model was also explored to enhance the relationship between face movements and speech acoustics. In general, the computational expense increases exponentially with the number of features used. Barker and Berthommier (1999) added the 1^{st}-order derivative to static LSPs in the estimation process. A small enhancement (from 0.36 to 0.37) was found, because the 1^{st}-order derivative captures only short-term correlations in the signal. Lee et al. (2001) applied dynamical information by dividing sentences into segments. The new dynamical model proposed in this dissertation is, however, a causal and a non-causal filter based on the

autocorrelations of the acoustics and those of the face movements.

1.4. The Relationship between Visual Speech Perception and Physical Measures

Potential applications of optical speech are now widely acknowledged, but there are several issues that need to be addressed. The first issue is the lack of sufficient optical measurements (Munhall and Vatikiotis-Bateson, 1998). Another issue is the scarcity of studies that attempt to correlate physical (optical) features and visual speech perception. Work related to this question has been carried out in the auditory phonetic perception literature, where stimuli have been synthesized using parametric changes in various acoustic control parameters, and perceivers have been tested to determine the relationship between the physical parameters and their perceptual structure (Benkí, 2001; Cooper et al., 1952; Liberman, 1957; Lisker, 1975; Lisker and Abramson, 1964, 1970; Nearey, 1990, 1992, 1997). For example, acoustic speech synthesizers based on parameters (e.g., formants) were built for perceptual experiments, and the corresponding perceptual results were analyzed against physical parameters (Cooper et al., 1952; Liberman, 1957). This method (analysis-by-synthesis; Liberman and Mattingly, 1985) is limited, however, because the stimulus space is designed to investigate a very restricted number of segments: The method does not scale to the level of the general similarity structure across all segments or all words in a language. A method is needed to capture physical stimulus similarity across a wide range of stimuli and relate it to perception. In the study of

auditory phonetic perception, a common method for obtaining knowledge about phonetic perception has been essentially descriptive and has three main components: (a) presenting speech stimuli to perceivers for their responses, (b) obtaining physical measures (formants, voice onset time, etc.) from these stimuli, and (c) examining the relationship between these perceptual data and physical measures (Chen and Alwan, 2000; Chen and Alwan, 2001; Soli and Arabie, 1979).

Like its counterpart in acoustics, optical phonetics can be explored using optical speech synthesis. Several researchers have performed visual speech perception on synthetic visual speech (Benoît et al., 1996; Le Goff et al., 1994; Massaro, 1998; Walden et al., 1987). Le Goff et al. (1994) and Massaro (1998) reported significant gain in intelligibility by adding a synthetic face (or lips) to an acoustic speech synthesizer. Benoît et al. (1996) conducted audio-visual perceptual experiments and showed that speech identification accuracy increased, as more of the face (lip, jaw, and cheek) was made visible. This result is predictable from the fact that speech articulators move somewhat independently to convey phonemic distinctions. For example, cheek movements often show clear evidence of deformation of the back cavity in the vocal tract. However, Benoît et al. (1996) and Le Goff et al. (1994) did not quantitatively examine the relationship between audio-visual perception and physical measures, and the study was general in nature. Several other studies focused on generating more controllable, realistic, and natural face movements, but they did not perform perceptual experiments (Parke, 1974; Waters, 1987). Similarly, Massaro (1998) and Walden et al. (1987) did not make any physical measurements and thus did not examine the relationship.

Using optical speech synthesizers to study optical phonetics would limit the results due to the assumptions made in optical speech synthesis, because optical phonetics is still in its early stage. Therefore, the descriptive method is more appropriate for the current study. Several researchers have examined the visual perception of CV syllables, vowel-consonant-vowel (VCV) disyllables, or (nonsense) words (Bernstein et al., 2000b; Franks, 1972; Iverson et al., 1998; Kricos and Lesner, 1982; Owens and Blazek, 1985). Owens and Blazek (1985) examined differences in visemes for hearing-impaired and normal-hearing adult viewers using VCV disyllables, while Bernstein et al. (2000b) used CV syllables. Note that groups of phonemes that are visually indistinguishable, at some degree of accuracy, are called "visemes" (Benoît et al., 1992; Fisher, 1968). Auer, Bernstein, and their colleagues (Auer and Bernstein, 1997; Iverson et al., 1998) further computationally analyzed the effect of obtained visemes on lexical structure. Kricos and Lesner (1982) examined differences in visual intelligibility across talkers. These studies, however, fall short of exploring underlying physical (facial) characteristics. The only complete example in this direction is a study by Montgomery and Jackson (1983). Their study (four female talkers, 10 participants, and 15 vowels in /hVg/ nonsense words) showed that the physical measures were moderately successful as predictors of vowel perception (approximately 50% of the variance was accounted for), but the predictions were talker dependent. These predictions illustrated that visual vowel perception was cued by some physical measures. However, their study used only a few physical measures (lip height, lip width, lip aperture, acoustic duration, and visual duration), and there was no measurement in the cheek or chin areas. Furthermore, a

similar experiment was not done for consonants, and little is known about the relationship between visual consonant perception and corresponding face movements.

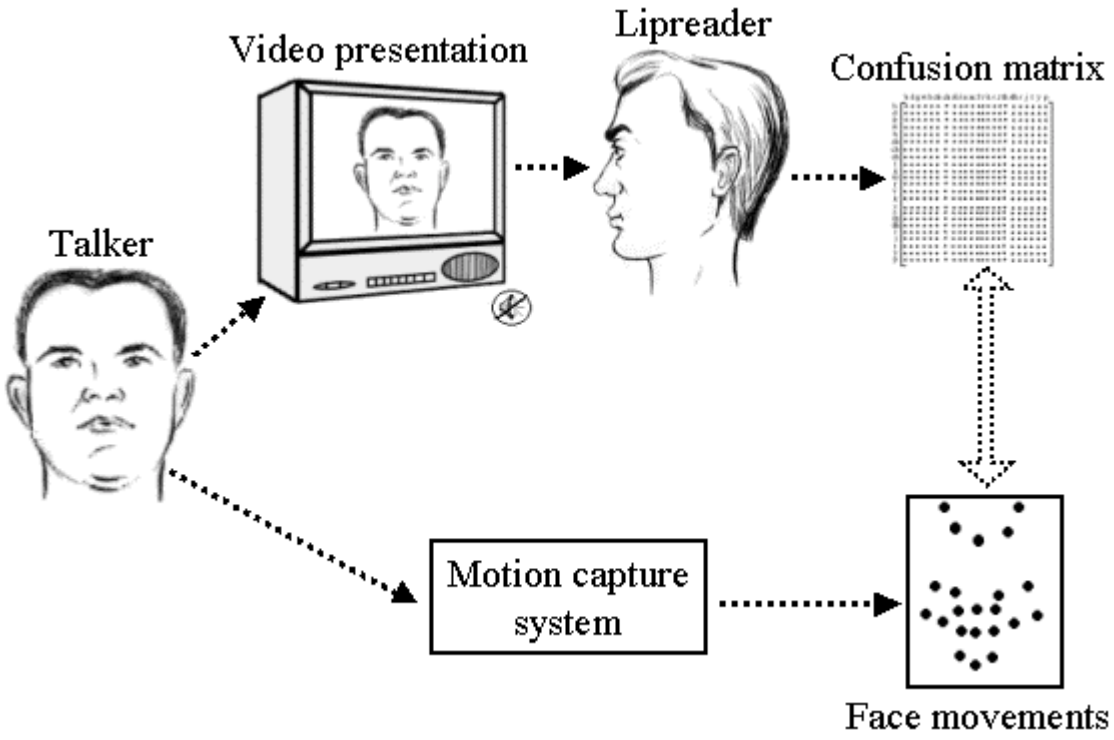


Figure 1.2 Finding optical cues to visual speech perception.

In this dissertation, a descriptive method was used to examine the relationship between visual speech perception of spoken nonsense syllables (with a number of consonants and vowels) and the physical attributes of those stimuli (see Figure 1.2). This work extended Montgomery and Jackson's (1983) study in two directions: It focused on consonants and used more physical measures. Unlike vowels, physical attributes of consonants are difficult to describe, because the fast transition between the consonant and the vowel involves significant dynamical characteristics of the face movements. To date, most measurements on talkers' faces have applied to the lip borders. For example, studies

on computer audio-visual speech recognition have focused the region of interest around the lips (Chan et al., 1998; Luettin et al., 1996; Potamianos et al., 2001). Although the lip area is the most informative, lipreading also depends on the other parts of the face (Benoît et al., 1996). This is especially true for back phonemes, where cheek movements are prominent. Therefore, a thorough examination of face movements should include the whole face. The need for investigating areas other than the lips does, however, impose problems for image processing techniques, as such movements are more difficult to detect automatically. For the current study, an optical motion capture system was used to record accurately the dynamical positions of markers on talkers' faces (including chin and cheeks).

Visual consonant perception was examined in CV syllables, and only face (including chin and cheeks) movements were recorded (no tongue movements). The similarity structure of the visual perceptual data was defined by analyzing the forced-choice perceptual identifications. The consonant identification data were tallied in stimulus-response confusion matrices that were transformed into dissimilarity matrices using the phi-square transformation (Iverson et al., 1998). The similarity structure of the face measurement data was defined by analyzing the 3-D motion capture data for each consonant segment. Multilinear regression was then used to predict perceptual dissimilarity vectors from the physical distance matrices. In addition, multidimensional scaling (MDS; Kruskal and Wish, 1978) was applied to both the perceptual dissimilarity matrices as well as matrices predicted from physical measures, and a phoneme equivalence class (PEC) analysis was used to examine the grouping of consonants.

Using stimulus-response confusion matrices is a common way to get perceptual similarity and thus a visual perceptual space. By hypothesis, humans perceive speech (auditory or visual) by judging the similarity between stimuli and templates in memory. This is yet another issue complicating the understanding of speech perception. An ability to assess similarity lies close to the core of cognition, and geometric and feature models have been among the most influential approaches to analyzing similarity (Blough, 2001; Goldstone, 1999). Geometric models assume three basic metric axioms: minimality, symmetry, and triangular inequality and are exemplified by multidimensional scaling (MDS) models (Shepard, 1962, 1987; Torgerson, 1965). The geometric approach stresses the representation of similarity relationships among the members of a set of objects. However, geometric models' metric assumptions are open to question (Blough, 2001). Tversky (1977) suggested an alternative approach, the contrast model, wherein similarity is determined by matching features of compared entities and by integrating the features. The features are assumed to be independent and can be manipulated via logical or set operations. However, unlike acoustic phonetic features (Chomsky and Halle, 1968; Jakobson et al., 1952; Stevens, 1989; Stevens and Blumstein, 1981; Stevens and Keyser, 1989), visual phonetic features have not been adequately studied, and currently there is no convincing definition of features from the raw physical data. Therefore, a geometric model is used in this dissertation for similarity. The objective is to show that direct (geometric) physical measures suffice to predict visual speech perception, and thus mediating features are not required.

We predicted that the whole face would yield better correspondence between

visual speech perception and physical measures than parts of the face. If lipreading participants did not see visual stimuli but simply guessed, then the correlation should be about zero; if participants only see the lips, then the correlations should be higher; while the whole face would yield the best results. The level of correlations shows how much visual perception is derived from physical measures and consequently shows whether direct physical measures (other than phonological features) can be used to construct a parsimonious explanation for visual speech perception. In addition, these correlations can support future studies that examine which regions on the face are important for visual speech perception. It is desirable to have a uniform relationship between visual speech perception and physical measures across talkers. However, talkers differ in physical conformation and pronunciation style. Therefore, talker differences were also investigated by examining how each part of the face contributes differently to visual speech perception for each talker. For nonsense syllables, a higher perceived intelligibility (or a larger face) may lead to a more detailed examination of face characteristics during lipreading (less guessing) and thus result in higher correlations.

1.5. Outline of This Dissertation

Chapter 2 describes the database used for the studies of the relationship between face movements, tongue movements, and speech acoustics and the relationship between optical speech perception and physical measures. Chapter 3 describes the basic mathematical techniques used in this dissertation. Chapter 4 presents a detailed analysis

of correlations between face movements, tongue movements, and speech acoustics. A dynamical model for the enhancement of the correlation between face movements and speech acoustics is also presented in Chapter 4. The relationship between visual consonant perception and optical measures is explored in Chapter 5. Chapter 6 presents a mutual-information-face algorithm for analyzing the correlation between face movements and speech acoustics. Finally, the implications and future directions are discussed in Chapter 7.

CHAPTER 2. DATA COLLECTION AND PRE-PROCESSING

2.1. Introduction

In this dissertation, two aspects of visual speech were examined. One was the relationship between face movements, tongue movements, and speech acoustics. The other was the similarity structure between visual speech perception and optical (physical) measures. For the correlation analysis, face movements, tongue movements, speech acoustics, and video (for reference only) were recorded simultaneously. For the similarity analysis, face movements, speech acoustics, and video (for presentations) were recorded simultaneously. This chapter describes the acquisition and processing of the databases.

The author of this dissertation participated in most recording sessions and perceptual experiments. To make this dissertation informative and complete, the author uses some information provided by House Ear Institute (HEI) and the UCLA Phonetics Laboratory. Note that several pictures used in this dissertation have been adapted from ones on the two websites:

http://www.hei.org/research/projects/comneur/speechdata.htm and

http://www.linguistics.ucla.edu/faciliti/uclaplab.html.

For example, the pictures about the data recording were adapted from those of the HEI website, and the pictures about the Electromagnetic Articulography (EMA) system and the head profile (including the one used in Chapter 4) were adapted from those of the

UCLA Phonetics Laboratory website.

2.2. Background

This work represents the first attempt to record several speech-related data streams simultaneously. Yehia et al. (1998) recorded face movements and tongue movements in different sessions and then aligned them using Dynamical Time Warping. The limitation of their method is that in different sessions, the recording configurations were different, and speakers cannot speak in the exact same way. However, recording high quality multiple data streams simultaneously was very difficult and time consuming for the studies in this dissertation. First, these recording systems needed to be synchronized, and a customized circuit was designed for this purpose. Second, wearing an EMA helmet with retro-reflectors on the face and EMA pellets on the tongue was stressful for the talkers. Third, a crew was needed to operate different equipment and monitor the process. Fourth, EMA and Qualisys^{TM} systems' effects on acoustic recording needed to be compensated for, and the microphone's location be carefully chosen. Fifth, there was extensive preparation before the recording: putting retro-reflectors on the face, putting pellets on the tongue, setting up the lighting, and calibration of the systems. Sixth, there was extensive work on post-processing of the raw data. The advantage of this simultaneous recording was that all data streams were recorded in the same recording configurations.

2.3. Recording a Database for the Correlation Analysis

2.3.1. Talkers

The talkers (with English as a native language) were selected from a larger pool that had been initially screened for their visual intelligibility. Each talker was video-recorded saying 20 different sentences. Five deaf adults were asked to transcribe the video recording (without audio) in regular English orthography for each talker and assign a subjective intelligibility rating to each sentence. Thus, each talker received 20 intelligibility-rating scores from each lipreader, and a mean was calculated. As a result, two male (M1 and M2) and two female (F1 and F2) talkers with different subjective intelligibility ratings were selected. Among the four talkers, M2 was perceived to be the most intelligible of the talkers, F1 the least intelligible, and M1 and F2 moderately intelligible. The mean sentence intelligibility ratings were 3.6, 8.6, 1.0, and 6.6 for talkers M1, M2, F1, and F2, respectively. These sentence intelligibility ratings are on a scale ranging from 1 (*not intelligible*) to 10 (*very intelligible*). The average percent words correct for the talkers were: 46% for M1, 55% for M2, 14% for F1, and 58% for F2. The correlation between the objective English orthography results and the subjective intelligibility ratings was 0.89. Note that F2 had the highest average percent words correct, but she did not have the highest intelligibility rating.

Subsequently, extensive additional visual-only speech perception testing, with 16 normal-hearing human subjects, of 320 sentences produced by each of these four talkers

showed that F2 was the most intelligible, then M1, followed by M2 and F1. These results were replicated with eight deaf lipreaders, except that M2 was more intelligible than M1 (i.e., F2 > M2 > M1 > F1).

Note that visual intelligibility ratings used in this dissertation refer to the more extensive objective ratings, but not the initial subjective ratings.

2.3.2. Materials

The experimental corpus then obtained with the four selected talkers consisted of 69 consonant-vowel (CV) syllables in which the vowel was one of /a, i, u/, and the consonant was one of the 23 American English consonants /y, w, r, l, m, n, p, t, k, b, d, g, h, θ, ð, s, z, f, v, ʃ, ʒ, tʃ, dʒ/. Each syllable was produced at least four times in a pseudo-randomly ordered list. In addition to the CVs, three sentences were recorded and produced at least four times by each talker:

1. When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow.
2. Sam sat on top of the potato cooker and Tommy cut up a bag of tiny potatoes and popped the beet tips into the pot.
3. We were away a year ago.

Sentences 1 and 2 are the same sentences used by Yehia et al. (1998). Sentence 3 contains only voiced sonorants (and /g/). During recording, the crew monitored each CV or sentence production. If one syllable/sentence was not produced well, then that syllable/sentence and the following syllables/sentences in the same take were repeated.

For example, if the first syllable/sentence in a take needed to be repeated, then the whole take was repeated; if the last syllable/sentence in a take needed to be repeated, then only that syllable/sentence was repeated. Therefore, each syllable/sentence had at least four good repetitions. Note that sentences were recorded after all CV syllables were recorded.

2.3.3. Recording Facilities

The recordings were made at HEI. Researchers at HEI screened talkers and lipreading participants, set up the recording facilities (video, Qualisys^{TM}, acoustic tiles, lighting equipment, teleprompter, etc.), designed synchronization schemes, processed the raw data, set up the perceptual experiments, and managed and oversaw the whole process. Researchers from the UCLA Phonetics Laboratory were responsible for the EMA data recording and processing (they provided the EMA recording system) and the placement of retro-reflectors. Researchers from the UCLA Speech Processing and Auditory Perception Laboratory (SPAPL) were responsible for acoustic settings (they provided acoustic recording equipment), participated in the recording sessions and processing of raw data, and administered the perception experiments.

Four data streams were recorded simultaneously: acoustics, video, tongue motion, and 3-D face point motion. The recording facilities were described in (Bernstein et al., 2000a). The equipment used in the current study is listed in Table 2.1.

Table 2.1 Summary of equipment used in the recordings.

Data streams	Equipment	Description
Audio	DAT recorder Sennheiser microphone	HHB Portadat PDR1000 (recorder) MKH416P48U (microphone)
Video	VTR	Sony UVW-1800 (recorder) Sony DXC-D30 (camera)
Face motion	Qualisys™ system	MCU120/240Hz CCD Imager Passive retro-reflectors (4 mm) Infrared flash www.qualisys.se
Tongue motion	Carstens EMA system	Medizinelektronik AG100 10 sensor recording Midsaggital plane www.articulograph.de

DAT recorder. During recording, a DAT recorder and a directional Sennheiser microphone were used to obtain the acoustic signals. The sampling frequency of the audio was 44.1 kHz. Because the microphone had to be out of the way of video and facial motion recording, the microphone was put at an angle of about 30 degrees below the chin and at a distance of about 5 inches. For the DAT recorder, one channel was used to record the audio signal; the other one was used to record a sync pulse.

Qualisys™ Motion Capture System. Face motion was captured with a Qualisys™ optical motion capture system using three infrared emitting-receiving cameras. During recording, passive retro-reflectors glued on the talker are illuminated by infrared flashes (not perceptible to the talker) emitted from the three cameras. Each camera produces size and two-dimensional position information (in the imager plane) about the retro-reflectors, at the rate of 120 Hz. Using the information gathered from each camera, as well as

calibration data based on camera position, the 3-D position of each retro-reflector is calculated for each sampling period with a precision of 10 μm. The reconstruction of a retro-reflector's position depends on having data from at least two of the three cameras. When retro-reflectors are only seen by a single camera, dropouts in the motion data occur (missing data). In addition, when two retro-reflectors are too close to one another, dropouts occur. Usually, dropouts were only a few frames in duration, and only one or two retro-reflectors were missing at a time.

VTR. A Sony recorder and a Sony digital video camera were used to record the video. The video was recorded onto BETACAM SP tape. Lighting and positioning were carefully adjusted so as to get the best view, and talkers were lighted with a spotlight at both sides at the same height as their face and with an overhead fill light. During recording, talkers watched the screen of a teleprompter that was mounted below the camera.

EMA system. A Carstens AG-100 EMA system was used to track the tongue movements. The principle of the system is illustrated in Figure 2.1 (*also see UCLA Phonetics Lab's website* ***http://www.linguistics.ucla.edu/faciliti/uclaplab.html***). During recording, a talker wears a helmet with three fixed transmitters: one near the forehead, one near the chin, and one near the back neck. The positions of these transmitters form an equilateral triangle whose sides limit the measurement plane of the system. These transmitters produce alternating magnetic fields at three different frequencies, and the corresponding alternating magnetic fields induce three alternating currents in the sensors. The distance between each sensor and each transmitter can be determined from the

strength of the current at a certain frequency (the electromagnetic field strength in a receiver is inversely proportional to the cube of its distance from a transmitter). It is then possible to calculate each sensor's XY coordinates and its tilt relative to the transmitter axes. To track vocal tract motion, a number of receiver pellets (sensors) are placed on talkers' articulators (such as tongue, jaw, lips and/or teeth) along the midsaggital plane. The induced voltages on these receiver pellets are sampled at 666 Hz. Thus each pellet's movements can be recovered. This EMA system could track up to 10 pellets, but fewer were used in the current experiments.

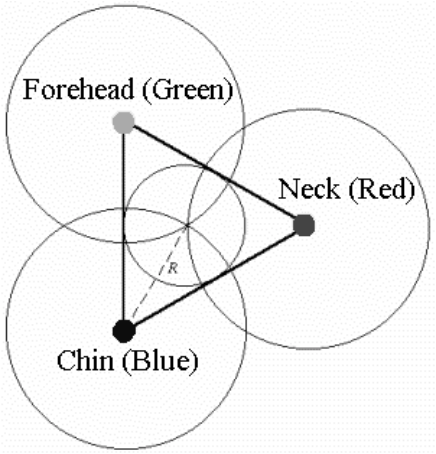


Figure 2.1 How the EMA system works.

2.3.4. Placement of EMA Pellets and Qualisys^{TM} Retro-Reflectors

Physical attributes were actually space- and time- sampled face (or tongue) movements. Space sampling was performed by carefully choosing a number of points of interest on the face (or tongue). Due to system limitations, only a small number of markers were used to sample the face (or tongue) movements. Figure 2.2 shows the overall

configuration of helmet, receivers, wires etc. when a talker is set up for simultaneous EMA and Qualisys^{TM} recording.

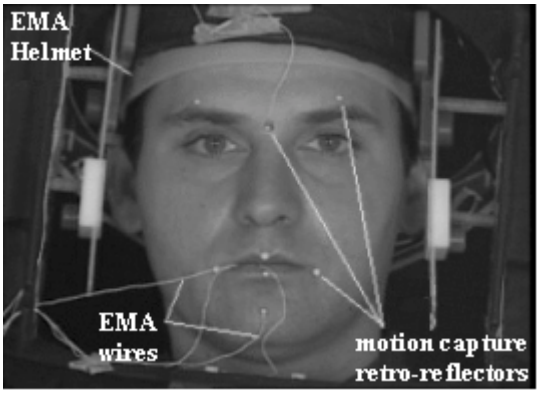


Figure 2.2 Optical retro-reflectors and EMA pellets.

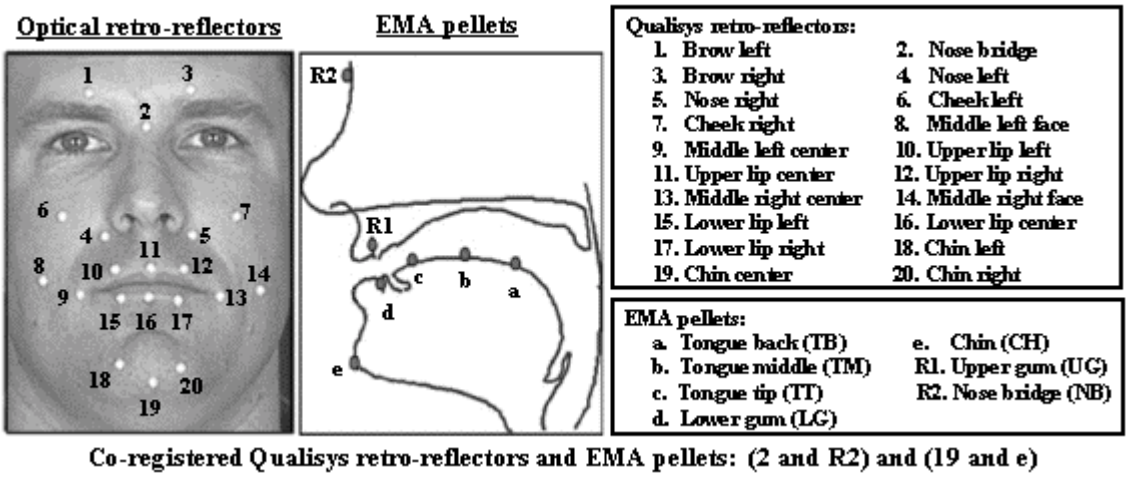


Figure 2.3 Placement of optical retro-reflectors and EMA pellets.

Figure 2.3 shows the placement of the optical retro-reflectors and EMA pellets in the current experiments. Three EMA pellets (tongue back, tongue middle, and tongue tip) were placed on the tongue, one on the lower gum (for jaw movement), one on the upper gum, one on the chin, and one on the nose bridge. One EMA channel, which was used for

synchronization with the other data streams, and two pellets, which were used only at the beginning of each session for defining the bite plane, are not shown in Figure 2.3. The pellets on the nose bridge (R2) and upper gum (R1), the most stable sensors available, were used for reference only (robust head movement correction). The pellet on the chin (e), which was co-registered with an optical retro-reflector, was used only for synchronization of tongue and face motion, because a chin retro-reflector (19) was used to track face movements. The chin generally moves with the jaw, except when the skin is pulled by the lower lip. This can happen, for example, when bilabial stops are produced; in such cases, the chin sometimes rises while the jaw stays still or moves down. The pellet on the lower gum (d), which is highly correlated with the chin pellet (e), was not used in analysis. Hence, only the movements from the three pellets on the tongue (a, b, c in Figure 2.3) went into the analysis of tongue movements.

There were 20 optical retro-reflectors. They were placed on the nose bridge (one), eyebrows (two), lip contour (eight), chin (three), and cheeks (six). The retro-reflectors on nose bridge and eyebrows (retro-reflectors 1, 2, and 3) were used for head movement compensation only (they were not used in the analyses). The 17 retro-reflectors used to quantify speech movement were divided into three sets: lip retro-reflectors (9, 10, 11, 12, 13, 15, 16, and 17), chin retro-reflectors (18, 19, and 20), and cheek retro-reflectors (4, 5, 6, 7, 8, and 14).

2.3.5. Synchronization

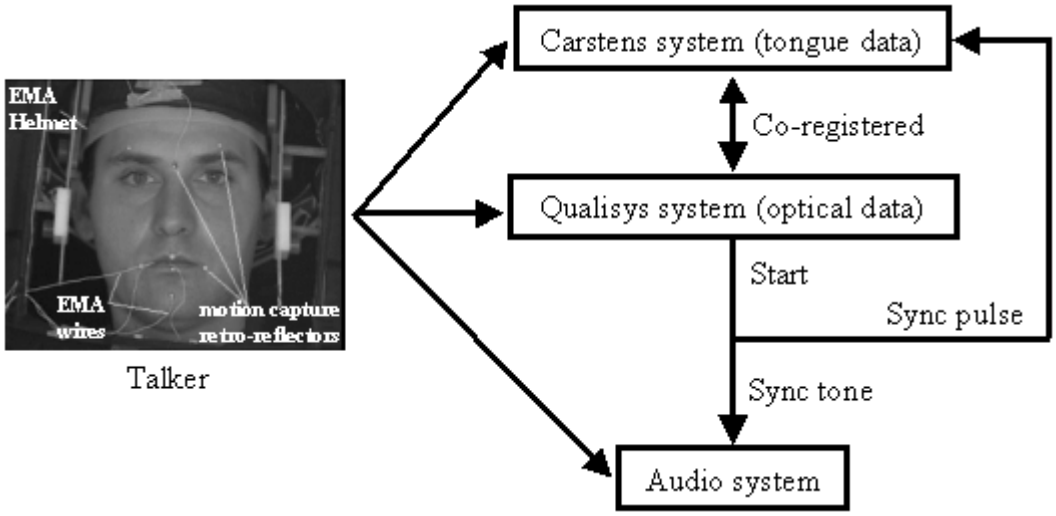


Figure 2.4 Synchronization of optical, EMA, and audio.

EMA and optical data were temporally aligned by the co-registered EMA pellet and optical retro-reflector on the chin as well as by a special time-sync signal. At the beginning of each recording, a custom circuit (Bernstein et al., 2000a), which analyzed signals from the optical and video systems, invoked a 100-ms pulse that was sent to one EMA channel and a 100-ms 1-kHz pure tone that was sent to the DAT line input for synchronization. Figure 2.4 illustrates the synchronization scheme. The Qualisys^{TM} system initiated the sync tone and sync pulse first; the audio system was then synchronized by finding the tone position. The sync pulse in the EMA data can help to find an approximate starting point, and then an alignment between the co-registered chin pellet and retro-reflector gives an exact synchronization between EMA and optical data.

2.3.6. Recording Procedure

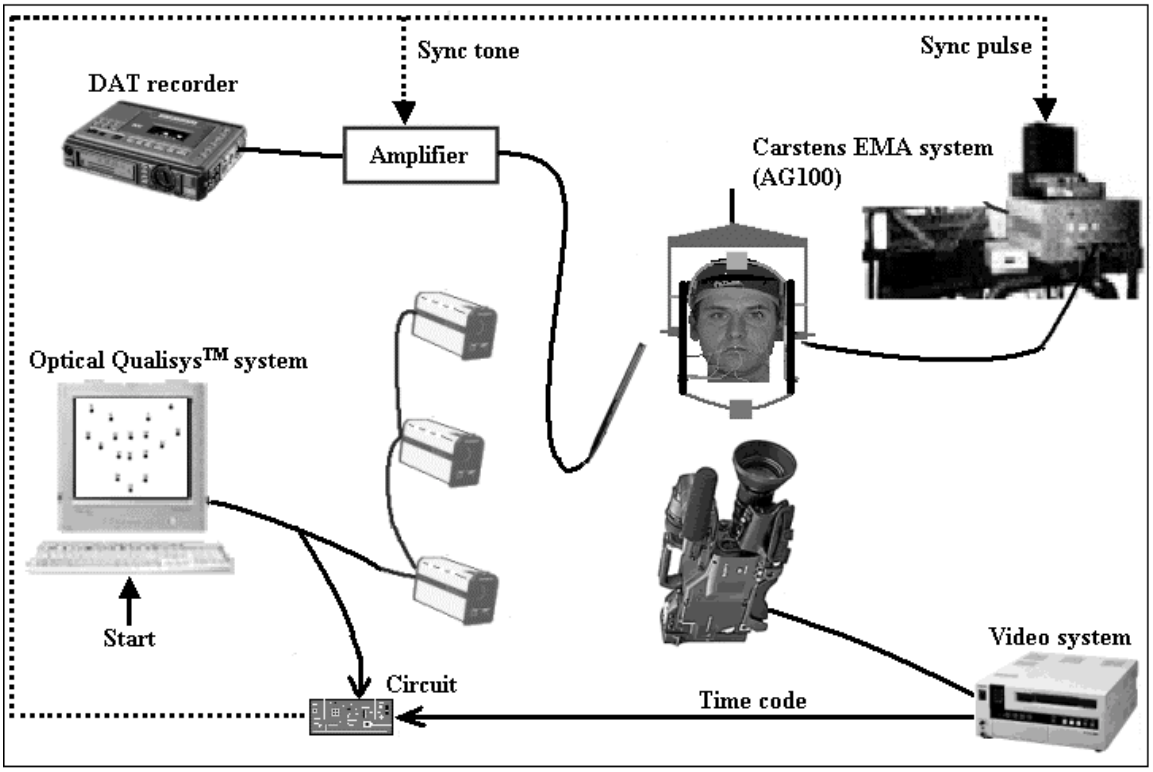


Figure 2.5 Overall recording scene.

The overall recording scene is illustrated in Figure 2.5. Talkers sat inside a sound-treated booth. The recording crew stayed in another room, monitored the talker through the video display, and communicated with the talker through microphones and speakers (the speaker inside the sound-treated booth was turned off during recording). In each recording session, a crew of about 5-6 people were involved and assigned to EMA, Qualisys™, teleprompter, audio, overall direction, and communication with subject. Prior to recording, they practiced producing each syllable. Talkers were coached prior to recording to produce the stimuli with a falling intonation. A pseudo-randomly ordered

syllable or sentence list was presented on the teleprompter. They looked directly into the camera, and their face filled the video monitor.

Each session lasted about 4 hours. It takes 1.5 - 2 hours to place retro-reflectors, calibrate, and confirm data quality. Note that the recording did not go on continuously. Due to memory limitations of the Qualisys^{TM} and Carstens systems, each recording (one take) lasted 22 seconds. After each take, we saved the optical and EMA data, and then began a new take. A green light inside the sound-treated booth indicated recording. After each session, the data related to this dissertation were processed as follows: Raw video was copied to dubs; video in and out point time codes were identified for each take, and a take-list was compiled; the 3-D retro-reflectors were tracked across frames and labeled; the EMA data were filtered and rotated into the midsaggital plane; a master listing was generated to provide the synchronization information across different data streams; audio data on the DAT tape were transferred to PC. Note that each optical file represented data for one take (22 seconds), while audio data were recorded continuously. Each EMA file represents data for one take but subject to synchronization.

2.4. Recording a Database for the Perceptual Similarity Analysis

The database for the perceptual similarity analysis was similar to that used in the correlation analysis. Acoustic signals were recorded and were used to select the starting and ending points for the analyses of optical speech. However, there are several major differences:

1. Tongue movements were not recorded, and the EMA system was not used.
2. Only CV syllables were recorded (no sentences).
3. Each CV syllable was repeated at least twice. The crew monitored each CV production. If one syllable was not produced well, then that syllable and the following syllables in the same take was repeated. For example, if the first syllable in a take needed to be repeated, then the whole take was repeated; if the last syllable in a take needed to be repeated, then only that syllable was repeated. Therefore, each syllable had at least two good repetitions.

2.5. Perceptual Experiments

2.5.1. Participants

The participants were six individuals (four women and two men; average age 32 years; range 22 to 43) with normal hearing, normal or corrected-to-normal vision, English as a native language, and average or better lipreading.

2.5.2. Video Presentations

For the similarity database, each CV syllable had at least two good repetitions. However, for the perceptual experiments, only two repetitions were selected for the current study; each began with the mouth closed and ended with a closed mouth. During the perceptual identification testing, only the video recordings were presented to participants. The

selected tokens were stored on BETACAM SP video in lists that were blocked by talker. Tokens were pseudo-randomized across consonants and vowels to counterbalance the effects of token order. Two tapes were recorded. On Tape 1, tokens were pseudo-randomized, and the talker order was M1-F2-M2-F1. On Tape 2, tokens were pseudo-randomized and, the order was F2-M1-F1-M2.

2.5.3. Procedure

Testing took place in a sound-treated IAC booth. A Sony BETACAM videotape player (outside the sound-treated booth) was controlled by a personal computer that was placed on the table (inside the sound-treated booth). Stimuli were presented on a 19" high-resolution color monitor (Sony Trinitron) placed next to the PC monitor at a distance of about one meter from the participants.

Visual perception was assessed using high-quality video recordings as described in Section 2.5.2. The two videotapes (without sound) were presented to each participant five times. Therefore, for each CV syllable, there were 480 responses (2 repetitions, 2 tapes, 5 trials, 4 talkers, and 6 participants).

The participants' task was to make visual-only 23-alternative forced-choice identification. Testing was administered in blocks of 138 pseudo-randomized items (2 repetitions x 69 tokens) for each of the four talkers. At the beginning of each session, instructions were displayed on the PC monitor. After the participants acknowledged reading the instructions, a computer program proceeded to present each of the video stimuli. A practice set of 10 trials was given on Day 1 only. No feedback was given at

any time. There was no time limitation on how fast the participants should respond to each visual stimulus.

After each stimulus presentation, the video monitor became black. At the same time, the graphical display on the PC monitor was activated, showing the 23 consonant labels (in an arbitrary order) with corresponding sample words to exemplify pronunciation. The participant then chose (guessing if necessary) a response using a computer mouse. Once the participant responded, no modification was allowed. Following the response, the computer program automatically switched to the presentation of the next syllable. Typically, the participants responded to one list for each talker on each day of testing. Each list took about 16 minutes to finish, and a five-minute break was given between lists. The order of viewing the tapes was counter-balanced across participants. Half of the participants began with Tape 1 and the other half began with Tape 2, and then they viewed the other tape. Occasionally, participants received more than four blocks of trials but not more than eight. A half-hour rest was always given when the number of blocks exceeded four. Each participant contributed ten responses for each stimulus token. Five participants completed testing within three weeks, and the sixth within eight.

2.6. Conditioning the Data

2.6.1. Audio

The audio signals were continuously recorded. Therefore, the first step was to find the

starting point for each take. The synchronization was done by finding the position of the pure tone as shown in Figures 2.6 and 2.7. After locating the tone position, a length of 22 seconds audio was cut out for each take.

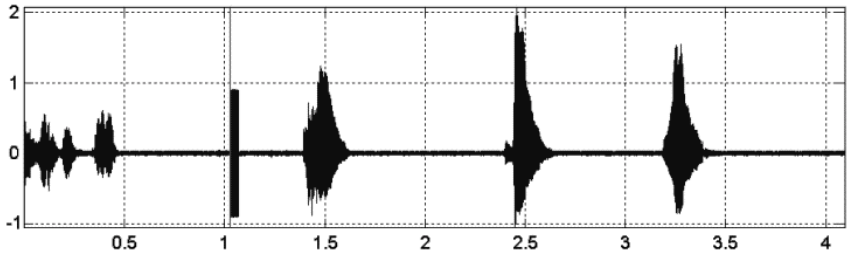


Figure 2.6 Sync tone in audio signals.

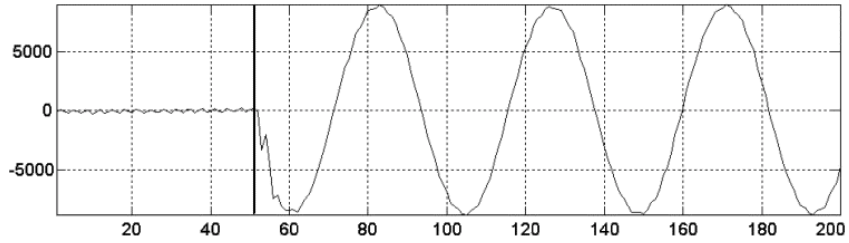


Figure 2.7 A close look at the sync tone.

Speech acoustic signals were originally sampled at 44.1 kHz and then down-sampled to 14.7 kHz. Speech signals were then divided into frames. The frame length and shift were 24 ms and 8.33 ms, respectively. Thus the frame rate was 120 Hz, which was consistent with the Qualisys^{TM} sampling rate. For each acoustic frame, pre-emphasis was applied. Then a covariance-based LPC algorithm (Rabiner and Schafer, 1978) was used to obtain 16th-order Line Spectral Pair (LSP) parameters (eight pairs) (Sugamura and Itakura, 1986). If the vocal tract is modeled as a non-uniform acoustic tube of p sections of equal length (p = 16 in this dissertation), the LSP parameters indicate the resonant

frequencies at which the acoustic tube shows a particular structure under a pair of extreme artificial boundary conditions: complete opening and closure at the glottis. LSPs are derived from the LPC polynomial $A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}$ that can be decomposed into:

$$P(z) = A(z) + z^{-(p+1)} A(z^{-1}) \tag{2.1}$$

$$Q(z) = A(z) - z^{-(p+1)} A(z^{-1}) \tag{2.2}$$

The roots of $P(z)$ and $Q(z)$ lie on the unit circle and alternate. LSPs have a tendency to approximate the formant frequencies (Sugamura and Itakura, 1986; Yehia et al., 1998). LSP parameters have good temporal interpolation properties, which are desirable (Yehia et al., 1998). The RMS energy (in dB) was also calculated.

2.6.2. Optical Data

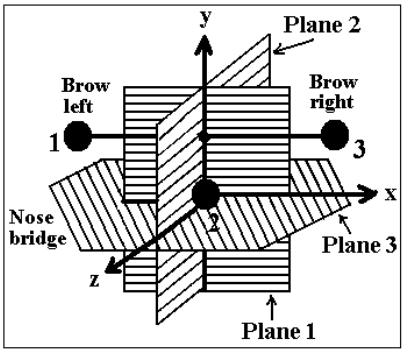


Figure 2.8 A new 3-D coordinate system defined by retro-reflectors 1, 2, and 3.

Compensation for head movements. Although the talkers were instructed to sit quietly

and focus on the camera (actually, the teleprompter), small head movements occurred. The recording session for each talker lasted about one hour. Over such a long period, the head movements were noticeable. In order to quantify and examine face movements, it was necessary to compensate for head movements. During recording, the retro-reflectors on the nose bridge (2) and eyebrows (1 and 3) were relatively stable across the session, and their movements were mainly due to head movements. Note that for spontaneous speech, eyebrows can be very mobile. Keating et al. (2000), however, found that eyebrow movements accompany the production of focused words in sentences, but not of isolated words. Therefore, these three retro-reflectors were used for head movement compensation as shown in Figure 2.8. Plane 1 was through retro-reflectors 1, 2, and 3. Plane 2 was defined as perpendicular to the line between retro-reflectors 1 and 3 and through retro-reflector 2. Plane 3 was perpendicular to planes 1 and 2 and through retro-reflector 2. These three planes were vertical to each other and thus defined a 3-D coordinate system with the origin at the nose bridge. In the new axes, the x axis was vertical to plane 2 and represented left and right movements; the y axis was vertical to plane 3 and represented up and down movements; and the z axis was vertical to plane 1 and represented near and far movements. Although the two retro-reflectors on the eyebrows had small movements, they usually moved in the same direction. Therefore, these planes were relatively stable.

Compensation for facial retro-reflector dropouts. During the recording of the CV syllables and sentences, there was a small percentage of dropouts of optical retro-reflectors (Table 2.2). Dropout means that in the reconstructed 3-D optical retro-reflector

movements, the motion data for one of the retro-reflectors disappeared for a few frames, as shown in Figure 2.9. Because the dropouts happened for only a few frames, and only one or two retro-reflectors were missing at the same time, the remaining movements of the retro-reflector and movements from other retro-reflectors were used to predict missing segments. One example is shown in Figure 2.9: Retro-reflector 8 (Middle left face) was missing for 12 frames. Although the face was not strictly symmetrical, retro-reflector 14 (Middle right face) was highly correlated with retro-reflector 8. Non-dropout frames from retro-reflectors 8 and 14 were used to predict the missing data using least-squares criterion. This method can be expressed as follows:

$$x_{\text{MLF\_good}} = ax_{\text{MRF\_good}} + b \tag{2.3}$$

$$x_{\text{MLF\_dropout}} = ax_{\text{MRF\_dropout}} + b \tag{2.4}$$

where x is a value on the X coordinate. The non-dropout frames were used to define transformation parameters a and b (using least-squares method). Then these parameters were applied to predict the missing data.

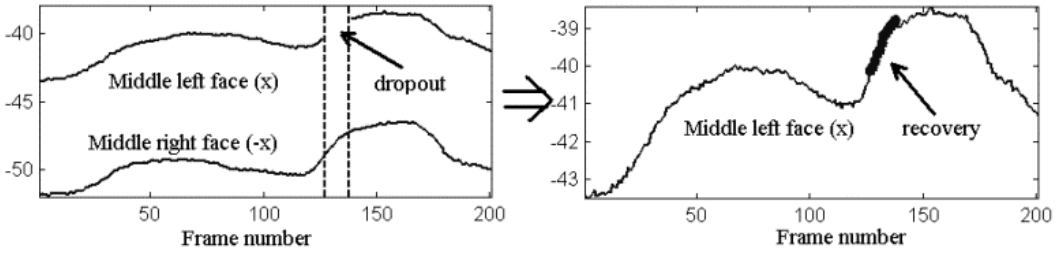


Figure 2.9 Missing data recovery for optical data.

Table 2.2 Statistics of retro-reflector dropouts during the recording of CV syllables.

	Percentage Retro-reflector Dropout (%)			
	Talkers			
	M1	M2	F1	F2
8. Middle left face	3.99	0	4.09	2.30
9. Middle left center	0	0	0.06	2.00
10. Upper lip left	0	0	0.09	0.08
11. Upper lip center	0.97	0	4.07	0
12. Upper lip right	0.15	0	1.88	1.94
13. Middle right center	0.08	0	0	0
14. Middle right face	2.45	0	0	0
15. Lower lip left	2.45	0	6.04	7.39
16. Lower lip center	3.56	9.86	0	0
17. Lower lip right	0.11	0	1.55	1.55
18. Chin left	0	0	21.05	0
19. Chin center	0	0	0.06	0
20. Chin right	0	0	13.14	0

2.6.3. EMA Data

Each EMA file represents data for one take. However, EMA recording did not start at the same time with Qualisys^{TM} data. Therefore, in the EMA data, we need to find when the optical data recording started. This was done by (a) finding the sync pulse in the EMA sync channel and (b) comparing the chin pellet and chin retro-reflector to get a fine alignment. In Figure 2.10, a sync pulse was found in one EMA channel and this pulse indicated the beginning of Qualisys^{TM} recording. A further correlation analysis between chin pellet and chin retro-reflector gave a more accurate alignment (Figure 2.11).

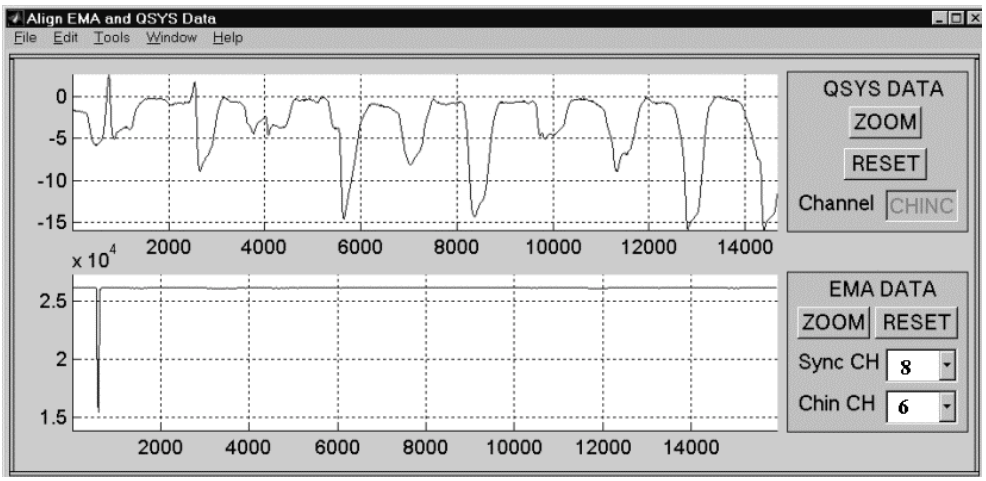


Figure 2.10 Finding the sync pulse in EMA data.

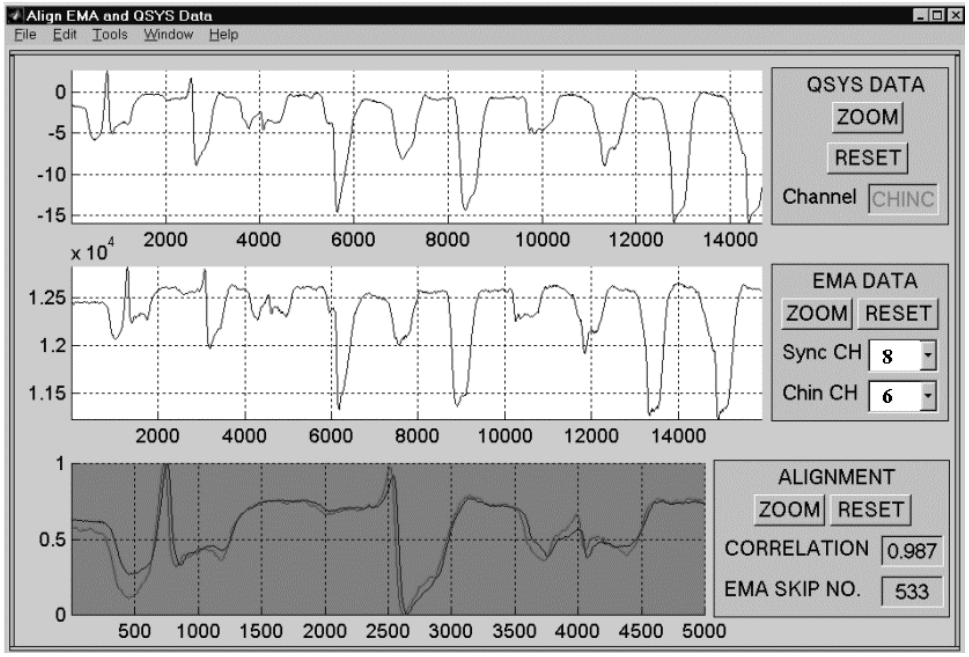


Figure 2.11 Alignment between the EMA chin pellet and optical chin retro-reflector.

2.6.4. All Three Data Streams

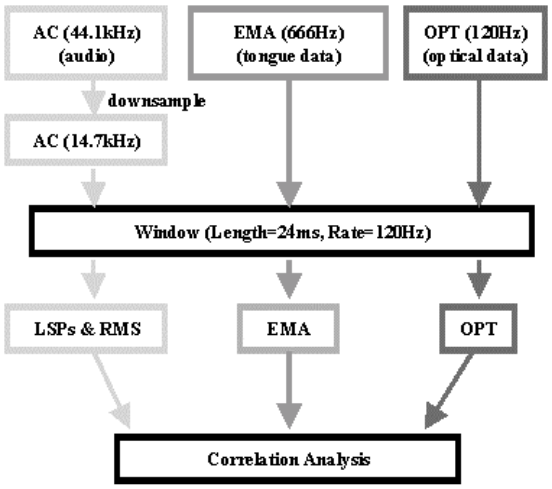


Figure 2.12 Conditioning of the three data streams.

Figure 2.12 shows how data were processed so that the frame rate was uniform. Hereafter, the following notation will be used: AC for acoustic data, OPT for optical Qualisys^{TM} data, EMA for magnetometer tongue data, LSP for Line Spectral Pairs, LSPE for LSP plus RMS energy, and RMS for RMS energy.

2.7. Summary of Physical Measures and Perceptual Data

2.7.1. Physical Measures

Table 2.3 A summary of the three datasets (four talkers).

Database	Corpus	Repetitions	Common data streams	EMA
DcorCV	69 CVs	≥4	Audio, Video, and Qualisys^{TM}	Yes
DcorSENT	3 sentences	≥4		Yes
DsimCV	69 CVs	≥2		None

There were three different datasets recorded for the correlation and similarity analyses as summarized in Table 2.3. Datasets *DcorCV* and *DcorSENT* were recorded for the correlation analysis. Because there were four or above four repetitions of each CV syllable or sentence, only the first four repetitions were used in the analysis. Dataset *DsimCV* was recorded for similarity analysis. Two repetitions of each CV syllable were chosen based on the criterion that the mouth was closed before and after producing the syllable. The difference between *DcorCV* and *DcorSENT* is that *DcorCV* consisted of productions of the 69 CV syllables, while *DcorSENT* consisted of productions of the three sentences. *DsimCV* differs from *DcorCV* and *DcorSENT* in three aspects: *DsimCV* does not have EMA data; *DsimCV* has two or above two repetitions of each CV syllable; and *DsimCV* has different selection criterion.

Table 2.4 summarizes the channels for each data stream used in this dissertation. For EMA data, we were interested in tongue movements, so only tongue back, tongue middle, and tongue tip were used in the analysis. For optical data, the retro-reflectors 1, 2, 3 were relatively stable and were used for head movement removal so that they were not used in the analyses. The optical channels were grouped according to their positions on the face: lips (9, 10, 11, 12, 13, 15, 16, and 17); chin (18, 19, and 20); cheeks (4, 5, 6, 7, 8, and 14). For acoustic data, LSP pairs 1-8 and RMS energy were used in the analysis.

Table 2.4 A summary of data channels used in the analysis.

Data streams	Channels used in the analysis
Optical data	Lip retro-reflectors (lip) (9, 10, 11, 12. 13, 15, 16, and 17)
	Chin retro-reflectors (chn) (18, 19, and 20)
	Cheek retro-reflectors (chk) (4, 5, 6, 7, 8, and 14)
EMA data	Tongue back (TB), Tongue middle (TM), and Tongue tip (TT)
Acoustic data	RMS energy (L0), LSP pairs 1-8 (L1-L8)

The data were first organized into matrices (Horn and Johnson, 1985). Each EMA frame is a six-dimensional vector (x and y coordinates of the three moving pellets: TB, TM, and TT). Each OPT frame is a 51-dimensional vector (x, y, and z coordinates of the 17 retro-reflectors). Each LSPE frame is a 17-dimensional vector (16 LSP parameters and RMS energy). Let $\boldsymbol{O}$, $\boldsymbol{E}$, and $\boldsymbol{L}$ represent the OPT, EMA, and LSPE matrices, respectively. Using $\boldsymbol{O}$ as an example, each matrix can be written in the form of

$$\boldsymbol{O} = \begin{bmatrix} o_{1,1} & \cdots & o_{1,\mathrm{N}} \\ \vdots & \vdots & \vdots \\ o_{51,1} & \cdots & o_{51,\mathrm{N}} \end{bmatrix} \tag{2.5}$$

where N is the number of frames. For optical data, there are three subsets: $\boldsymbol{O}_{\mathrm{lip}}$, $\boldsymbol{O}_{\mathrm{chn}}$, and $\boldsymbol{O}_{\mathrm{chk}}$. EMA data also have three subsets: $\boldsymbol{E}_{\mathrm{TB}}$, $\boldsymbol{E}_{\mathrm{TM}}$, and $\boldsymbol{E}_{\mathrm{TT}}$. For acoustic data, there are nine subsets: $\boldsymbol{L}_0$ (RMS energy) and $\boldsymbol{L}_1$-$\boldsymbol{L}_8$ (LSP pairs 1-8).

2.7.2. Perceptual Data

Table 2.5 An example of a stimulus-response confusion matrix.

		R	E	S	P	O	N	S	E										
		y	w	r	L	m	n	p	t	k	b	d	g	h	θ	ð	s	z	f	v	ʃ	ʒ	tʃ	dʒ
	y	301	54	1	139		153		53	194	1	121	89	189	17	9	30	5	32	14	16	5	6	11
	w	8	1284	93		39		1	1		8	1		1						3				1
	r	4	374	372	3	30	3	2	7	3	4	6		4	3		12	5	504	98	1		1	4
	l	111	10		855	1	89		66	43	1	72	39	42	41	11	41	10			2		3	3
	m			1	1	683		221			528	2		2			1					1		
	n	80	12	2	534		176		132	73		152	43	102	16	5	56	20	1		11	2	12	11
	p		5	4		506	1	349			569	1					3			1	1			
S	t	68	16		37	2	75		300	54	1	298	24	70	40	20	266	81			26	4	17	41
T	k	100	81	2	71	2	64		37	293	1	63	124	570	8	2	12	5	2		2		1	
I	b			2	2	487		351			595			1					1					1
M	d	80	19	1	127		118		266	21		286	26	31	28	7	263	71	2	2	27	5	19	41
U	g	147	50	2	109		90		51	291		83	142	342	11	3	71	17	1		11	3	3	13
L	h	29	124	1	11	4	44	2	8	105	1	21	49	1018	7		9	1	1	1	1		2	1
U	θ	3	9		80	3	34		8	3		7	6	1	780	502	1	1		1				1
S	ð	14	2		55	1	50		52	12	1	57	18	23	680	452	13	6	2		1	1		
	s	17	3		1		5		149	2	1	191	2	3	14	7	711	230		3	39	5	13	44
	z	20	1	1	3	1	6	5	145	3	14	191	4	2	7	3	683	220			46	12	24	49
	f			1	1		1		2		1	1		2	1	1	1	1	1104	323				
	v	2	1	4			1		1			1		1	4				1083	341	1			
	ʃ	29			3		10	1	66	7	1	81	36	6	4		282	73			294	67	216	264
	ʒ	22			8	1	6		78	1		78	64	4	1	5	221	69	1	1	324	54	225	277
	tʃ	63	3		5		7	2	116	12	2	90	61	8	4	2	224	80		1	256	38	242	224
	dʒ	41	7		4		4		79	5	1	72	63	3	2	1	226	83			306	46	233	264

In perceptual experiments, each stimulus can receive different responses. For example, a visual stimulus /b/ may be perceived as /p/ or /m/. In one stimulus-response condition, the same stimulus is presented a certain number of times. Thus a distribution of responses can be obtained. To display and further process this information, the results are put together in a stimulus-response confusion matrix as shown in Table 2.5. “STIMULUS” appears in the first column, which indicates that each row represents a distribution of responses to one stimulus. The stimuli are listed in the second column. “RESPONSE” appears in the first row, which indicates each column representing how many times the stimulus was perceived as this type. The names of response types are listed in the second row, and they are the same as the stimulus names. For example, in the table, across h

(row) and under k (column), the value is 105, which means that among 1440 (row total) presentations of the /h/ stimulus, the participants perceived it as /k/ 105 times.

Perceptual data consisted of six participants' identifications of 23 consonants through lipreading each of the four talkers. Across all of the testing, for each CV syllable, there were 480 responses (2 repetitions x 10 trials x 4 talkers x 6 participants). Results are shown in stimulus-response confusion matrices. The results were pooled in 12 stimulus-response confusion matrices (23x23) for each talker and vowel context. $\boldsymbol{C}_{T,V}$ represents one stimulus-response confusion matrix, where T is the talker, and V is the vowel context. In each of these confusion matrices, there were 120 responses (2 repetitions x 6 participants x 10 trials) per stimulus.

Ignoring either the talker or vowel factor, two types of confusion matrices can be computed, respectively, as:

$$\boldsymbol{C}_{\mathrm{ALL},V} = \boldsymbol{C}_{\mathrm{M1},V} + \boldsymbol{C}_{\mathrm{F1},V} + \boldsymbol{C}_{\mathrm{M2},V} + \boldsymbol{C}_{\mathrm{F2},V} \tag{2.6}$$

$$\boldsymbol{C}_{T,\mathrm{aiu}} = \boldsymbol{C}_{T,\mathrm{a}} + \boldsymbol{C}_{T,\mathrm{i}} + \boldsymbol{C}_{T,\mathrm{u}} \tag{2.7}$$

Table 2.6 lists the resulting confusion matrices (23x23; see Appendix A) that were analyzed in Chapter 5.

Table 2.6 A summary of the perceptual confusion matrices

Confusion matrix	Talker	Vowel	Response number	Confusion matrix	Talker	Vowel	Response number
$\boldsymbol{C}_{\mathrm{M1,a}}$	M1	/a/	120	$\boldsymbol{C}_{\mathrm{M1,i}}$	M1	/i/	120
$\boldsymbol{C}_{\mathrm{M1,u}}$	M1	/u/	120	$\boldsymbol{C}_{\mathrm{M1,aiu}}$	M1	/aiu/	360
$\boldsymbol{C}_{\mathrm{F1,a}}$	F1	/a/	120	$\boldsymbol{C}_{\mathrm{F1,i}}$	F1	/i/	120
$\boldsymbol{C}_{\mathrm{F1,u}}$	F1	/u/	120	$\boldsymbol{C}_{\mathrm{F1,aiu}}$	F1	/aiu/	360
$\boldsymbol{C}_{\mathrm{M2,a}}$	M2	/a/	120	$\boldsymbol{C}_{\mathrm{M2,i}}$	M2	/i/	120
$\boldsymbol{C}_{\mathrm{M2,u}}$	M2	/u/	120	$\boldsymbol{C}_{\mathrm{M2,aiu}}$	M2	/aiu/	360
$\boldsymbol{C}_{\mathrm{F2,a}}$	F2	/a/	120	$\boldsymbol{C}_{\mathrm{F2,i}}$	F2	/i/	120
$\boldsymbol{C}_{\mathrm{F2,u}}$	F2	/u/	120	$\boldsymbol{C}_{\mathrm{F2,aiu}}$	F2	/aiu/	360
$\boldsymbol{C}_{\mathrm{ALL,a}}$	ALL	/a/	480	$\boldsymbol{C}_{\mathrm{ALL,i}}$	ALL	/i/	480
$\boldsymbol{C}_{\mathrm{ALL,u}}$	ALL	/a/	480	$\boldsymbol{C}_{\mathrm{ALL,aiu}}$	ALL	/aiu/	1440

2.8. Summary

This chapter described (1) the method of recording databases for the correlation analysis covering the issues of talkers, materials, recording facilities, placement of optical retro-reflectors and EMA pellets, synchronization, and procedure, (2) the method for the second database for the perceptual similarity analysis, (3) methods for the perceptual experiment covering the issues of participants, video presentations, and procedure, and (4) methods for conditioning the physical streams (audio, optical data, EMA data) for the correlation and perceptual similarity analyses. Finally, the databases (physical and perceptual data) were summarized.

CHAPTER 3. MULTILINEAR REGRESSION, MULTIDIMENSIONAL SCALING, HIERARCHICAL CLUSTERING ANALYSIS, AND PHONEME EQUIVALENCE CLASSES

3.1. Introduction

This chapter describes the algorithms for the correlation and perceptual similarity analyses. For the correlation analysis, multilinear regression is used. For the perceptual similarity analysis, the perceptual stimulus-response confusion matrices are first transformed to dissimilarity matrices using phi-square transformation, and then the relationship between the perceptual and physical measures is examined using multilinear regression. For the perceptual data, hierarchical clustering analysis is applied to obtain phoneme equivalence classes (PECs), and multidimensional scaling (MDS) is used to analyze the structures of the dissimilarity matrices.

3.2. Multilinear Regression

3.2.1. Mean Subtraction

Mean subtraction is applied just before multilinear regression to eliminate bias. For each utterance, the channel mean is deducted from all the values in that channel.

3.2.2. Multilinear Regression

Multilinear regression is chosen as the method for deriving relationships between the various obtained measures. Multilinear regression fits a linear combination of the components of a multi-channel signal $\boldsymbol{X}$ to a single-channel signal $\boldsymbol{y}_j$ and a residual error vector:

$$\boldsymbol{y}_j = a_1\boldsymbol{x}_1 + a_2\boldsymbol{x}_2 + \cdots + a_I\boldsymbol{x}_I + \boldsymbol{b} \tag{3.1}$$

where $\boldsymbol{x}_i$ (i=1,2, …, I) is one channel of the multi-channel signal $\boldsymbol{X}$, a_i is the weighting coefficient, and $\boldsymbol{b}$ is the residual vector. In multilinear regression, the objective is to minimize the root mean square error $\|\boldsymbol{b}\|_2$, so that:

$$\boldsymbol{a} = \text{argmin}\left\{\left\|\boldsymbol{X}^T\boldsymbol{a} - \boldsymbol{y}_j^T\right\|_2\right\} \tag{3.2}$$

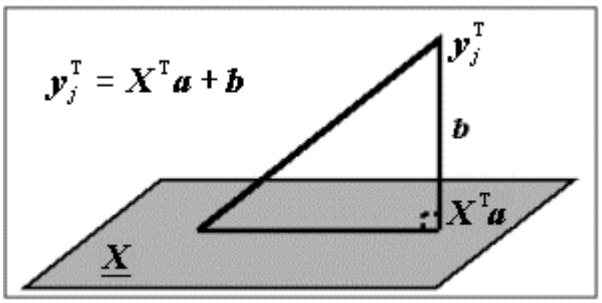


Figure 3.1 Illustration of multilinear regression.

This optimization problem has a standard solution (Kailath et al., 2000; Sen and Srivastava, 1990). Let $\underline{\boldsymbol{X}}$ represent the range of matrix $\boldsymbol{X}$ (linear combination of column vectors from $\boldsymbol{X}^T$). Thus $\boldsymbol{X}^T\boldsymbol{a}$ is one line in the $\underline{\boldsymbol{X}}$ plane. To obtain the most information about the target $\boldsymbol{y}_j$ from $\underline{\boldsymbol{X}}$, the error signal $\boldsymbol{b}$ should be vertical to the $\underline{\boldsymbol{X}}$ plane (see Figure

3.1):

$$\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a} - \boldsymbol{y}_j^{\mathrm{T}}) = 0 \tag{3.3}$$

Thus,

$$\boldsymbol{a} = (\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\boldsymbol{X}\boldsymbol{y}_j^{\mathrm{T}} \tag{3.4}$$

3.2.3. Jackknife Procedure

For correlation analysis between face movements, tongue movements, and speech acoustics, the data (*DcorCV* and *DcorSENT*) were limited, compared to the very large databases used in automatic speech recognition. Therefore, a leave-one-out Jackknife training procedure was applied to protect against bias in the prediction (Efron, 1982; Miller, 1974). First, data were divided into training and test sets. The training set was used to define a weighting vector ***a***, which was then applied to the test set.

Syllable-dependent, syllable-independent, and vowel-dependent predictions were performed for individual consonant-vowel (CV) syllable, all CV syllables, and vowel-grouped syllables (C/a/, C/i/, and C/u/ syllables), respectively. The differences between these prediction procedures were that, for syllable-dependent predictions, each syllable was treated separately; for syllable-independent predictions, all syllables were grouped together; and for vowel-dependent predictions, syllables sharing the same vowel were grouped together.

For syllable-dependent predictions, the data were divided into four sets, where

each set contained one repetition of a particular CV per talker. One set was left out for testing and the remaining sets were for training. A rotation was then performed to guarantee each utterance was in the training and test sets. For syllable-independent prediction, the data were divided into four sets, where each set had one repetition of every CV syllable per talker. For vowel-dependent prediction, the syllables were divided into four sets for each of the three vowels separately. For example, for C/a/ syllables, each set had one repetition of every C/a/ syllable per talker.

Figure 3.2 shows how the multilinear regression was performed for the prediction of LSP data from OPT data.

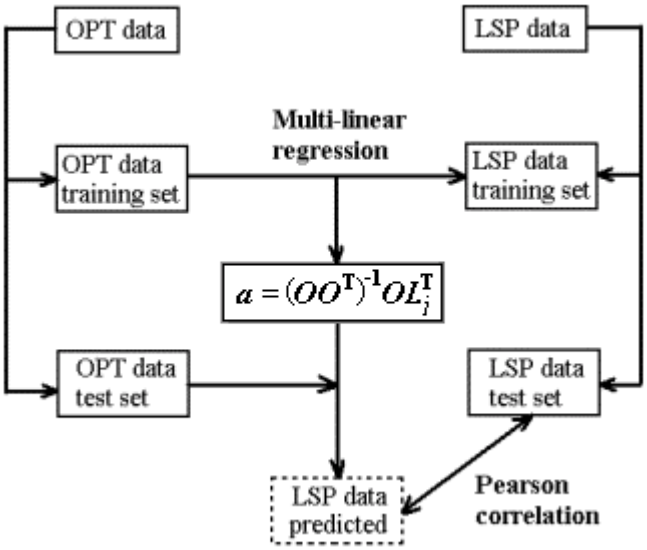


Figure 3.2 Diagram for the multilinear regression, where *O* and *L* stand for OPT and LSP, respectively.

3.2.4. Goodness of Fit

After applying the weighting vector ***a*** to the test data, a Pearson correlation is evaluated

between predicted (***Y***') and measured data (***Y***). The correlation was calculated as:

$$r_{Y'Y} = \frac{\sum\sum(y'_{j,n} - \overline{y'_j})(y_{j,n} - \overline{y_j})}{\sqrt{\sum\sum(y'_{j,n} - \overline{y'_j})^2} \cdot \sqrt{\sum\sum(y_{j,n} - \overline{y_j})^2}} \quad (3.5)$$

where ***Y***' is the predicted data, ***Y*** is the measured data, *j* is the channel number, and *n* is the frame number. For OPT and EMA data, all channels are used to calculate the correlation coefficients. For acoustic data, LSP channels are used to calculate correlation coefficients separately from the RMS channel. When examining the difference between different areas, such as the face areas, lip, chin, and cheeks, the related channels are grouped to compute correlation coefficients. For example, when estimating OPT data from LSPE, optical retro-reflectors 9, 10, 11, 12 13, 15, 16, and 17 are grouped together to compute the correlation coefficients for the lip area.

For the perceptual similarity analysis, ***Y***' and ***Y*** are one-dimensional (a vector of distances), so the correlation coefficients are simply calculated.

3.2.5. Maximum Correlation Criterion Estimation

Multilinear regression minimizes the root mean squared error between obtained and predicted measures. Correlation coefficients are used to measure the goodness of estimation. Hence, the maximum correlation coefficient can be used as an alternative criterion for linear estimation. Fortunately, it has been proven that the multilinear regression method is also optimized in the sense of maximum correlation when using linear programming techniques (Bertsimas and Tsitsiklis, 1997). The maximum

correlation criterion estimation can be formulated as:

$$\begin{aligned} &\text{maximize} \quad p \\ &\text{subject to} \quad \frac{\left|\boldsymbol{y}_j \boldsymbol{X}^{\mathrm{T}} \boldsymbol{a}\right|}{\left|\boldsymbol{y}_j\right| \cdot \left|\boldsymbol{X}^{\mathrm{T}} \boldsymbol{a}\right|} = p \end{aligned} \tag{3.6}$$

where p is the correlation coefficient, $\boldsymbol{a}$ is the estimator vector, $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}$ is the estimated target, $\boldsymbol{X}$ and $\boldsymbol{y}_j$ are the same as in Section 3.2.2. This operation can be written as:

$$\begin{aligned} &\text{maximize} \quad p \\ &\text{subject to} \quad \boldsymbol{a}^T (p^2 \left|\boldsymbol{y}_j\right|^2 \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} - \boldsymbol{X}\boldsymbol{y}_j^{\mathrm{T}} \boldsymbol{y}_j \boldsymbol{X}^{\mathrm{T}})\boldsymbol{a} = 0 \\ &\qquad \Rightarrow \boldsymbol{a}^{\mathrm{T}} \left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{\frac{1}{2}} \left[p^2 \left|\boldsymbol{y}_j\right|^2 \boldsymbol{I} - \left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-\frac{1}{2}} \boldsymbol{X}\boldsymbol{y}_j^{\mathrm{T}} \boldsymbol{y}_j \boldsymbol{X}^{\mathrm{T}} \left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-\frac{1}{2}} \right] \left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{\frac{1}{2}} \boldsymbol{a} = 0 \end{aligned} \tag{3.7}$$

where $p^2 \left|\boldsymbol{y}_j\right|^2 \boldsymbol{I} - \left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-\frac{1}{2}} \boldsymbol{X}\boldsymbol{y}_j^{\mathrm{T}} \boldsymbol{y}_j \boldsymbol{X}^{\mathrm{T}} \left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-\frac{1}{2}}$ is a positive semidefinite matrix or a regular matrix, depending on p. For the identity matrix $\boldsymbol{I}$, its eigenvectors can be any orthogonal unit vector set. Then $\left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-\frac{1}{2}} \boldsymbol{X}\boldsymbol{y}_j^{\mathrm{T}}$ can be one of its eigenvectors, and we can assume the corresponding other eigenvectors are $\boldsymbol{q_i}$ (i=1,2, M-1). Also, define,

$$\boldsymbol{q}_{\mathrm{M}} = \frac{\left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-\frac{1}{2}} \boldsymbol{X}\boldsymbol{y}_j^{\mathrm{T}}}{\left|\left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-\frac{1}{2}} \boldsymbol{X}\boldsymbol{y}_j^{\mathrm{T}}\right|} \tag{3.8}$$

Then,

$$
\begin{aligned}
& p^2\left|\boldsymbol{y}_j\right|^2 \boldsymbol{I}-\left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-\frac{1}{2}} \boldsymbol{X}\boldsymbol{y}_j^{\mathrm{T}} \boldsymbol{y}_j \boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-\frac{1}{2}} \\
&= p^2\left|\boldsymbol{y}_j\right|^2\left(\boldsymbol{q}_{\mathrm{M}}\boldsymbol{q}_{\mathrm{M}}^{\mathrm{T}}+\sum_{i=1}^{\mathrm{M}=1} \boldsymbol{q}_i\boldsymbol{q}_i^{\mathrm{T}}\right)-\left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-\frac{1}{2}} \boldsymbol{X}\boldsymbol{y}_j^{\mathrm{T}} \boldsymbol{y}_j \boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-\frac{1}{2}} \\
&= p^2\left|\boldsymbol{y}_j\right|^2\left(\sum_{i=1}^{\mathrm{M}=1} \boldsymbol{q}_i\boldsymbol{q}_i^{\mathrm{T}}\right)+p^2\left|\boldsymbol{y}_j\right|^2 \boldsymbol{q}_{\mathrm{M}}\boldsymbol{q}_{\mathrm{M}}^{\mathrm{T}}-\left|\left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-\frac{1}{2}} \boldsymbol{X}\boldsymbol{y}_j^{\mathrm{T}}\right|^2 \boldsymbol{q}_{\mathrm{M}}\boldsymbol{q}_{\mathrm{M}}^{\mathrm{T}} \\
&= p^2\left|\boldsymbol{y}_j\right|^2\left(\sum_{i=1}^{\mathrm{M}=1} \boldsymbol{q}_i\boldsymbol{q}_i^{\mathrm{T}}\right)+\left(p^2\left|\boldsymbol{y}_j\right|^2-\left|\left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-\frac{1}{2}} \boldsymbol{X}\boldsymbol{y}_j^{\mathrm{T}}\right|^2\right) \boldsymbol{q}_{\mathrm{M}}\boldsymbol{q}_{\mathrm{M}}^{\mathrm{T}}
\end{aligned}
\tag{3.9}
$$

Therefore, the eigenvalues of the matrix are $\lambda_1=\lambda_2=\cdots=\lambda_{\mathrm{M}-1}=p^2\left|\boldsymbol{y}_j\right|^2>0$ and $\lambda_{\mathrm{M}}=p^2\left|\boldsymbol{y}_j\right|^2-\left|\left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-\frac{1}{2}} \boldsymbol{X}\boldsymbol{y}_j^{\mathrm{T}}\right|^2$. λ_{M} can be positive or negative, depending on *p*. If *p* is small, λ_{M} is negative and the matrix $p^2\left|\boldsymbol{y}_j\right|^2 \boldsymbol{I}-\left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-\frac{1}{2}} \boldsymbol{X}\boldsymbol{y}_j^{\mathrm{T}} \boldsymbol{y}_j \boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-\frac{1}{2}}$ is not positive semidefinite. If *p* is large, λ_{M} is positive and the matrix is positive definite. If the matrix is positive definite, then there is no solution for the estimator $\boldsymbol{a}$. Therefore, the matrix should not be positive definite. The maximum *p* must then satisfy this equation:

$$
p^2\left|\boldsymbol{y}_j\right|^2-\left|\left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-\frac{1}{2}} \boldsymbol{X}\boldsymbol{y}_j^{\mathrm{T}}\right|^2=0 \tag{3.10}
$$

or

$$
p^2=\frac{\left|\left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-\frac{1}{2}} \boldsymbol{X}\boldsymbol{y}_j^{\mathrm{T}}\right|^2}{\left|\boldsymbol{y}_j\right|^2} \tag{3.11}
$$

In Section 3.2.2, we showed that $\boldsymbol{a}=\left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-1} \boldsymbol{X}\boldsymbol{y}_j^{\mathrm{T}}$. The resulting correlation from multilinear regression is:

$$
\begin{aligned}
p^2_{\text{multiR}} &= \frac{\boldsymbol{y}_j \boldsymbol{X}^{\mathrm{T}} \boldsymbol{a}\boldsymbol{a}^{\mathrm{T}} \boldsymbol{X}\boldsymbol{y}_j^{\mathrm{T}}}{\left|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}\right|^2 \cdot \left|\boldsymbol{y}_j\right|^2} \\
&= \frac{\boldsymbol{y}_j \boldsymbol{X}^{\mathrm{T}} \left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-1} \boldsymbol{X}\boldsymbol{y}_j^{\mathrm{T}} \boldsymbol{y}_j \boldsymbol{X}^{\mathrm{T}} \left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-1} \boldsymbol{X}\boldsymbol{y}_j^{\mathrm{T}}}{\boldsymbol{y}_j \boldsymbol{X}^{\mathrm{T}} \left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-1} \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} \left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-1} \boldsymbol{X}\boldsymbol{y}_j^{\mathrm{T}} \left|\boldsymbol{y}_j\right|^2} \\
&= \frac{\boldsymbol{y}_j \boldsymbol{X}^{\mathrm{T}} \left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-1} \boldsymbol{X}\boldsymbol{y}_j^{\mathrm{T}} \boldsymbol{y}_j \boldsymbol{X}^{\mathrm{T}} \left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-1} \boldsymbol{X}\boldsymbol{y}_j^{\mathrm{T}}}{\boldsymbol{y}_j \boldsymbol{X}^{\mathrm{T}} \left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-1} \boldsymbol{X}\boldsymbol{y}_j^{\mathrm{T}} \left|\boldsymbol{y}_j\right|^2} \\
&= \frac{\boldsymbol{y}_j \boldsymbol{X}^{\mathrm{T}} \left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-1} \boldsymbol{X}\boldsymbol{y}_j^{\mathrm{T}}}{\left|\boldsymbol{y}_j\right|^2} \\
&= \frac{\boldsymbol{y}_j \boldsymbol{X}^{\mathrm{T}} \left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-\frac{1}{2}} \left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-\frac{1}{2}} \boldsymbol{X}\boldsymbol{y}_j^{\mathrm{T}}}{\left|\boldsymbol{y}_j\right|^2} \\
&= \frac{\left|\left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-\frac{1}{2}} \boldsymbol{X}\boldsymbol{y}_j^{\mathrm{T}}\right|^2}{\left|\boldsymbol{y}_j\right|^2}
\end{aligned}
\tag{3.12}
$$

The above equations (3.11 and 3.12) show that the multilinear regression is also optimal in the sense of maximum correlation.

3.3. Phi-Square Transformation

Prior to the applications of multilinear regression, MDS, and PEC analyses, a perceptual stimulus-response confusion matrix (e.g., Table 2.5 in Chapter 2) is transformed into dissimilarity estimates using the phi-square statistics (Iverson et al., 1998). The phi-square coefficient replaces the original confusion data by examining the overall response distribution of a pair of stimuli, thus compensating for response biases and asymmetries in the stimulus-response confusion matrices (Iverson et al., 1998). The resulting dissimilarity matrices represent the dissimilarity structure in visual perceptual space and are then used in hierarchical clustering and MDS analyses. Phi-square transformation is a

normalized version of the chi-square test of equality for the two response distributions (identical distributions result in a distance of 0; non-overlapping distributions result in a distance of 1). The phi-square measure is:

$$PH2(x,y)=\sqrt{\frac{\sum_i \frac{(x_i-\mathrm{E}_{xy}(x_i))^2}{\mathrm{E}_{xy}(x_i)}+\sum_i \frac{(y_i-\mathrm{E}_{xy}(y_i))^2}{\mathrm{E}_{xy}(y_i)}}{\mathrm{N}}} \tag{3.13}$$

where x_i and y_i are the frequencies that phonemes x and y were identified as response category i. N equals the total number of responses to phonemes x and y. $E(x_i)$ and $E(y_i)$ equal the expected frequencies of responses for x_i and y_i, if phonemes x and y are equivalent. For example, $E_{xy}(x_i)$ can be expressed as:

$$\mathrm{E}_{xy}(x_i)=\frac{\sum_i x_i}{\sum_i x_i+\sum_i y_i}\cdot(x_i+y_i) \tag{3.14}$$

For a stimulus-response confusion matrix as follows:

$$\begin{array}{c} \\ x \\ y \\ z \end{array}\begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 3 & 4 & 8 \\ 6 & 7 & 5 \\ 8 & 3 & 4 \end{bmatrix} \end{array} \tag{3.15}$$

The $E_{xy}(x_1)$ is

$$\mathrm{E}_{xy}(x_1)=\frac{\left(\sum x_i\right)(x_1+y_1)}{\sum x_i+\sum y_i}=\frac{(3+4+8)(3+6)}{(3+4+8)+(6+7+5)}=\frac{135}{33} \tag{3.16}$$

Note that the above $E_{xy}(x_1)$ applies only to computing the distance between stimuli

x and *y*. The distance between *x* and *z* is computed using $E_{xz}(x_1)$, which is 4.

Other than phi-square transformation, there is another simple method that examines only the response categories associated with the individual phoneme pairs. For example, a consonant C_1 is perceived as C_2 for a percentage of p_{1_2} and as C_1 for a percentage of p_{1_1}. Consonant C_2 is perceived as C_1 for a percentage of p_{2_1} and as C_2 for a percentage of p_{2_2}. Then the similarity between C_1 and C_2 can be estimated as an average of $(p_{1\_2}+p_{2\_1})/2$. However, this similarity estimate has disadvantages (Iverson et al., 1998). First, $(p_{1\_1}+p_{1\_2})$ and $(p_{2\_1}+p_{2\_2})$ are not equal to 1 and are affected by the possibility of other consonants. For example, if C_1 and C_2 are identical and very distinguishable from other consonants, then $(p_{1\_2}+p_{2\_1})/2$ would be 50%. If a similar condition applies to consonants C_1, C_2, and C_3, then $(p_{1\_2}+p_{2\_1})/2$ is about 33% instead. The numbers are different, but the magnitude of similarity is actually equivalent. The phi-square transformation can correct this problem by comparing the response distributions across all available response categories. For both cases, the phi-square distance between C_1 and C_2 should be near zero, because the two distributions are similar. Therefore, the phi-square coefficient for an individual consonant pair (C_1 and C_2) is then independent of other consonants in the experiment. Second, the phi-square transformation can protect against biases and asymmetries (Iverson et al., 1998). For example, /p/, /b/, and /m/ have similar distributions, and p_{b_b}, p_{b_m}, p_{m_b}, and p_{m_m} are equal to zero. The phi-square distance between /b/ and /m/ would be 1, but the percentage method would yield a result of 0.

3.4. Multidimensional Scaling

	ATL	CHI	DEN	HOU	LAX	MIA	NYC	SFO	SEA	WAS
ATL	0	587	1212	701	1936	604	748	2139	2182	543
CHI	587	0	920	940	1745	1188	713	1858	1737	597
DEN	1212	920	0	879	831	1726	1631	949	1021	1494
HOU	701	940	879	0	1374	968	1420	1645	1891	1220
LAX	1936	1745	831	1374	0	2339	2451	347	959	2300
MIA	604	1188	1726	968	2339	0	1092	2594	2734	923
NYC	748	713	1631	1420	2451	1092	0	2571	2408	205
SFO	2139	1858	949	1645	347	2594	2571	0	678	2442
SEA	2182	1737	1021	1891	959	2734	2408	678	0	2329
WAS	543	597	1494	1220	2300	923	205	2442	2329	0

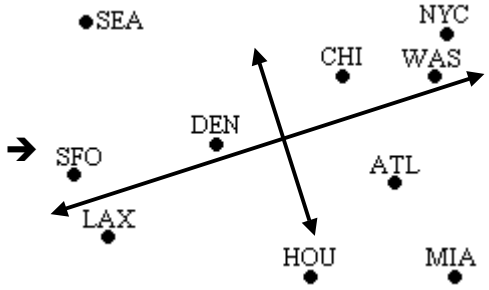


Figure 3.3 Distances between 10 American cities and their 2-D MDS representations [data from (Kruskal and Wish, 1978)].

After the phi-square transformation, 23x23 dissimilarity matrices are obtained. However, from these matrices, it is difficult to examine the relationship between the 23 consonants. If we can put these 23 consonants in a 2-D or 3-D space, then their relationships would be very easy to observe. One of the advantages of geometric based similarity is that it can be examined in detail with MDS (Kruskal and Wish, 1978; Young and Hamer, 1987). The purpose of using MDS is to reduce the number of dimensions while keeping the most important ones so that researchers can explain observed similarities or dissimilarities between the investigated objects. Figure 3.3 illustrates a typical scenario for an MDS analysis. Given a matrix of distances between major US cities from a map, a 2-D MDS can be applied. As a result of the MDS analysis, a 2-D representation of the locations of the cities is obtained. Obviously, it is very clear how the cities are related to each other.

In general, MDS is an optimization problem. In a Euclidean space, every object is assigned a location in the space. An optimization program is then used to move these objects around so as to produce a configuration that best approximates the observed

distances. It actually moves objects around in the space defined by the requested number of dimensions and then checks how well the distances between objects can be reproduced by the new configuration. In more technical terms, it uses a function minimization algorithm that evaluates different configurations with the goal of maximizing the goodness-of-fit. This procedure is illustrated in Figure 3.4.

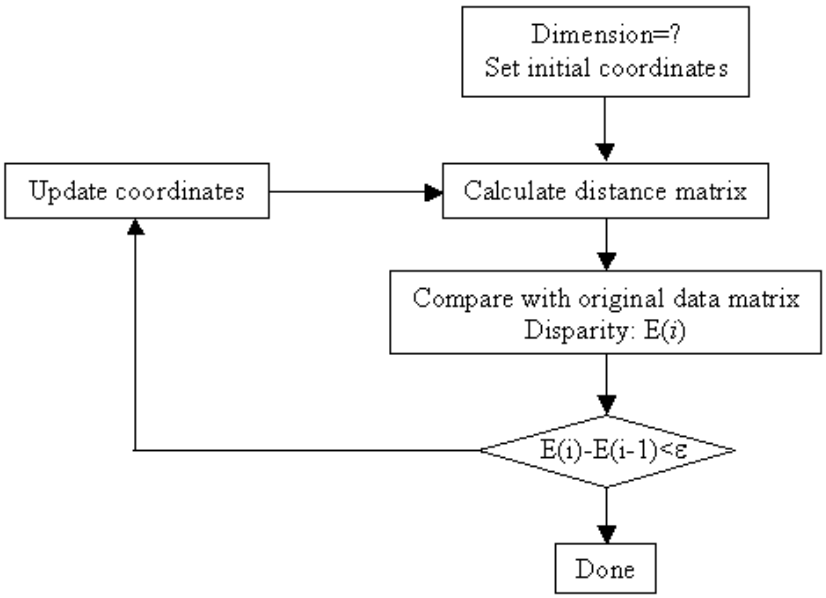


Figure 3.4 Diagram for MDS process.

The goodness-of-fit function is the “stress value”, which is used to evaluate how well (or poorly) a newly produced matrix matches the observed distance matrix. The smaller the stress value, the better the fit. The raw stress value E of a configuration is defined as:

$$\mathrm{E} = \sum\sum \left[d_{ij} - \mathrm{f}\left(\delta_{ij}\right)\right]^2 \tag{3.17}$$

In Equation 3.17, d_{ij} stands for the reproduced distances, given the respective

number of dimensions, and δ_{ij} stands for the input data (i.e., observed distances). The expression f(δ_{ij}) can indicate a transformation (e.g., phi-square transformation) of the observed input data. The stress value in Equation 3.17 is just one of the available measures for the goodness-of-fit of MDS. In this dissertation, the stress value was computed using SPSS' built-in function, and the actual form of the stress equation is unknown.

Another important issue with MDS is to determine the number of dimensions. In general, the more dimensions we use, the better the fit. On the other hand, our goal is to reduce the number of dimensions, that is, to adequately characterize the distance matrix in terms of fewer dimensions. A common way to decide how many dimensions to use is to plot the stress value against different numbers of dimensions to find a place where the stress values begin to decrease smoothly (Cattell, 1966). This procedure can be illustrated in Figure 3.5. From two dimensions to three dimensions, there is a large decrease. From three dimensions to four dimensions, there is also a large decrease. However, after four dimensions, stress values decrease smoothly and appear to level off to the right of the plot. Therefore, a dimension of four can be used in the analysis. Nevertheless, if the dimensionality is too high, then it is difficult to display the data to get a clear interpretation. A compromise can be reached by looking at both the stress value and the variance accounted for. For example, in Figure 3.5, if the variance accounted for (VAF) is acceptable for a dimension of three, and the stress value decrease from three dimensions to four dimensions is not so large. A dimension of three can then be used to approximate the data.

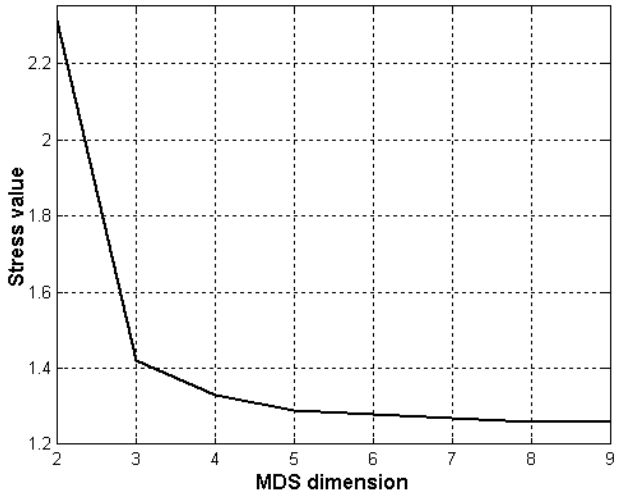


Figure 3.5 Stress value vs. MDS dimension.

Once an MDS analysis is completed, it is important to find and interpret its dimensions. An analytical way of interpreting dimensions is to use multiple regression techniques to seek some weighted combination of the coordinates of a configuration that agree with a meaningful variable as well as possible (described in Kruskal and Wish, 1978). It is extremely valuable to find the relationship between a perceptual MDS space and some physical measures: For example, does the lip height, lip width, lip area, chin, place of articulation, or manner of articulation have a strong effect on perception? Note that for MDS results, the original dimensions have no meaning at all, because if you rotate the axis, the results still fit the observed data. Other than the underlying dimensions, one should also look for clusters of points or particular patterns. For example, /b, m, p/ should be very close to each other.

In this dissertation, MDS is used to yield spatial representations of visual consonant perception from which the Euclidean distances between all possible pairs of consonants in a 3-D space are calculated (the number of distances is 253 for the 23

consonants). At the same time, MDS is also applied to perceptual dissimilarity matrices predicted from different levels of physical measures (the whole face or part of the face). By comparing the spatial structures of visual consonant perception from lipreading and those predicted from physical measures, how the visual consonant perception is predicted from physical measures can be examined. Furthermore, given these visual perceptual distances and those predicted from physical measures, the predictability of visual perception from physical measures could be well assessed as a function of the number of dimensions. For example, when a high dimensional space collapses into a low dimensional one, does the predictability change?

3.5. Hierarchical Clustering Analysis and Phoneme Equivalence Classes

Hierarchical clustering analysis (HCA; Aldenderfer and Blashfield, 1984) is a powerful method for analyzing the structure of data using a multi-stage process. HCA starts with an M-by-M matrix of similarities or dissimilarities and then forms a tree diagram by iteratively clustering two closest entities/clusters or splitting one cluster into two distinguishing entities/clusters. Therefore, this procedure can be developed in two directions: One can start with each entity as a single cluster and then merge two closest entities/clusters iteratively (agglomerative; method used in this dissertation); or one can start with all entities as one large cluster and then split one cluster into two distinguishing clusters iteratively (divisive). One property of HCA is that, the previous step is the basis for the next step, and it cannot be discarded or changed later. In other words, for each

step, the optimization is performed on the results of the previous step. As a result, the procedure is not robust and does not protect against outliers. For example, extraneous data or noise can result in a totally different structure. Misclassification in an early stage can happen, and a check on the final result is highly necessary. Another factor affecting HCA is the possibility of having the same minimum distance between different clusters. In this case, choosing either one as the minimum distance is theoretically correct, but the resulting hierarchical structures would be totally different.

During each iterative step, a distance measure should be used to determine which two entities/clusters are closest and furthest to each other. There are four popular methods, and they are single linkage, complete linkage, average linkage, and the centroid method. Let between-groups distances be $d(P,Q)$, x be one entity from cluster P, and y be one entity from cluster Q. The single linkage distance is $\min_{x,y}\{\|x-y\|\}$, and the complete linkage distance is $\max_{x,y}\{\|x-y\|\}$. Average linkage distance is the mean of $\|x-y\|$. Centroid linkage distance is the distance between the centroids of clusters P and Q.

Agglomerative HCA generates an inverted tree structure in which phonemes join classes based on dissimilarity. At the lowest level of the structure, no phonemes are joined together. At each succeeding level, the most similar pair of classes is joined together. This continues until, at the highest level, all phonemes join a single equivalence class. For a stimulus-response confusion matrix (23x23), after phi-square transform, a dissimilarity matrix (23x23) is obtained, and HCA can be applied. In this dissertation, an average-linkage-between-groups method is used to determine which classes to join at each level in the hierarchy. As the level of the hierarchy goes from low to high, the

average between-class distances become larger, and the number of classes becomes fewer. The procedures are (a) to create 23 clusters where each consonant is a cluster, (b) to compute distances between clusters and to find the smallest distance, and (c) to merge the two closest clusters. Then steps (b) and (c) are run iteratively until a single cluster is formed.

For a given stimulus-response confusion matrix, a phi-square transformation is applied first. Then HCA is applied to the resulting dissimilarity matrix, and consequently, a dendrogram is derived. For visual speech perception, groups of phonemes that are visually highly similar (given a particular level of accuracy) have been referred to as *visemes* (Fisher, 1968). Auer and Bernstein (1997) generalized the notion of visually similar segments to the *phoneme equivalence class* (PEC), which explicitly acknowledges that visual speech similarity varies across talkers and vowel contexts. Based on the dendrogram and original stimulus-response confusion matrix, PECs are chosen by finding the first level in which at least 75% of all the responses are within-class, similar to (Iverson et al., 1998; Walden et al., 1977). As an example, if the phonemes /b/ and /m/ are in the same class, then a /b/ response to a /m/ stimulus should be considered to be a within-class response. Therefore, stimuli with excellent intelligibility would yield many distinct classes of phonemes, because the criterion of 75% should be met at a relatively low level of the hierarchy. On the contrary, less intelligible stimuli would result in fewer PECs, because the criterion of 75% should be more difficult to meet. An example is in Figure 3.6 (PEC analysis on the stimulus-response confusion matrix in Table 2.5). The bold line means the cluster it represents has

75% or above 75% in-class responses; the other line indicates that the cluster it represents has below 75% in-class responses. At the lowest level, only /f/ and /v/ meet the criterion. Then the two closest clusters join, and all clusters are checked whether or not the clusters meet the 75% criterion. This procedure continues until HCA reaches the level that the dashed horizontal line represents. At this level, the six clusters all have 75% or above 75% in-class responses. Therefore, PECs are {f v r}, {w}, {θ ð}, {ʒ dʒ ʃ tʃ s z t d}, {l n k g y h}, and {p b m}. PECs can be used to examine the characteristics of visual perception.

For the visual consonant perception predicted from physical measures, the same procedure can be applied. By this means, PECs can serve as an additional metric to examine how visual consonant perception can be predicted from physical measures. By examining the number of PECs, and how phonemes are grouped, the goodness of the structure can be easily determined.

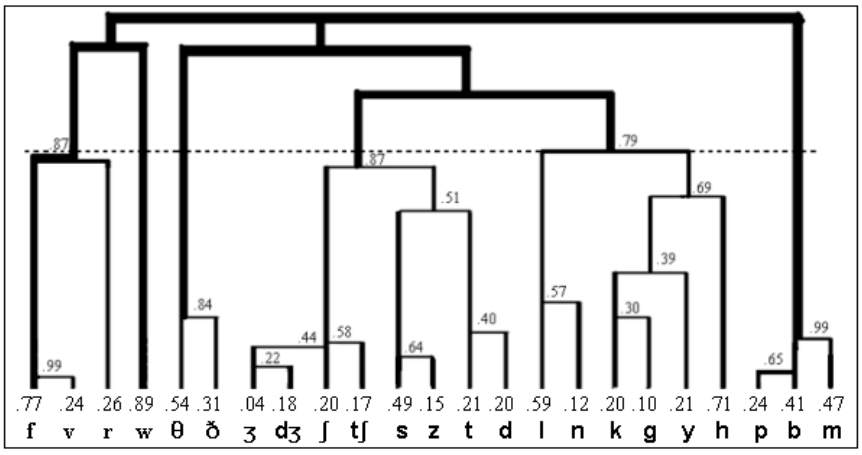


Figure 3.6 PEC analysis of the confusion matrix in Table 2.5.

3.6. Summary

This chapter describes several algorithms used in data analyses: Multilinear regression is used for examining the relationships between data streams; Pearson correlation is used for evaluation of goodness-of-fit. A Jackknife procedure is used to compensate for the lack of sufficient data. Further, it is proven that multilinear regression also meets the maximum correlation criterion (other than minimum root mean square error). For the perceptual similarity analysis, phi-square transformation, MDS, HCA, and PECs are used.

CHAPTER 4. ON THE RELATIONSHIP BETWEEN FACE MOVEMENTS, TONGUE MOVEMENTS, AND SPEECH ACOUSTICS

4.1. Introduction

In this chapter, the correlations between face movements, tongue movements, and speech acoustics were analyzed using multilinear regression for CV syllables (*DcorCV*) and sentences (*DcorSENT*). The correlations were then analyzed for the reduced data sets from the two cases (*DcorCV* and *DcorSENT*). In an effort to improve the predictability of face movements from speech acoustics, the spectral dynamics were modeled, and the new dynamical model enhanced the relationship between face movements and speech acoustics. Part of Chapter 4 is based on a journal paper that appeared in the EURASIP Journal on Applied Signal Processing (November, 2002, pp. 1174-1188) entitled “On the relationship between face movements, tongue movements, and speech acoustics” by Jintao Jiang, Abeer Alwan, Patricia A. Keating, Edward T. Auer, and Lynne E. Bernstein.

4.2. Background

Yehia et al. (1998) used multilinear regression to examine relationships between tongue movements, external face movements (lips, jaw, and cheeks), and speech acoustics for two or three sentences repeated four or five times by a native male talker of American

English and by a native male talker of Japanese. For the English talker, results showed that tongue movements predicted from face movements accounted for 61% of the variance of measured tongue movements (correlation coefficient r = 0.78), while face movements predicted from tongue movements accounted for 83% of the variance of measured face movements (r = 0.91). Furthermore, acoustic line spectral pairs (LSPs; Sugamura and Itakura, 1986) predicted from face movements and tongue movements accounted for 53% and 48% (r = 0.73 and r = 0.69) of the variance in measured LSPs, respectively. Face and tongue movements predicted from the LSPs accounted for 52% and 37% (r = 0.72 and r = 0.61) of the variance in measured face and tongue movements, respectively.

Barker and Berthommier (1999) examined the correlation between face movements and the LSPs of 54 French nonsense words repeated ten times. Each word had the form $V_1CV_2CV_1$ in which V was one of /a, i, u/, and C was one of /b, j, l, r, v, z/. Using multilinear regression, the authors reported that face movements predicted from LSPs and RMS energy accounted for 56% (r = 0.75) of the variance of obtained measurements, while predicted acoustic features from face movements accounted for only 30% (r = 0.55) of the variance.

The databases (*DcorCV* and *DcorSENT*) differed from those in (Barker and Berthommier, 1999; Yehia et al., 1998) in several respects. First, all data streams were recorded simultaneously. Yehia et al. (1998) recorded face movements and tongue movements in different sessions and then used Dynamic Time Warping (DTW) to align them, while Barker and Berthommier (1999) did not record tongue movements. Second,

an optical Qualisys^{TM} system was used to capture facial motion. Note that Yehia et al. (1998) used an OPTOTRAK system to capture facial motion, while Barker and Berthommier (1999) extracted physical measures from video images. Third, the results were analyzed in terms of vowel context, place of articulation, and individual articulatory (EMA or Optical) or acoustic (LSP) channel. Finally, the talkers recorded had different intelligibility ratings as shown by the visual perception scores of normal hearing and hearing-impaired individuals.

For sentences, other than segmenting sentences into smaller segments for correlation analysis (Barbosa and Yehia, 2001; Lee et al., 2001), the dynamical information in speech acoustics can be used to enhance the relationship between face movements and speech acoustics. Section 4.6 introduced a new dynamical model that enhanced the relationship between face movements and speech acoustics on the database *DcorSENT*. Based on the autocorrelations of the speech acoustics and those of the face movements, a causal and a non-causal filter were proposed to approximate dynamical features in the speech signals.

4.3. Analysis of CV Syllables

4.3.1. Consonants: Place and Manner of Articulation

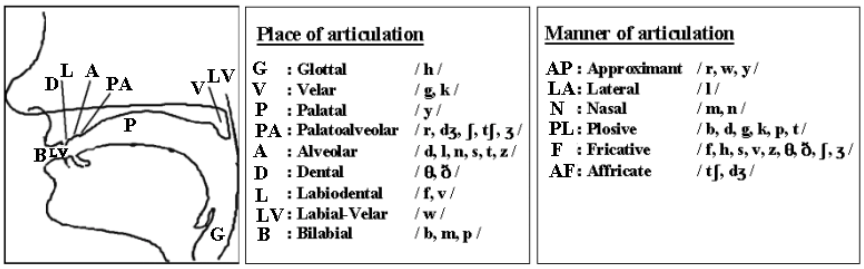


Figure 4.1 Definition of place and manner of articulation for consonants.

One objective in this dissertation was to observe whether linguistic features (place, manner, and voicing) of the consonants affect the correlations between the corresponding data streams. The 23 consonants were grouped in terms of place of articulation (position of maximum constriction), manner of articulation, and voicing (Chomsky and Halle, 1968; Jakobson et al., 1952). The places of articulation are, from back to front, Glottal (G), Velar (V), Palatal (P), Palatoalveolar (PA), Alveolar (A), Dental (D), Labiodental (L), Labial-Velar (LV), and Bilabial (B). The manners of articulation are Approximant (AP), Lateral (LA), Nasal (N), Plosive (PL), Fricative (F), and Affricate (AF) (see Figure 4.1; Ladefoged, 2001).

4.3.2. Syllable-Dependent Predictions

Syllable-dependent correlations (using *DcorCV*), that is, training and prediction

performed for each syllable type separately, are reported first. For each talker, four repetitions of each syllable were analyzed, and a mean correlation coefficient was computed. Table 4.1 summarizes the results averaged across the 69 syllables. The correlations between EMA and OPT data were moderate to high: 0.70-0.88 when predicting OPT from EMA, and 0.74-0.83 when predicting EMA from OPT. Table 4.1 also shows that LSPs were not predicted particularly well from articulatory data, although they were better predicted from EMA data (correlations ranged from 0.54 to 0.61) than from OPT data (correlations ranged from 0.37 to 0.55). However, OPT and EMA data can be recovered reasonably well from LSPE (correlations ranged from 0.74 to 0.82). As for speaker differences, in general, the data from talker F2 resulted in higher correlations than the data from talker F1, and results were similar for talkers M1 and M2.

Table 4.1 Correlation coefficients averaged over all CVs (N=69) and the corresponding standard deviation. The notation X→Y means that X data were used to predict Y data.

	M1	M2	F1	F2	Mean
OPT→EMA	0.83 (0.14)	0.81 (0.15)	0.81 (0.17)	0.74 (0.18)	0.80 (0.16)
OPT→LSP	0.50 (0.16)	0.55 (0.16)	0.37 (0.16)	0.42 (0.13)	0.46 (0.17)
OPT→E	0.75 (0.16)	0.79 (0.17)	0.57 (0.24)	0.70 (0.18)	0.70 (0.21)
EMA→OPT	0.88 (0.12)	0.71 (0.22)	0.70 (0.18)	0.77 (0.19)	0.76 (0.19)
EMA→LSP	0.61 (0.13)	0.61 (0.14)	0.54 (0.15)	0.57 (0.13)	0.59 (0.14)
EMA→E	0.76 (0.18)	0.70 (0.18)	0.65 (0.22)	0.78 (0.14)	0.72 (0.19)
LSPE→OPT	0.82 (0.13)	0.76 (0.17)	0.74 (0.12)	0.79 (0.14)	0.78 (0.14)
LSPE→EMA	0.80 (0.11)	0.79 (0.13)	0.78 (0.15)	0.75 (0.15)	0.78 (0.13)
Mean	0.74 (0.18)	0.71 (0.19)	0.65 (0.22)	0.69 (0.20)	

In order to assess the effects of vowel context, voicing, place, manner, and channels, the results were re-organized and are shown in Figures 4.2 - 4.6.

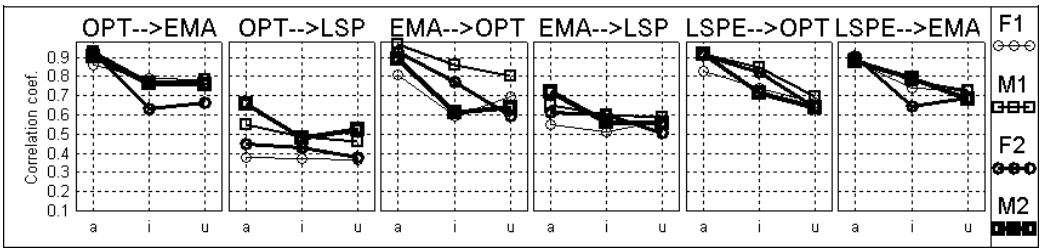


Figure 4.2 Correlation coefficients averaged as a function of vowel context, C/a/, C/i/, or C/u/. Line width represents intelligibility rating level. Circles represent female talkers, and squares represent male talkers.

Figure 4.2 illustrates the results as a function of vowel context, /a, i, u/. It shows that C/a/ syllables were better predicted than C/i/ [$t(23) = 6.2$, $p < 0.0001$] and C/u/ [$t(23) = 9.5$, $p < 0.0001$] syllables for all talkers. Also, several predictions (all but predicting LSP parameters) had near-max correlations, but only for C/a/. [Note that paired T-Tests were used to test the significance of the difference (Sachs, 1984), where *p* refers to significant level, t(N-1) refers to the calculated t-statistic value, and N-1 is the degrees of freedom. In Figure 4.2, there were 24 correlation coefficients for C/a/ (four talkers and six predictions) so that N-1=23.]

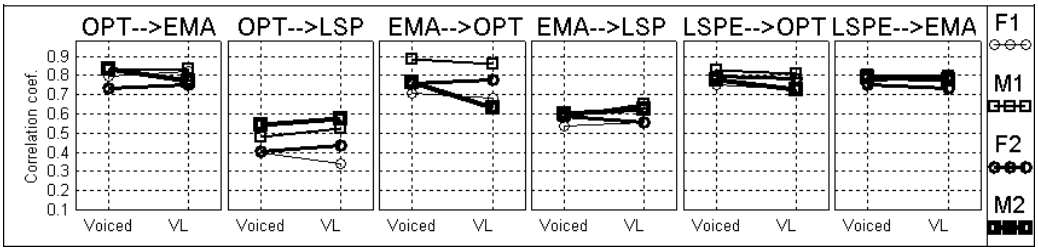


Figure 4.3 Correlation coefficients averaged according to voicing (VL: voiceless).

Figure 4.3 illustrates the results as a function of voicing and shows that voicing had little effect on the correlations.

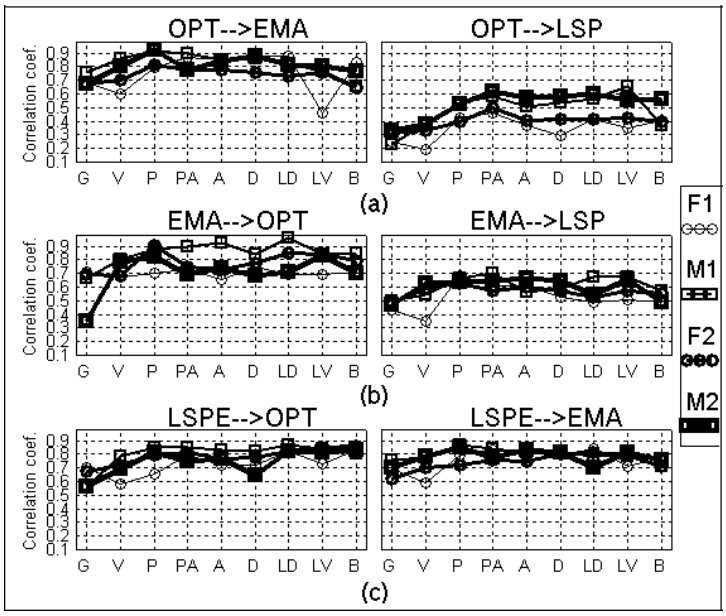


Figure 4.4 Correlation coefficients averaged according to place of articulation. Refer to Figure 4.1 for place of articulation definitions.

Figure 4.4 shows that the correlations for the lingual places of articulation (V, P, PA, A, D, LD, and LV) were in general higher [$t(22)=4.66$, $p<0.0001$] than glottal (G) and bilabial (B) places.

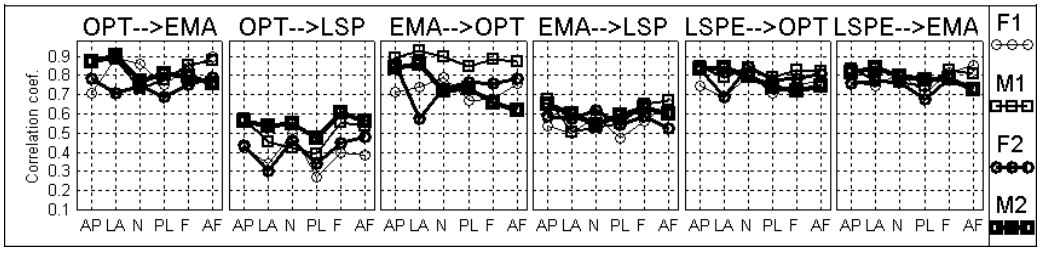


Figure 4.5 Correlation coefficients averaged according to manner of articulation. Refer to Figure 4.1 for manner of articulation definitions.

Figure 4.5 shows the results based on manner of articulation. In general, the prediction of one data stream from another for the plosives was worse than for other

manners of articulation. This trend was stronger between the articulatory data and speech acoustics.

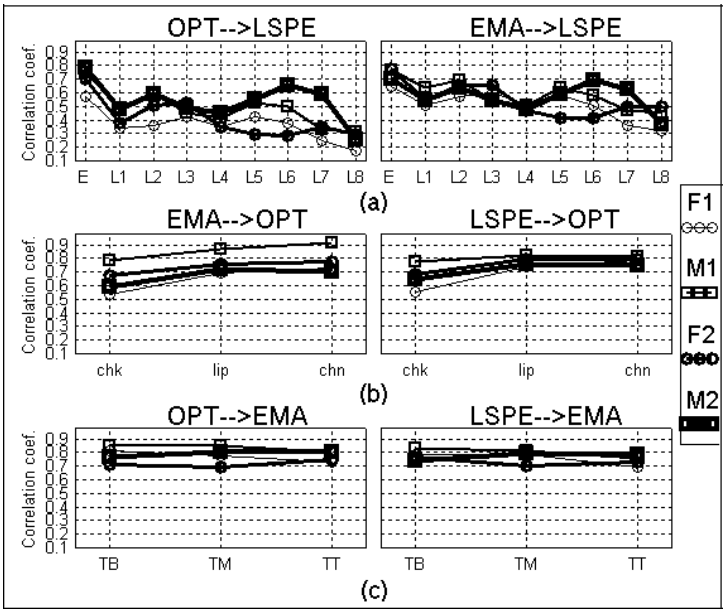


Figure 4.6 Correlation coefficients averaged according to individual channels: (a) LSPE, (b) retro-reflectors, and (c) EMA pellets. Refer to Table 2.4 for definition of individual channels.

Figures 4.6(a-c) illustrate the results based on individual channels. Figure 4.6(a) shows that the RMS energy (E) was the best-predicted acoustic feature from articulatory (OPT and EMA) data, followed by the 2nd LSP pair. Also note that there was a dip around the 4th LSP pair. For talker F1, who had the smallest mouth movements, correlations for RMS energy were much lower than those from the other talkers, but still higher than the LSPs. For the OPT data [Figure 4.6(b)], chin movements were the easiest to predict from speech acoustics or EMA, while cheek movements were the hardest. When predicting EMA data [Figure 4.6(c)], there was not much difference among the

EMA pellets. However, correlations for TB and TM are, in general, slightly higher than for TT.

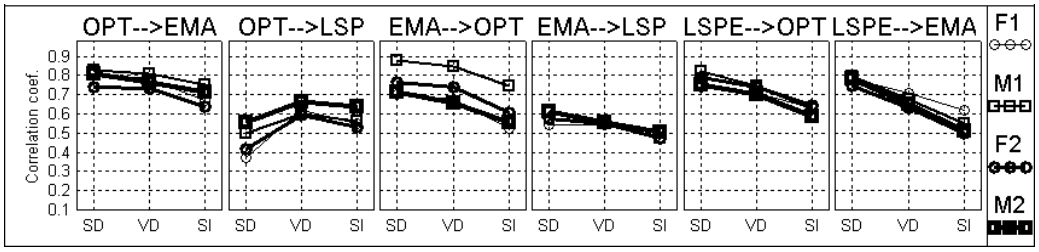


Figure 4.7 Comparison of syllable-dependent (SD), vowel-dependent (VD), and syllable-independent (SI) prediction results.

Syllable-dependent, syllable-independent, and vowel-dependent predictions are compared in Figure 4.7. In general, syllable-dependent prediction yielded the best correlations, followed by vowel-dependent prediction, and then syllable-independent prediction. The only exception occurred when predicting LSPs from OPT data, when the syllable-dependent prediction yielded the lowest correlations.

4.3.3. Discussion

The correlations between internal movements (EMA) and external movements (OPT) were moderate to high, which can be readily explained inasmuch as these movements were produced simultaneously and are physically related (in terms of muscle activities) in the course of producing the speech signal. Yehia et al. (1998) also reported that facial motion is highly predictable from vocal-tract motion. However, Yehia et al. (1998) reported that LSPs were better recovered from OPT than from EMA data. This is not true in the current study for CVs, but the differences might be due to talkers having different

control strategies for CVs than for sentences. For example, sentences and isolated CVs have different stress and co-articulation characteristics.

Note that talker F2, who had higher visual intelligibility than talker F1, produced speech that resulted in higher correlations. However, for the male talkers, intelligibility ratings were not predictive of the correlations. On the other hand, the two male talkers were perceived with different visual intelligibility by 16 normal hearing lipreaders (M1 > M2) and eight deaf lipreaders (M2 > M1).

C/a/ syllables were better predicted than C/i/ and C/u/ syllables for all talkers. This can be explained as an effect of the typically large mouth opening for /a/ and as an effect of co-articulation; articulatory movements are more prominent in the context of /a/. Note that Yehia et al. (1998) also reported that the lowest correlation coefficients were usually associated with the smallest amplitudes of motion. As expected, voicing had little effect on the correlations, because the vocal cords, which vibrate when the consonant is voiced, are not visible. The correlations for the lingual places of articulation were in general higher than glottal and bilabial places. This result can be explained by the fact that, during bilabial production, the maximum constriction is formed at the lips, the tongue shape is not constrained, and therefore one data stream cannot be well predicted from another data stream. Similarly, for /h/, with the maximum constriction at the glottis, the tongue shape is flexible and typically assumes the shape of the following vowel.

In general, the prediction of one data stream from another was worse for the plosives than for other manners of articulation. This result is expected, because plosive production involves silence, a very short burst, and a rapid transition into the following

vowel, which may be difficult to capture. For example, during the silent period, the acoustics contain no information, while the face is moving into position. In this work, the frame rate was 120 Hz, which is not sufficient to capture rapid acoustic formant transitions. In speech coding and recognition, a variable frame rate method is used to deal with this problem by capturing more frames in the transition regions (Zhu and Alwan, 2000).

Figure 4.6(a) shows that the RMS energy (E) and the 2nd LSP pair, which approximately corresponds to the 2nd formant frequency, were better predicted from articulatory (OPT and EMA) data than other LSP pairs as also reported by Yehia et al. (1998). One hypothesis is that the RMS energy is highly related to mouth aperture, and mouth aperture is well represented in both EMA and OPT data. In addition, the 2nd formant has been shown to be related to acoustic intelligibility (Langereis et al., 2000) and lip movements (Grant and Seitz, 2000; Yehia et al., 1998). Grant and Seitz (2000) reported that the lip area functions were correlated best with the F2 envelope and correlated least with the F1 envelope. Again this is presumably because chin movements are directly related to mouth opening, while cheek movements are not directly related to any articulation. Not surprisingly, the larger movements of the tongue body have a greater effect on the acoustics (especially on F1 and F2), than do the smaller tongue tip movements.

Syllable-dependent prediction shows that vowel effects were prominent for all CVs. Hence, if a universal estimator was applied to all 69 CVs, correlations should decrease. This hypothesis was tested, and the results are shown in Figure 4.7. In general,

syllable-independent prediction yielded the best correlations, followed by vowel-dependent prediction. An exception is that for the prediction of LSPs from OPT data, the syllable-dependent prediction yielded the lowest correlations. These results show that there were significant differences between the predictions of the different CV syllables so that syllable-independent prediction gave the worst results. Although vowel-dependent predictions gave lower correlations than syllable-dependent predictions, they were much better than syllable-independent predictions, suggesting that the vowel context effect was significant in the relationship between speech acoustics and articulatory movements. Note that, compared with syllable-independent predictions, vowel-dependent predictions were performed with smaller data sets defined by vowel context.

4.4. Examining the Relationships between Data Streams for Sentences

Sentences (*DcorSENT*) were also analyzed to examine similarity with results obtained from the CV database.

4.4.1. Analysis

For sentence-independent predictions, the 12 utterances (three sentences repeated four times) were divided into four parts, where each part had one repetition of each sentence, and then a Jackknife training and testing procedure was used.

4.4.2. Results

Table 4.2 lists the results for the sentence-independent predictions for the four talkers. Note that talker F1, who had the lowest intelligibility rating based on sentence stimuli, gave the poorest prediction of one data stream from another. In general, the relationship between EMA and OPT data was relatively strong. The predictions of articulatory data from LSPs were better than the predictions of LSPs from articulatory data. LSP data were better predicted from OPT data than from EMA data. OPT data, which includes mouth movements, yielded better prediction of RMS energy than did EMA data.

Table 4.2 Sentence-independent prediction.

	M1	M2	F1	F2
OPT→EMA	0.61	0.68	0.47	0.71
OPT→LSP	0.57	0.61	0.47	0.57
OPT→E	0.71	0.70	0.67	0.63
EMA→OPT	0.65	0.51	0.50	0.61
EMA→LSP	0.36	0.45	0.42	0.49
EMA→E	0.27	0.43	0.45	0.52
LSPE→OPT	0.59	0.67	0.59	0.62
LSPE→EMA	0.54	0.68	0.61	0.68

Figures 4.8(a-c) illustrate predictions for sentences as a function of individual channels. For OPT data predicted from EMA or LSPE, chin movements were best predicted and cheek data were worst predicted, as were found for CVs. For EMA data predicted from OPT data, the tongue tip pellet (TT) was better predicted than tongue back (TB) and tongue middle (TM), which is different from for CVs, suggesting that the tongue tip (TT) was more coupled with face movements during sentence production. For

EMA data predicted from LSP, however, TT was the worst predicted among the three tongue pellets, as was found for CVs. For the acoustic features, the 2nd LSP pair was more easily recovered than other LSP pairs, and unlike for CVs, even better than the RMS energy (E).

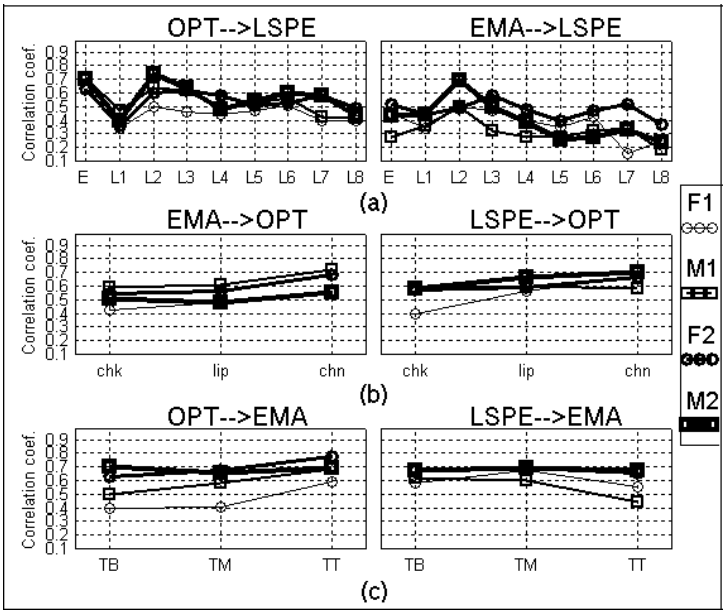


Figure 4.8 Prediction of individual channels for the sentences.

4.4.3. Discussion

As with the CV database, the data from talker F1 gave the poorest predictions of one data stream from another, while the data from talker F2, who had the highest intelligibility rating, did not always give the best predictions. Hence, it is not clear the extent to which the obtained correlations are related to the factors that drove the intelligibility ratings for these talkers. In general, the data from the two males gave better predictions than those from the two females. This may be related to gender or some other effect like talkers'

face sizes. Note that talker M1 had the largest face among the four talkers. As with CVs, the tongue motion could be recovered quite well from facial motion, as also reported by Yehia et al. (1998). Unlike with CVs, LSP data were better predicted from OPT than from EMA data. This is somewhat surprising, because the tongue movements should be more closely, or better, related to speech acoustics than to face movements. However, as discussed by Yehia et al. (1998), this may be due to incomplete measurements of the tongue (sparse data). It may also be due to the fact that the tongue's relationship to speech acoustics is nonlinear. OPT data, which include mouth movements, yielded better prediction of RMS energy than did EMA data. Compared to CV syllables, the predictions of sentences from one data stream to another were much worse than those of the syllables. This is expected, because multilinear regression is more applicable to short segments, where the relationships between two data streams are approximately linear and context information should affect the correlations to some extent.

For EMA data predicted from OPT data, TT was better predicted than TB and TM, which was different from with CVs, suggesting that the tongue tip was more coupled with face movements during sentence production. A possible reason is that, during sentence production, the TT was more related to the constriction and front cavity than were the TB and TM. For EMA data predicted from LSPE, however, TT was the least predicted among the three tongue pellets, as was found for CVs.

The correlations in Table 4.2 were lower than those of the two similar studies (Barker and Berthommier, 1999; Yehia et al., 1998). The differences, however, may result from different databases and different channels considered for analysis. In (Yehia

et al., 1998), four pellets on the tongue, one on the lower gum, one on the upper lip, and one on the lower lip were used in analysis, which should give better prediction of face movements and speech acoustics because more EMA pellets, including several co-registered pellets, were used. Other differences include the fact that the EMA data were filtered at a low frequency (7.5 Hz), audio was recorded at 10 kHz, and 10^{th}-order LSP parameters were used. However, there are several common observations in (Yehia et al., 1998) and this work. For example, the correlations between face and tongue movements were relatively high, articulatory data can be well predicted from speech acoustics, and speech acoustics can be better-predicted from face movements than from tongue movements for sentences.

In (Barker and Berthommier, 1999), nonsense $V_1CV_2CV_1$ phrases were used, and face movements were represented by face configuration parameters from image processing. It is difficult to interpret results about correlations with face movements unless it is known what points on the face being modeled. More specifically, it is difficult to include all the important face points and exclude unrelated points. Therefore, if the previous studies tracked different face points, then of course they would have different results; differences could be also due to talker differences. In this work, results were talker-dependent. This is understandable, given that different talkers have different biomechanics and control strategies.

4.5. Prediction Using Reduced Data Sets

4.5.1. Analysis

In Sections 4.3 and 4.4, all available channels of one data stream were used to estimate all available channels of another data stream. For the EMA data, each channel represents a single pellet (TB, TM, or TT). For the optical data, the retro-reflectors were classified into three groups (lips, chin, and cheeks). In the analyses above, all three EMA pellets and all three optical groups were used. As a result, it is difficult to find out how much each channel contributes to the prediction of another data set. For example, how many EMA pellets were crucial for predicting OPT data? Predictions using EMA and OPT data were re-calculated using only one of the three EMA pellets (TB, TM, or TT) and only one of the three optical sets (cheeks, chin, or lips) for syllable-dependent and sentence-independent predictions.

4.5.2. Results

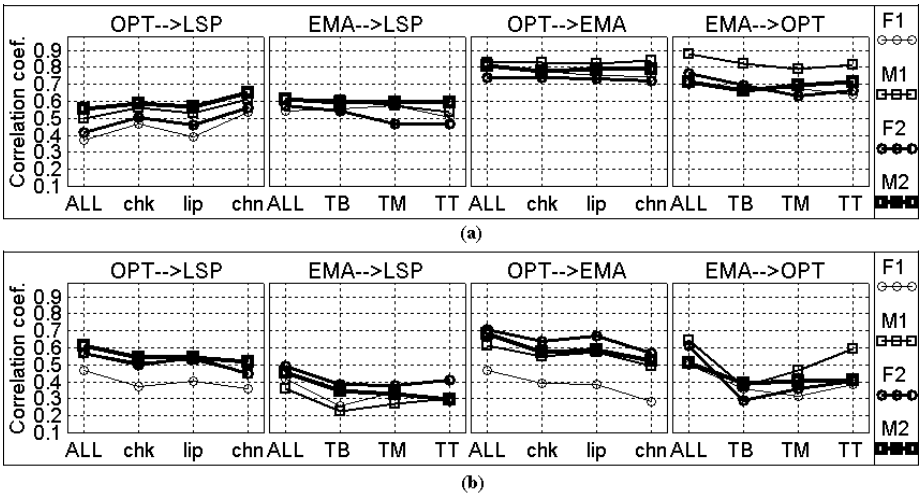


Figure 4.9 Using reduced data sets for (a) syllable-dependent and (b) sentence-independent predictions of one data stream from another.

Figures 4.9(a-b) show the prediction results using reduced EMA and optical sets in syllable-dependent and sentence-independent predictions.

For syllable-dependent prediction, when predicting LSP from OPT, the chin retro-reflectors were the most informative channels, followed by the cheek, and then the lips. Surprisingly, using all OPT channels did not yield better prediction of LSP. When predicting LSP or OPT from EMA, the TB, TM, and TT did not function differently and the all-channel prediction yielded only slightly better correlations. With only one pellet on the tongue, OPT data were still predicted fairly well. When predicting EMA from OPT, different optical channels function similarly and all-channel prediction did not yield higher correlations.

For sentence-independent prediction, when predicting LSP or EMA from OPT, the lip retro-reflectors were more informative channels than cheek and chin retro-

reflectors and the all-channel prediction yielded more information about LSP or EMA. This was different from the predictions of CVs. When predicting OPT or LSP from EMA, the all-channel predictions yielded higher correlations than just one EMA channel. Note that TT provided the most information about OPT data, followed by TM, and then TB.

4.5.3. Discussion

For syllable-dependent predictions, the TB, TM, and TT did not function differently and using all channels yielded slightly better prediction, which implies a high redundancy among the three EMA pellets (Here, redundancy means that when predicting face movements or speech acoustics from tongue movements, combined channels did not give much better predictions than one channel. To examine the absolute level of redundancy between channels, correlation analysis should be applied among EMA pellets and among chin, cheek, and lip retro-reflectors). Such redundancy resulted in a fairly good prediction of OPT data using only one pellet on the tongue. When predicting EMA from OPT, different optical channels function similarly and all-channel prediction did not yield higher correlations, which implies either part of the face contains enough information about EMA. Note that using cheek movements alone can predict tongue movements well. This shows the strong correlations of cheek movements with midsagittal movements as also reported by Yehia et al. (1998).

For sentence-independent prediction, when predicting OPT or LSP from EMA, the all-channel predictions yielded higher correlations than just one EMA channel. This

implies that the difference among the three tongue pellets was stronger for sentences than for CVs. This may be because during CV production, talkers may have attempted to emphasize every sound, which resulted in more constrained tongue movements; this is also presumably because for CVs the big variation (spatially) was in the vowels whose prediction score would be high. When predicting LSP or EMA from OPT, the lip retro-reflectors were more informative channels than cheek and chin retro-reflectors and all channels prediction yielded better prediction of LSP or EMA. This is different from results for CVs. Note that TT provided the most information about OPT data, followed by TM, and then TB.

4.6. Predicting Face Movements from Speech Acoustics Using Spectral Dynamics

4.6.1. Analysis

A dynamical model based on the autocorrelation of speech acoustics and face movements was proposed to model the articulatory dynamics. The objective was to use dynamical features, and at the same time to maintain the low dimensionality of the system. A database of three sentences (*DcorSENT*) was used. Figures 4.10 and 4.11 illustrate the autocorrelation of LSPs and face movements for each talker and for each of the 12 sentences spoken (superimposed lines). The autocorrelations were calculated for each sentence as a function of the autocorrelation lag. These figures verify that each frame is

correlated with its neighboring frames, and after about 15 frames, no more correlations are evident. Recall that the frame length is 24 ms and the frame shift is 8.33 ms. The correlations for face movements decrease slightly more slowly than those for LSPs. A simple curve can be used to approximate the autocorrelations:

$$R(n) = 0.9^{|n|},\ n \in [-\infty, +\infty] \tag{4.1}$$

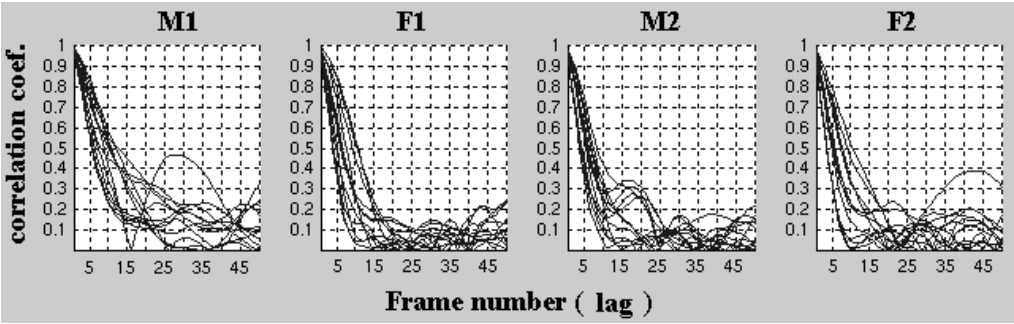


Figure 4.10 Autocorrelations of LSPs.

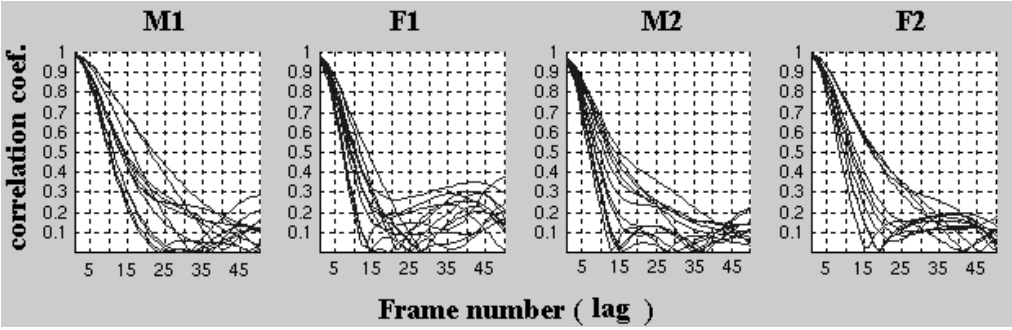


Figure 4.11 Autocorrelations of face movements.

Hence, the following causal (forward: F) filter and non-causal (backward: B) filter can be used to model the effect of speech dynamics on the LSPs:

$$H_{\mathrm{F}}(z) = \frac{1}{1-0.9z^{-1}} \tag{4.2}$$

$$H_{\mathrm{B}}(z) = \frac{1}{1-0.9z} \tag{4.3}$$

$$H_{\mathrm{BF}}(z) = \frac{1}{1-0.9z} + \frac{1}{1-0.9z^{-1}} \tag{4.4}$$

For LSP parameters, by applying the filters in Equations 4.2, 4.3, and 4.4, three new data streams were obtained. They are denoted as FD (forward dynamical), BD (backward dynamical), and BFD (backward-forward dynamical) LSP streams, respectively. The static LSP parameters (same as those in Chapter 2) are denoted as S features. Figure 4.12 shows the frequency response of the BF filter, which is effectively a low-pass filter with a cutoff frequency of less than 10 Hz.

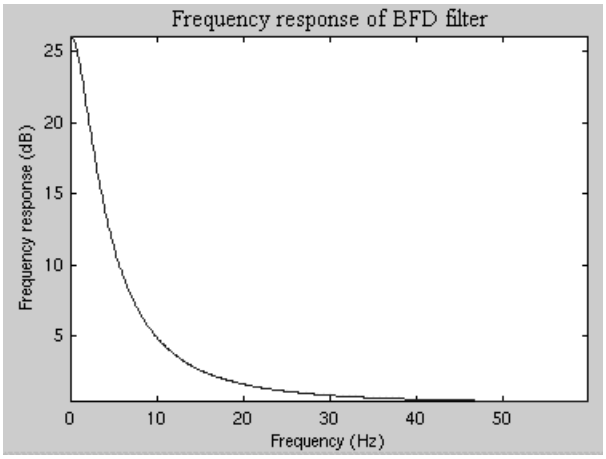


Figure 4.12 Frequency response of the BF filter.

The autocorrelations for face movements and LSPs are also related to the speaking rates of the talkers. Table 4.3 lists the sentence durations for the four talkers.

Talkers M1 and F2 spoke relatively slowly, while talker F1 spoke the fastest. Speaking rates were also reflected in Figures 4.10 and 4.11. The curves for talkers F1 and M1 decline more quickly than those for talkers M1 and F2. Speaking rate did not seem to influence visual intelligibility significantly. For example, talker M2 spoke fast, but he had a higher perceived intelligibility rating than talker M1, as reported by deaf lipreaders.

Table 4.3 Sentence durations (in seconds) for the four talkers.

	Sentence 1	Sentence 2	Sentence 3
M1	5.6 (0.1)	7.6 (0.3)	2.6 (0.3)
F1	4.9 (0.3)	6.4 (0.2)	1.7 (0.1)
M2	4.8 (0.1)	7.1 (0.2)	2.1 (0.1)
F2	6.1 (0.3)	7.8 (0.2)	2.6 (0.3)

4.6.2. Results of Correlation Analysis Using Dynamical Information

In (Barker and Berthommier, 1999; Yehia et al., 1998), training was performed on one data set, and testing was performed on another data set. In this section, training and testing for predicting OPT from LSPE (see Figure 3.2) were performed on the same utterance. Thus, the variability incurred by recording was no longer a factor in the analyses. A correlation coefficient between face movements and those predicted from speech acoustics was calculated for each utterance, and a mean and a standard deviation were computed for the four repetitions of each sentence.

In this section, there were four different LSP streams: static, forward-dynamical, backward-dynamical, and forward-backward dynamical features. These LSP streams alone or in combination can be used to predict optical data. Detailed results are listed in

the Tables 4.4 – 4.11.

Table 4.4 Correlation coefficients obtained using S features.

	Sentence 1	Sentence 2	Sentence 3
M1	0.68 (0.03)	0.69 (0.02)	0.78 (0.04)
F1	0.69 (0.01)	0.64 (0.01)	0.90 (0.02)
M2	0.71 (0.02)	0.67 (0.03)	0.86 (0.04)
F2	0.68 (0.02)	0.67 (0.03)	0.81 (0.05)

Table 4.5 Correlation coefficients obtained using FD features.

	Sentence 1	Sentence 2	Sentence 3
M1	0.77 (0.02)	0.66 (0.02)	0.88 (0.03)
F1	0.74 (0.01)	0.67 (0.02)	0.95 (0.01)
M2	0.75 (0.01)	0.65 (0.02)	0.88 (0.02)
F2	0.79 (0.01)	0.75 (0.02)	0.90 (0.04)

Table 4.6 Correlation coefficients obtained using BD features.

	Sentence 1	Sentence 2	Sentence 3
M1	0.79 (0.05)	0.77 (0.02)	0.90 (0.03)
F1	0.72 (0.03)	0.70 (0.03)	0.95 (0.01)
M2	0.74 (0.03)	0.69 (0.03)	0.93 (0.01)
F2	0.65 (0.03)	0.66 (0.03)	0.93 (0.02)

Table 4.7 Correlation coefficients obtained using BFD features.

	Sentence 1	Sentence 2	Sentence 3
M1	0.85 (0.03)	0.79 (0.02)	0.93 (0.01)
F1	0.81 (0.02)	0.74 (0.03)	0.96 (0.01)
M2	0.83 (0.03)	0.74 (0.03)	0.95 (0.01)
F2	0.81 (0.02)	0.80 (0.03)	0.96 (0.01)

Table 4.8 Correlation coefficients obtained using S+FD features.

	Sentence 1	Sentence 2	Sentence 3
M1	0.82 (0.02)	0.76 (0.03)	0.92 (0.04)
F1	0.82 (0.01)	0.77 (0.01)	0.98 (0.01)
M2	0.84 (0.01)	0.77 (0.02)	0.95 (0.01)
F2	0.82 (0.01)	0.80 (0.02)	0.95 (0.02)

Table 4.9 Correlation coefficients obtained using S+BD features.

	Sentence 1	Sentence 2	Sentence 3
M1	0.84 (0.03)	0.81 (0.02)	0.93 (0.02)
F1	0.82 (0.01)	0.77 (0.01)	0.98 (0.01)
M2	0.85 (0.01)	0.78 (0.02)	0.96 (0.01)
F2	0.80 (0.01)	0.77 (0.02)	0.95 (0.01)

Table 4.10 Correlation coefficients obtained using S+BFD features.

	Sentence 1	Sentence 2	Sentence 3
M1	0.88 (0.02)	0.82 (0.02)	0.97 (0.01)
F1	0.86 (0.01)	0.79 (0.02)	0.99 (0.00)
M2	0.87 (0.03)	0.79 (0.02)	0.97 (0.01)
F2	0.84 (0.02)	0.82 (0.02)	0.98 (0.01)

Table 4.11 Correlation coefficients obtained using S+FD+BD features.

	Sentence 1	Sentence 2	Sentence 3
M1	0.93 (0.01)	0.87 (0.01)	0.98 (0.01)
F1	0.91 (0.02)	0.86 (0.02)	0.99 (0.00)
M2	0.93 (0.01)	0.85 (0.02)	0.99 (0.00)
F2	0.90 (0.02)	0.87 (0.02)	0.99 (0.00)

Table 4.4 lists the correlations obtained using only static LSP features that had 17 channels. For Sentences 1 and 2, the correlations ranged from 0.64 to 0.71, while the correlations ranged from 0.78 to 0.90 for Sentence 3. If training and testing were performed on different utterances, these correlations would be lower. Tables 4.5-4.7 list

the correlations obtained using FD, BD, and BFD LSP features, respectively. Tables 4.8-4.11 list the correlations obtained using static LSP features together with filtered LSP features or their combination.

To compare the overall performance, an average number was computed for all talkers and all sentences. These results are listed in Tables 4.12 and 4.13.

Table 4.12 Average correlation coefficients using individual LSP stream.

	S	FD	BD	BFD
Average	0.73	0.78	0.79	0.85
Improvement	-	7%	8%	16%

Table 4.12 lists correlations obtained using S, FD, BD, or BFD features (LSP) individually. This table shows that BFD filtering was better than using FD or BD filtering alone although they had the same number of channels (17).

Table 4.13 Average correlation coefficients using combined features.

	S+FD	S+BD	S+FD+BD	S+BFD
Average	0.85	0.86	0.92	0.88
Improvement	16%	18%	26%	21%

Table 4.13 lists correlations obtained using static LSP features together with FD, BD, FD+BD, and BFD LSP features, respectively. Both FD and BD features yielded about 17% improvements with additional 17 channels for each case. With an additional 34 channels (FD+BD), the improvement was about 26%.

4.6.3. **Discussion**

Using FD or BD LSP features together with static features can improve the correlations between face movements and those predicted from speech acoustics by about 17%. When using FD and BD together with static features, the improvement was about 26%. Using the backward-forward filter was better than using the backward or forward filter alone. When only using BFD LSP features, the correlations were higher than using static LSP features, although the number of channels is the same. This is because face movements are low-frequency movements, and hence, filtering the LSPs should result in better correlations. The results demonstrate that dynamical information in speech acoustics is important for prediction of face movements. However, dynamical constraints should differ from phoneme to phoneme.

4.7. Summary

In this chapter, relationships among face movements, tongue movements, and acoustic data were quantified through correlation analyses on CVs and sentences using multilinear regression. These correlations were further examined using reduced datasets. Finally, a dynamical acoustic spectral model was proposed to enhance the relationship between face movements and speech acoustics.

Predictions for syllables yielded higher correlations than those for sentences, which suggests a strong context effect in the relationships. The relationships between face movements and speech acoustics varied from vowel to vowel and from consonant to

consonant. This suggests that the relationships are most likely nonlinear, or at least they are locally linear.

For CV syllables, multilinear regression was successful in predicting articulatory movements from speech acoustics, and the correlations between tongue and face movements were high. LSP data were better predicted from EMA than from OPT. Another fact about these correlations was asymmetry of the predictions. In general, articulatory movements were easier to predict from speech acoustics than the reverse. One reason for the asymmetry in estimating acoustic vs. articulatory data can be that acoustic data are more redundant than facial or tongue data. The results reported in this chapter did not show a clear effect of intelligibility of the talker, while the data from the two males gave better predictions than those from the two females. Note that the data from talker M1, who had the largest face among the four talkers, yielded reasonably good predictions. The talker F1, who has lowest intelligibility rating, has the lowest correlations. However, the talker with highest intelligibility rating did not always result in the highest correlations.

Results also showed that the prediction of C/a/ syllables was better than C/i/ and C/u/. Furthermore, vowel-dependent predictions produced much better correlations than syllable-independent predictions. Across different places of articulation, lingual places in general resulted in better predictions of one data stream from another compared to bilabial and glottal places. Among the manners of articulation, plosive consonants yielded lower correlations than others, while voicing had no influence on the correlations. For both syllable-dependent and sentence-independent predictions, prediction of

individual channels was also examined. The chin movements were the best predicted, followed by the lips, and then the cheeks. In regards to the acoustic features, the 2nd LSP pair and RMS energy were better predicted than other LSP pairs. The tongue and face movements are more related to the front cavity of the vocal tract, and thus correlated well with the 2nd formant. The RMS energy can be reliably predicted from face movements. The internal tongue movements cannot predict the RMS energy and LSP well over long periods (sentences), while they were predicted reasonably well for short periods (CVs). When predicted from LSP, TB and TM were better predicted than TT. When predicted from OPT, TT was the best predicted for sentences, while the worst for CVs.

Another question we examined was the magnitude of predictions based on a reduced data set. For both CVs and sentences, a large level of redundancy among TB, TM, and TT and among chin, cheek, and lip movements was found. One implication is that the cheek movements could convey significant information about the tongue and speech acoustics, but these movements were redundant to some degree if chin and lip movements were present. For CVs, using one channel or all channels did not make a difference, except when predicting LSPs from OPT, where the chin movements were the most informative. For sentences, using all channels usually resulted in better prediction; lip movements were the most informative when predicting LSP or EMA; when predicting LSP or OPT, TT was the most informative channel.

The results in Section 4.6 demonstrated that dynamical information in speech acoustics was important for prediction of face movements. However, dynamical constraints appear to be phoneme-dependent.

CHAPTER 5. THE RELATIONSHIP BETWEEN VISUAL SPEECH PERCEPTION AND PHYSICAL MEASURES

5.1. Introduction

In this chapter, the relationship between visual speech perception and physical measures (*DsimCV*; see Chapter 2) was examined. Visual perceptual distances were obtained through the phi-square transform of perceptual stimulus-response confusion matrices. Physical distances were obtained by examining the 3-D face movements of pairs of consonants. Multilinear regression (see Chapter 3) was used to quantify the relationships between visual consonant perception and physical measures. The relationships were further examined using multidimensional scaling (MDS) and phoneme equivalence class (PEC) analyses (see Chapter 3 for a description of these algorithms).

5.2. Background

A descriptive method was used to examine the relationship between physical measures and visual speech perception. The only related study was by Montgomery and Jackson (1983). The authors examined the relationship between visual vowel perception and physical characteristics in an experiment with four female talkers, 10 viewers, and 15 vowels in /hVg/ nonsense words. Because these investigators assumed a traditional description of vowels characterized as relatively steady and of long duration, the authors

used a set of static descriptors to define physical characteristics: lip height (H), lip width (W), lip aperture (AREA), acoustic duration (AD), and visual duration (VD) (see Figure 5.1). However, their study was limited in that it only considered vowels, the results were talker-dependent, few physical measures were used, and the dynamic information was completely absent from the analysis.

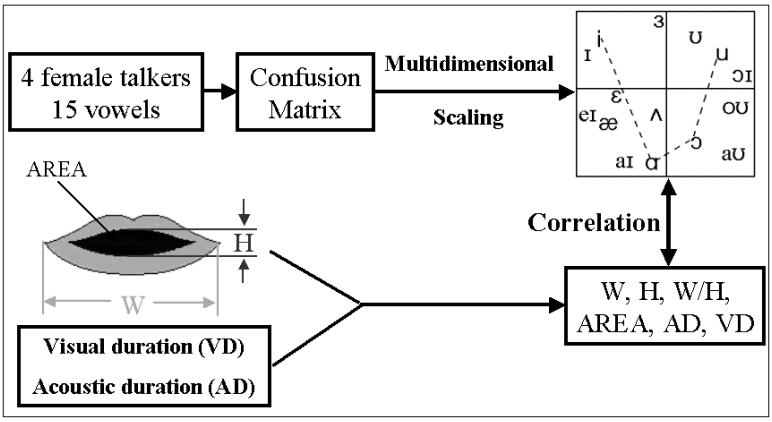


Figure 5.1 A diagram of the Montgomery and Jackson's (1983) study.

The analysis in this chapter extended Montgomery and Jackson's (1983) study in two directions: it focused on consonants and used additional physical measures. Relationships among the perception of spoken nonsense syllables were examined with a number of consonants and vowels and physical attributes of the stimuli (see Figure 5.2). Visual consonant perception was examined using consonant-vowel (CV) syllables. An optical motion capture system was used to capture accurately talkers' face (including chin and cheeks) markers as they moved (see Chapter 2 for details). Physical measures were characterized in terms of the dynamic positions of the face markers (retro-reflectors). A forced-choice perceptual task was used to obtain stimulus-response confusion matrices

(see Chapter 2 for details).

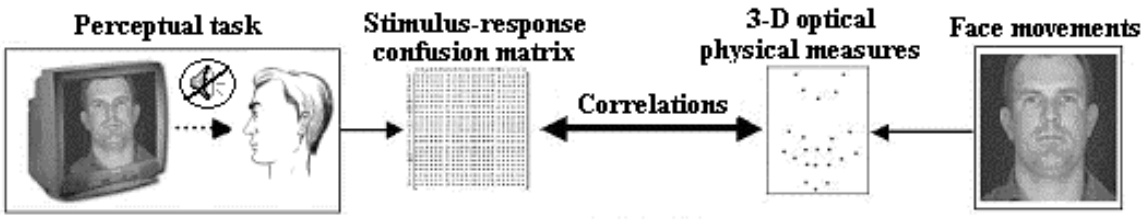


Figure 5.2 The relationship between visual consonant perception and physical measures.

The analyses were organized as follows: (1) The optical recordings were segmented first to extract the consonant segments. (2) Euclidean distances were computed between every consonant pair for each face marker. As a result, physical distance matrices (each row for a face marker coordinate) were obtained. (3) The perceptual stimulus-response confusion matrices were transformed into dissimilarity matrices using the phi-square transformation, and dissimilarity vectors were obtained using the symmetry of dissimilarity matrices. (4) Multilinear regression was then used to map physical distance matrices to perceptual dissimilarity vectors, and Pearson correlations were computed to assess these relationships. These correlations, however, can only indicate the general validity of the relationships. Thus, (5) further analyses were necessary to study in detail these relationships.

During the process of predicting perceptual data from physical measures, perceptual dissimilarity matrices and those predicted from physical measures were obtained. These matrices, however, can be regarded as raw data and provide no obvious relationships between visual consonant perception and physical measures among the consonants. If the consonants can be placed in a 2-D or 3-D space, then their

relationships will be very easy to observe (Eriksen and Hake, 1955; Walden and Montgomery, 1975). Therefore, to examine the details of the relationship, MDS (Kruskal and Wish, 1978; see Chapter 3) was applied to both the perceptual dissimilarity matrices and those predicted from physical measures.

Multidimensional scaling (MDS) provides a reduction in the dimensionality of the data, while keeping the most important ones and thus facilitating understanding relationships. The reduction in dimensionality is particularly important, given that movements of speech articulators (i.e., tongue, lips, jaw, larynx, and velum) are semi-independent. The jaw alone has six-dimensional movements (Vatikiotis-Bateson and Ostry, 1995). In this dissertation, MDS was used to yield spatial representations of visual consonant perceptions and those predicted from physical measures. By comparing their spatial structures, one can examine how the visual consonant perception is predicted from physical measures. Furthermore, MDS can be seen as collapsing high dimensional data into low dimensional data, and thus the correlations can be also examined in different dimensional spaces to assess the stability of these relationships. Nevertheless, MDS provides too many details, and the high-level classification is not straightforward. Therefore, it is desirable to emphasize results at a certain level, while ignoring the internal structure.

To further examine the lipreading results at a higher level, a phoneme equivalence class (PEC) analysis (Auer and Bernstein, 1997; Iverson et al., 1998; see Chapter 3) was used (for both perceptual data and those predicted from physical measures) to examine the grouping of consonants, while ignoring the internal structure. Human speech

perception is often assessed according to the grouping of acoustic sounds or visual gestures. PECs can serve as a high-level (phoneme categorization) metric to examine how visual consonant perception can be predicted from physical measures. By examining the number of PECs, and how phonemes are grouped, the goodness of the structure can be easily seen, and we could also examine how each part of the face contributes to visual speech perception qualitatively.

5.3. Method

5.3.1. Analyses of Perceptual Data

Prior to the application of multilinear regression, MDS, and PEC analyses, each of the perceptual confusion data matrices (see Chapter 2) was transformed using the phi-square statistic (Iverson et al., 1998; see Chapter 3 for a description of these algorithms). The resulting matrices were symmetric. This yielded 253 different stimulus pairs for the 23 consonants. The notation, $\boldsymbol{VD}_{T,V}$, represents visual perceptual distances for talker *T* in vowel context *V*, which was a 253-component, one-dimensional vector. Twenty such vectors for the corresponding 20 raw stimulus-response confusion matrices were computed.

5.3.2. 3-D Optical Signal Analyses

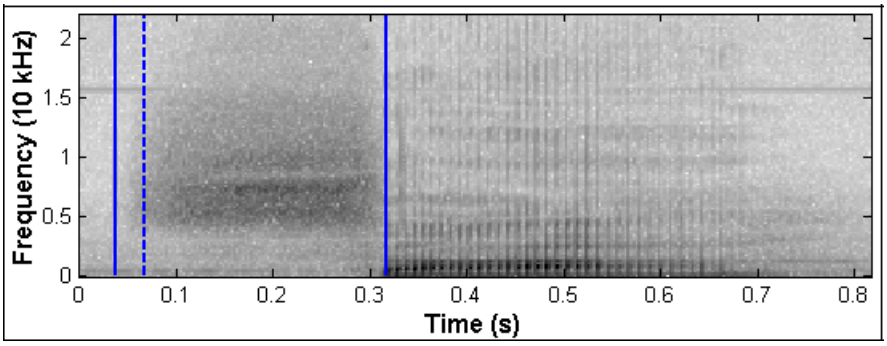


Figure 5.3 Consonant segmentation in a /sa/ syllable.

The sampling frequency for the optical data (*DsimCV*) was 120 Hz. Analyses of 3-D motion were restricted to the initial part of the CV syllables, during the consonant onset, transition, and the initial part of the vowel. In order to have an objectively selected starting point for the analysis of each syllable, the acoustic onset was used because the starting point of the audio signal was a reliable event. However, speech movements often are initiated prior to acoustic signal onsets. Therefore, the beginning point for optical signal analysis was set at 30 ms prior to the acoustic onset (dashed line in Figure 5.3, that is, in the middle of the stop closure) to capture onset movements. Analyses were then applied to the 280-ms segment (between the two solid lines in Figure 5.3). At 120 frames/sec, the 280-ms analysis window was equivalent to 34 optical frames. Figure 5.3 illustrates this with a /sa/ token. Although the 23 consonants differed in duration, the 280-ms segment was sufficient to capture the CV transition and the initial portion of the vowel for all of the tokens. For example, for /ba/, a large interval of /a/ is included, while for /sa/, only a short interval of /a/ is included.

The data for each optical segment were organized into a matrix as follows:

$$\boldsymbol{O}^{T,CV,\beta}_{(1:51,1:34)} = \begin{bmatrix} o_{1,1} & \cdots & o_{1,34} \\ \vdots & \vdots & \vdots \\ o_{51,1} & \cdots & o_{51,34} \end{bmatrix} \tag{5.1}$$

where *T*, *CV*, and β, refer to the talker, CV syllable, and token number, respectively. For example, $\boldsymbol{O}^{\mathrm{M1,ba,1}}_{(1:51,1:34)}$ represents data for the first repetition of syllable /ba/ for talker M1. Each matrix had 34 columns that represented 34 frames, and 51 rows that represented the optical channels (17 retro-reflectors in a 3-D space).

The physical Euclidean distance between a pair of consonants (C_1, C_2), with vowel context *V* for talker *T*, was measured as follows:

$$PO^{T,C_1-C_2,V}_{m} = \sqrt{\Sigma^{2}_{j=1}\left[\Sigma^{34}_{n=1}(O^{T,C_1,V,j}_{m,n} - O^{T,C_2,V,j}_{m,n})^2\right]} \tag{5.2}$$

where *m* is the channel number (1-51), *n* is the frame number, and *j* is the repetition number. If all the Euclidean distances between the 23 consonants in a vowel context for talker *T* are put together, a 51 (row) by 253 (column) matrix can be obtained as $\boldsymbol{PO}^{T,V}$, where the rows represent optical channels, and columns represent consonant pairs. The distance between consonants in a pair was calculated first frame by frame (across channels), and then the mean was taken across the 34 frames. Therefore, dynamical and geometric information was captured to some degree in the distance measures but was weakened by taking the mean across frames.

If either the talker or the vowel factor is collapsed, then two types of physical

distances can be computed, respectively, as:

$$PO_{m,n}^{\mathrm{ALL},V} = \sqrt{\left(PO_{m,n}^{\mathrm{M1},V}\right)^2 + \left(PO_{m,n}^{\mathrm{F1},V}\right)^2 + \left(PO_{m,n}^{\mathrm{M2},V}\right)^2 + \left(PO_{m,n}^{\mathrm{F2},V}\right)^2} \quad (5.3)$$

$$PO_{m,n}^{T,\mathrm{aiu}} = \sqrt{\left(PO_{m,n}^{T,\mathrm{a}}\right)^2 + \left(PO_{m,n}^{T,\mathrm{i}}\right)^2 + \left(PO_{m,n}^{T,\mathrm{u}}\right)^2} \quad (5.4)$$

where *T* can be talker M1, F1, M2, F2, or all talkers. For these distances, three subsets of distances were derived based on the physical distance matrices computed for the lip, cheek (chk), and chin (chn) retro-reflectors, respectively. The matrices were labeled $\boldsymbol{PO}_{\mathrm{lip}}^{T,V}$ (for the lip markers), $\boldsymbol{PO}_{\mathrm{chk}}^{T,V}$ (for the cheeks), and $\boldsymbol{PO}_{\mathrm{chn}}^{T,V}$ (for the chin). In summary, the distance calculations thus resulted in estimates of the dissimilarity of the visual stimuli in terms of the talker, the vowel, and the partition of the talkers' faces into sub-regions (Table 5.1).

These three subsets were not independent of each other, although they did not contain markers in common. Therefore, when the cheek movements were used to predict visual speech perception, some of the lip and chin information was also included due to the correlations between face movements. Nevertheless, the level of correlations can still show the importance of the cheek movements.

Table 5.1 Physical distance matrices as a function of vowel context and talker.

	C/a/	C/i/	C/u/	C/aiu/	C/a/	C/i/	C/u/	C/aiu/
	All retro-reflectors				Lip area			
M1	$\boldsymbol{PO}^{M1,a}$	$\boldsymbol{PO}^{M1,i}$	$\boldsymbol{PO}^{M1,u}$	$\boldsymbol{PO}^{M1,aiu}$	$\boldsymbol{PO}_{lip}^{M1,a}$	$\boldsymbol{PO}_{lip}^{M1,i}$	$\boldsymbol{PO}_{lip}^{M1,u}$	$\boldsymbol{PO}_{lip}^{M1,aiu}$
F1	$\boldsymbol{PO}^{F1,a}$	$\boldsymbol{PO}^{F1,i}$	$\boldsymbol{PO}^{F1,u}$	$\boldsymbol{PO}^{F1,aiu}$	$\boldsymbol{PO}_{lip}^{F1,a}$	$\boldsymbol{PO}_{lip}^{F1,i}$	$\boldsymbol{PO}_{lip}^{F1,u}$	$\boldsymbol{PO}_{lip}^{F1,aiu}$
M2	$\boldsymbol{PO}^{M2,a}$	$\boldsymbol{PO}^{M2,i}$	$\boldsymbol{PO}^{M2,u}$	$\boldsymbol{PO}^{M2,aiu}$	$\boldsymbol{PO}_{lip}^{M2,a}$	$\boldsymbol{PO}_{lip}^{M2,i}$	$\boldsymbol{PO}_{lip}^{M2,u}$	$\boldsymbol{PO}_{lip}^{M2,aiu}$
F2	$\boldsymbol{PO}^{F2,a}$	$\boldsymbol{PO}^{F2,i}$	$\boldsymbol{PO}^{F2,u}$	$\boldsymbol{PO}^{F2,aiu}$	$\boldsymbol{PO}_{lip}^{F2,a}$	$\boldsymbol{PO}_{lip}^{F2,i}$	$\boldsymbol{PO}_{lip}^{F2,u}$	$\boldsymbol{PO}_{lip}^{F2,aiu}$
ALL	$\boldsymbol{PO}^{ALL,a}$	$\boldsymbol{PO}^{ALL,i}$	$\boldsymbol{PO}^{ALL,u}$	$\boldsymbol{PO}^{ALL,aiu}$	$\boldsymbol{PO}_{lip}^{ALL,a}$	$\boldsymbol{PO}_{lip}^{ALL,i}$	$\boldsymbol{PO}_{lip}^{ALL,u}$	$\boldsymbol{PO}_{lip}^{ALL,aiu}$
	Cheek area				Chin area			
M1	$\boldsymbol{PO}_{chk}^{M1,a}$	$\boldsymbol{PO}_{chk}^{M1,i}$	$\boldsymbol{PO}_{chk}^{M1,u}$	$\boldsymbol{PO}_{chk}^{M1,aiu}$	$\boldsymbol{PO}_{chn}^{M1,a}$	$\boldsymbol{PO}_{chn}^{M1,i}$	$\boldsymbol{PO}_{chn}^{M1,u}$	$\boldsymbol{PO}_{chn}^{M1,aiu}$
F1	$\boldsymbol{PO}_{chk}^{F1,a}$	$\boldsymbol{PO}_{chk}^{F1,i}$	$\boldsymbol{PO}_{chk}^{F1,u}$	$\boldsymbol{PO}_{chk}^{F1,aiu}$	$\boldsymbol{PO}_{chn}^{F1,a}$	$\boldsymbol{PO}_{chn}^{F1,i}$	$\boldsymbol{PO}_{chn}^{F1,u}$	$\boldsymbol{PO}_{chn}^{F1,aiu}$
M2	$\boldsymbol{PO}_{chk}^{M2,a}$	$\boldsymbol{PO}_{chk}^{M2,i}$	$\boldsymbol{PO}_{chk}^{M2,u}$	$\boldsymbol{PO}_{chk}^{M2,aiu}$	$\boldsymbol{PO}_{chn}^{M2,a}$	$\boldsymbol{PO}_{chn}^{M2,i}$	$\boldsymbol{PO}_{chn}^{M2,u}$	$\boldsymbol{PO}_{chn}^{M2,aiu}$
F2	$\boldsymbol{PO}_{chk}^{F2,a}$	$\boldsymbol{PO}_{chk}^{F2,i}$	$\boldsymbol{PO}_{chk}^{F2,u}$	$\boldsymbol{PO}_{chk}^{F2,aiu}$	$\boldsymbol{PO}_{chn}^{F2,a}$	$\boldsymbol{PO}_{chn}^{F2,i}$	$\boldsymbol{PO}_{chn}^{F2,u}$	$\boldsymbol{PO}_{chn}^{F2,aiu}$
ALL	$\boldsymbol{PO}_{chk}^{ALL,a}$	$\boldsymbol{PO}_{chk}^{ALL,i}$	$\boldsymbol{PO}_{chk}^{ALL,u}$	$\boldsymbol{PO}_{chk}^{ALL,aiu}$	$\boldsymbol{PO}_{chn}^{ALL,a}$	$\boldsymbol{PO}_{chn}^{ALL,i}$	$\boldsymbol{PO}_{chn}^{ALL,u}$	$\boldsymbol{PO}_{chn}^{ALL,aiu}$

5.3.3. Consonant Classification, Traditional Visemes, and Phoneme Equivalence Classes

The 23 consonants were also grouped in terms of place of articulation (position of maximum constriction), manner of articulation, and voicing (see Figure 4.1). It is well known that visual speech provides significant information about place of articulation and not as much about manner of articulation or voicing (Binnie et al., 1974; Fisher, 1968; Green and Kuhl, 1991; Kricos and Lesner, 1982). In this dissertation, we attempted to determine quantitatively how much information there was in the signal about articulatory features, and how much information was recovered by perceivers.

Traditionally, in the literature on lipreading, consonants have been grouped into

12 visual categories (“visemes”; Kricos and Lesner, 1982): glottal {h}, velars {k g}, palatal {y}, palatoalveolars {ʃ ʒ tʃ dʒ} and {r}, alveolars {l}, {s z}, and {t d n}, dentals {θ ð}, labiodentals {f v}, labial-velar {w}, bilabials {p b m}. These groupings reflect the fact that many phonemic distinctions are acoustic but not visual. These 12 traditional visemes agree well with the different places of articulation and they suggest that, in normal situations, it is difficult for perceivers to accurately label/categorize phonemes within one viseme group. For example, if {p b m} are not visually distinct, then lipreading /b/ as /p/ is not a visual error. Note that in actual lipreading sessions, the categories depend on the talker (Owens and Blazek, 1985), and the number of categories is often fewer than 12 (Binnie et al., 1974; Fisher, 1968; Kricos and Lesner, 1982). However, the 12 visemes can be used to evaluate lipreading performance (Kricos and Lesner, 1982). For example, if /tʃ/ and /dʒ/ are classified into different visemes, then it would suggest that the talker produces speech in an unusual way or that the lipreading participant has a bias.

5.3.4. Analysis Approach

The analysis approach is outlined in Figure 5.4. Physical measures were first organized into matrices, and then physical distances between consonants were computed. Perceptual stimulus-response confusion matrices were transformed into visual perceptual distances using a phi-square transformation. A multilinear regression method was used to predict visual perceptual distances from physical distances.

MDS and PEC analyses were applied to examine the perceptual structure and the

predicted perceptual structure from physical measures. The measures of visual perception were obtained through phi-square transformation and are those described in Section **5.3.1**, the measures of speech production are those described in Section **5.3.2**. For example, in the vowel /a/ context, these measures are referred to as $\boldsymbol{PO}^{T,V}$ (51x253, 17 retro-reflectors on the face), $\boldsymbol{PO}_{\text{lip}}^{T,V}$ (24x253, 8 retro-reflectors on the lips), $\boldsymbol{PO}_{\text{chk}}^{T,V}$ (18x253, 6 retro-reflectors on the cheek), and $\boldsymbol{PO}_{\text{chn}}^{T,V}$ (9x253, 3 retro-reflectors on the chin). The physical measures employed in this chapter included geometry, timing, duration, and, to some degree, dynamic information.

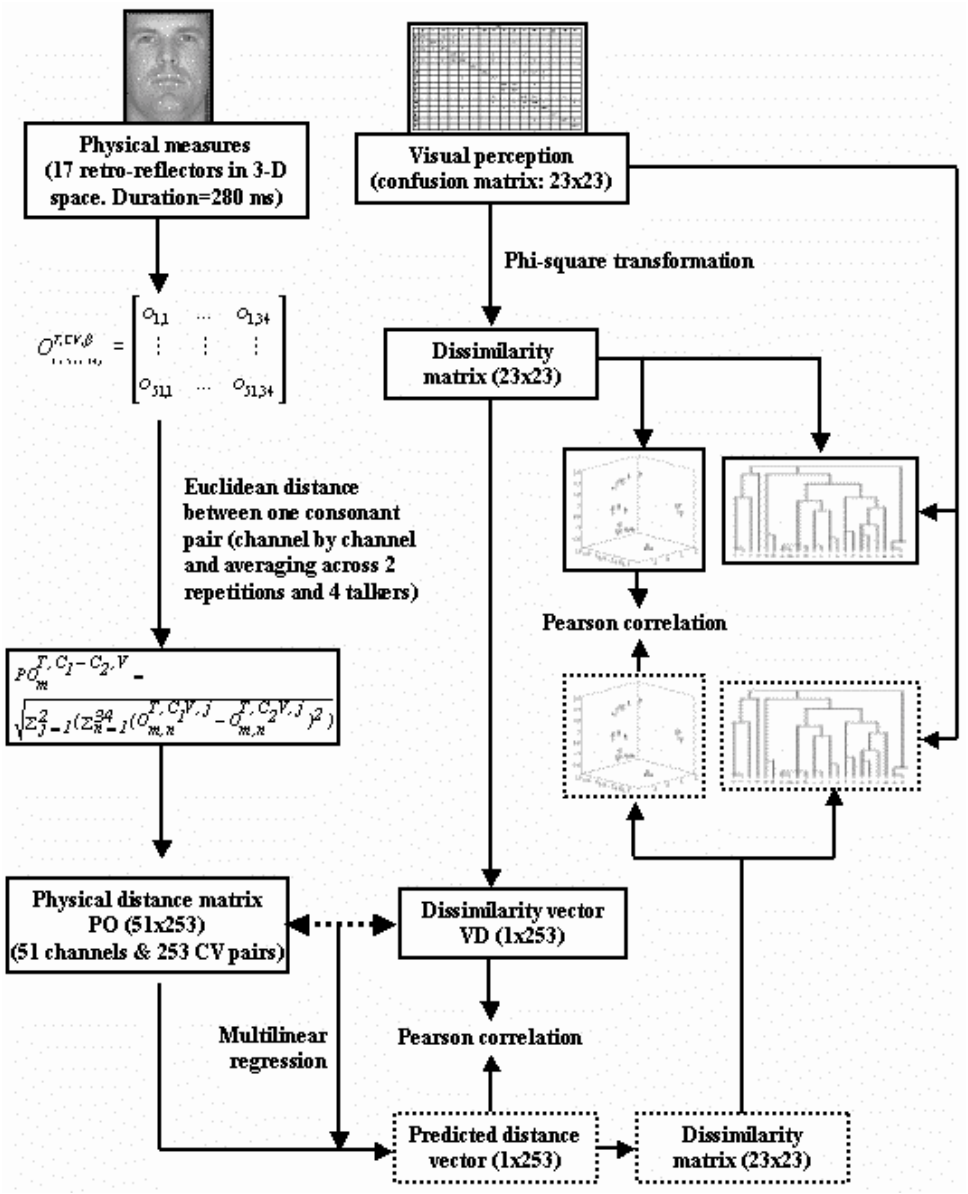


Figure 5.4 A diagram that shows how the analysis of the relationships between visual consonant perception and physical measures was done.

5.4. Results and Discussion

5.4.1. Overall Visual Perception Results

Table 5.2 Lipreading accuracy across talkers and vowel context.

	Vowel context							
	C/a/		C/i/		C/u/		All vowels	
Talker	Mean	Range	Mean	Range	Mean	Range	Mean	Range
M1	0.36	(0.31-0.45)	0.34	(0.30-0.40)	0.30	(0.26-0.41)	0.33	(0.29-0.42)
F1	0.32	(0.28-0.35)	0.32	(0.27-0.36)	0.31	(0.19-0.39)	0.31	(0.26-0.36)
M2	0.39	(0.32-0.47)	0.34	(0.30-0.39)	0.32	(0.20-0.41)	0.35	(0.28-0.42)
F2	0.37	(0.35-0.40)	0.37	(0.35-0.41)	0.32	(0.25-0.39)	0.35	(0.32-0.39)
All talkers	0.36	(0.32-0.42)	0.34	(0.31-0.38)	0.31	(0.23-0.40)	0.34	(0.29-0.40)

Mean percent correct scores for the perceptual identifications of the consonants are reported in Table 5.2. (A 4x3 repeated measures ANOVA was used in a fully nested design to examine effects of talker, vowel, and their interaction.) The talker effect was significant [$F(3, 3) = 23.644$, $p = 0.014$]. Identification was more accurate with talker F2 compared to talker F1 [$F(1, 5) = 29.915$, $p = 0.003$], but no significant difference as obtained between F2 and talkers M1 [$F(1, 5) = 3.240$, $p = 0.132$] and M2 [$F(1, 5) = 0.160$, $p = 0.706$]. Identification accuracy with talker M2 was high than with talkers M1 [$F(1, 5) = 7.731$, $p = 0.039$] and F1 [$F(1, 5) = 17.048$, $p = 0.009$]. Identification accuracy with talker M1 did not significantly differ from with talker F1 [$F(1, 5) = 1.568$, $p = 0.266$]. The vowel effect was also significant [$F(2, 3) = 10.412$, $p = 0.026$]. Accuracy for C/a/ syllables was higher than for C/i/ [$F(1, 5) = 7.665$, $p = 0.039$] and C/u/ syllables [$F(1, 5) = 11.744$, $p = 0.019$], accuracy for C/i/ and C/u/ syllables was not significantly

different [$F(1, 5) = 3.281$, $p = 0.130$]. The average lipreading accuracy was 34% (36% correct for /Ca/, 34% for /Ci/, and 31% for /Cu/ syllables). The talker and vowel interaction effect was "marginally" significant [$F(6, 30) = 2.412$, $p = 0.051$].

Table 5.3 Lipreading accuracy based on 12 Kricos and Lesner (1982) viseme groups across talkers and the vowel context.

	Vowel context			
Talker	C/a/	C/i/	C/u/	All vowels
M1	0.64	0.64	0.52	0.60
F1	0.65	0.61	0.60	0.62
M2	0.72	0.66	0.55	0.64
F2	0.74	0.71	0.64	0.70
All talkers	0.69	0.66	0.58	0.64

Lipreading accuracy was recomputed based on the 12 Kricos and Lesner (1982) visemes and is shown in Table 5.3. Obviously, a higher accuracy is then obtained: 0.69 for C/a/ syllables, 0.66 for C/i/ syllables, and 0.58 for C/u/ syllables.

5.4.2. Predicting Visual Perceptual Measures from Physical Measures

Figures 5.5(a-e) showed the Pearson correlations between visual perceptual distances and physical distances under different conditions. Figure 5.5(a) showed the correlations with unweighted physical measures, that is, each physical channel was equally weighted. Figure 5.5(b) showed the correlations using multilinear regression (each channel was unequally weighted). Figures 5.5(c-e) were similar to Figure 5.5(b) except that the perceptual data and physical measures were represented in a 6-D MDS space, a 3-D MDS space, and a 2-D MDS space, respectively.

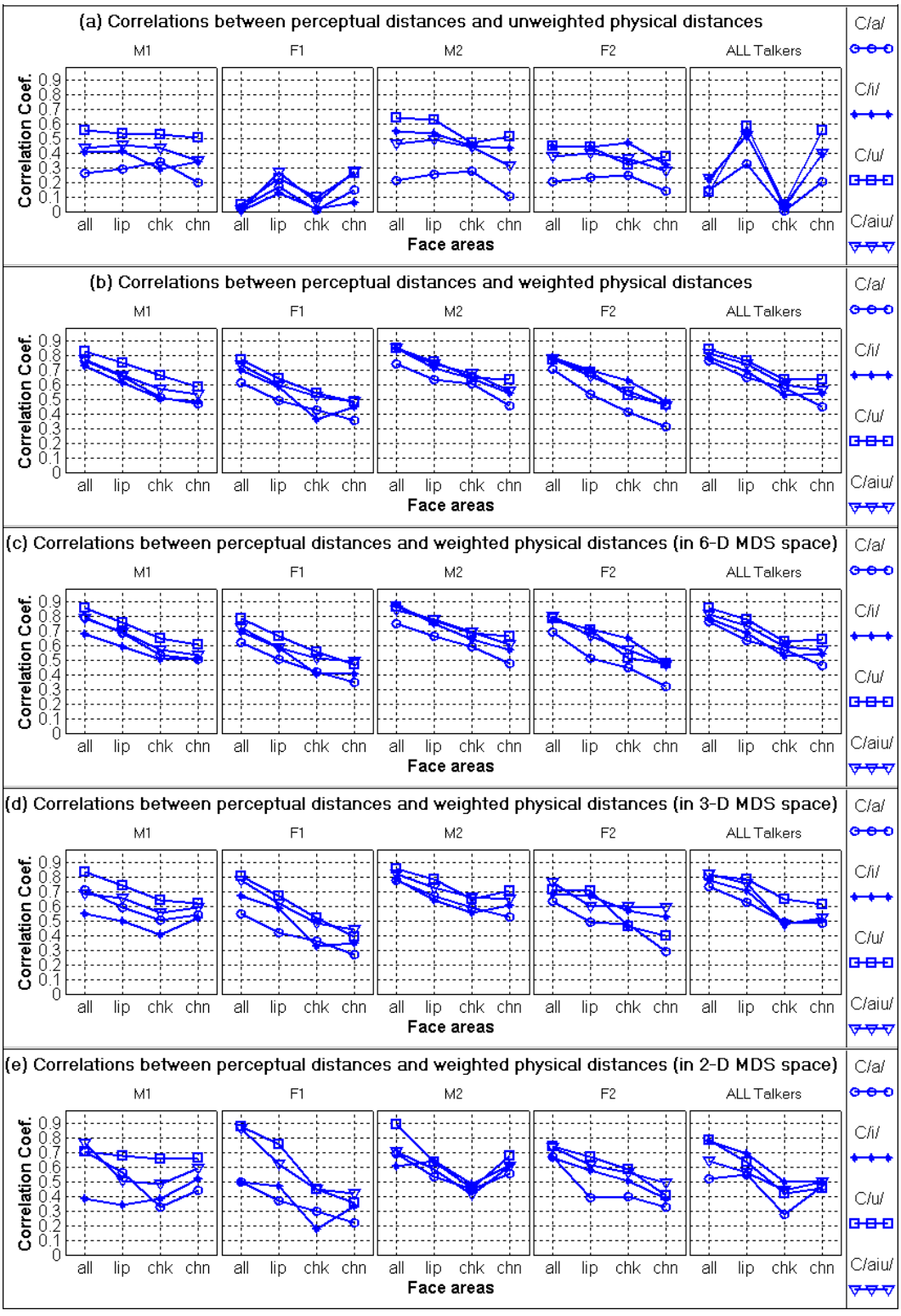


Figure 5.5 Predicting visual perceptual distances from physical distances.

As expected, the correlations in Figure 5.5(a) were not very high (below 0.65; correlations for talker F1 were below 0.30), because each channel was equally weighted. For example, information on the lips should be more important than that on the cheeks. Thus, weighting would result in a significant improvement on the correlations. In Figure 5.5(b), across all contexts and retro-reflectors, the correlations were 0.77 (z = 16.13, p = 0) for M1, 0.74 (z = 15.03, p = 0) for F1, 0.85 (z = 19.86, p = 0) for M2, and 0.78 (z = 16.53, p = 0) for F2. [Note that Fisher's z scores (normally distributed; see Hays, 1994) calculated from these correlations (not normally distributed) were used to assess the significance of the relationships.] Talker M2 had the highest correlations (z = 2.64, p = 0.008 for M1; z = 3.42, p = 0 for F1; z = 2.36, p = 0.018 for F2), and talker F1, with the lowest sentence intelligibility, had the lowest correlations (z = 0.79, p = 0.435 for M1; z = 3.42, p = 0 for M2; z = 1.06, p = 0.289 for F2). The two males tend to produce higher correlations than the two females (z = 2.36, p = 0.018 between M2 and F2; z = 0.79, p = 0.435 between M1 and F1), which could be due to the two males having larger faces and larger movements than the two females and thus leading to larger correlations. Across all talkers, the correlations were 0.77 (z = 16.13, p = 0) for C/a/, 0.78 (z = 16.53, p = 0) for C/i/, 0.84 (z = 19.31, p = 0) for C/u/, and 0.81 (z = 17.82, p = 0) for C/aiu/. Therefore, for the analysis across talkers and vowels, about 66% (r = 0.81 for C/aiu/) of the variance in visual perceptual distances was accounted for by physical distances. C/u/ syllables yielded higher correlations between visual perceptual distances and physical distances than C/a/ (z = 2.25, p = 0.025) and C/i/ syllables (z = 1.97, p = 0.049), even though the movements for /u/ were smaller.

Figure 5.5(b) also showed that both the lips and cheeks were important for visual perception (also true for MDS analyses). Considering the panel with data for all talkers, if one used lip data only, the correlation across all three vowels C/aiu/ was 0.74, which was better than those for cheek data only ($z = 2.88$, $p = 0.004$) and chin data only ($z = 3.39$, $p = 0.001$). Including the cheek and chin data improved the correlation to 0.81 ($z = 1.97$, $p = 0.048$). In addition, cheek data yielded the same level of correlations as chin data ($z = 0.51$, $p = 0.610$), suggesting that various cheek movements helped identify visual gestures, which are important for visual perception.

Figures 5.5(c-e) showed how the correlations varied with different MDS spaces. For the 6-D MDS, the correlations were very similar to those in Figure 5.5(b), suggesting that the 6-D representation of visual perception was sufficient to represent the perceptual confusions. A 3-D MDS can still produce acceptable results because the difference between Figure 5.5(b) and Figure 5.5(d) was not significant. However, a 2-D MDS space was not sufficient to represent the perceptual data and yielded significantly lower correlations in general. Figures 5.5(c-e) also showed a big decrease in correlations from three dimensions to two dimensions. Because 3-D MDS analysis yielded as large a correlation with articulation as can be had from the original data, the 3-D presentations must have included most of the important information.

The talker differences were the same magnitude and type across analyses except for the analysis using unweighted physical distances and 2-D MDS analysis.

5.5. Multidimensional Scaling Analysis

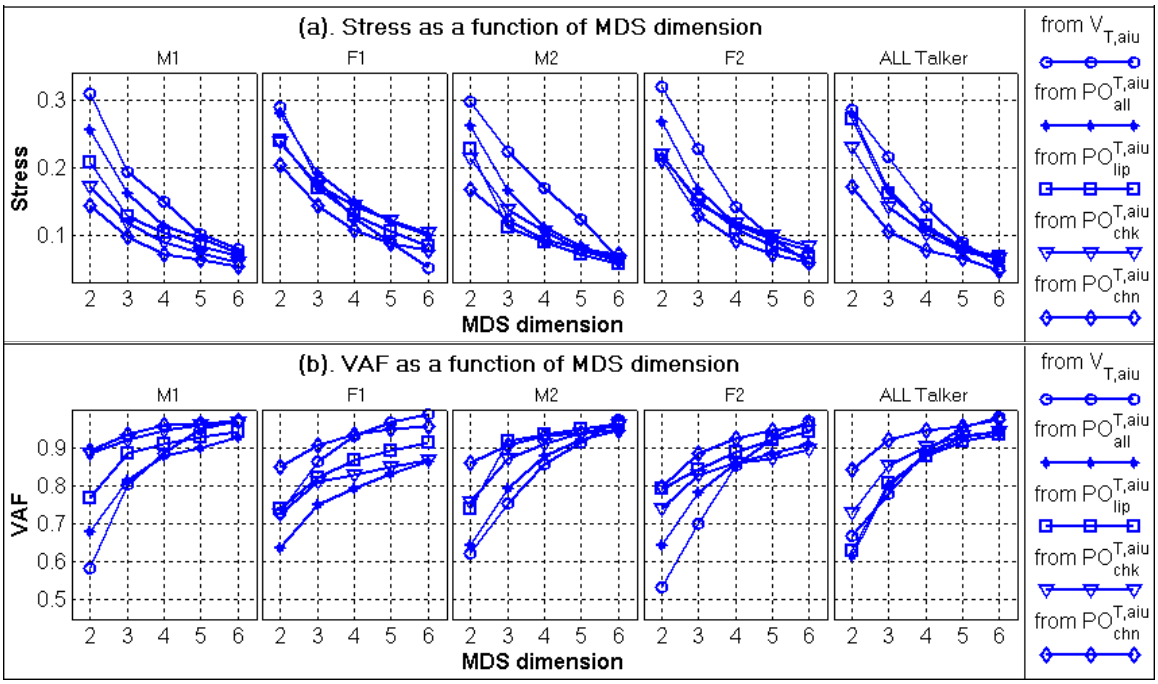


Figure 5.6 (a) Stress and (b) Variance accounted for (VAF) vs. MDS dimension.

MDS of different numbers (2-6) of dimensions were applied to the perceptual dissimilarity matrices and those predicted from physical measures and the resulting stress values and variance accounted for are plotted in Figure 5.6. In all situations, when the number of dimensions increases, the stress value decreases, and the variance accounted for (VAF) increases. Figure 5.6 also showed that the lipreading data had the most complex spatial structures (need more dimensions to represent the data because their stress values are largest), followed by the predicted perceptual data from the whole face, the lips, the cheeks, and then the chin. The number of dimensions for perception was higher than that predicted from the whole face, suggesting that the data recorded for this

dissertation (the 17 retro-reflectors on the face) did not capture all the dimensions used for visual perception. Adding the lips, cheeks, and chin yielded a higher number of dimensions, suggesting that they contained complementary information. However, the dimensionality numbers did not add up linearly, which confirmed that there was redundancy of information (partial independence of chin, cheeks, and lips). Montgomery and Jackson (1983) used 2-D maps to represent the visual vowel perceptual space. The perception of consonants appeared to be somewhat more complicated than the perception of vowels in Montgomery and Jackson's (1983) study. In Figure 5.6(a), there was a large decrease in stress values between two dimensions and three dimensions. Four dimensions would be enough to model visual perception (VAF was above 0.80, and the stress did not decrease much after four). Because a 3-D MDS analysis of data yielded as large a correlation with articulation as can be had from the original data, the 3-D presentations must have included most of the important information. Therefore, we can take a compromise solution that uses only three dimensions. When the number of the MDS dimensions was three, the VAF was above 0.70, which was considered reasonable. (Note that a 2-D MDS was not sufficient to represent consonant confusion.) When the number of dimensions was six, more than 96% of the variance was accounted for.

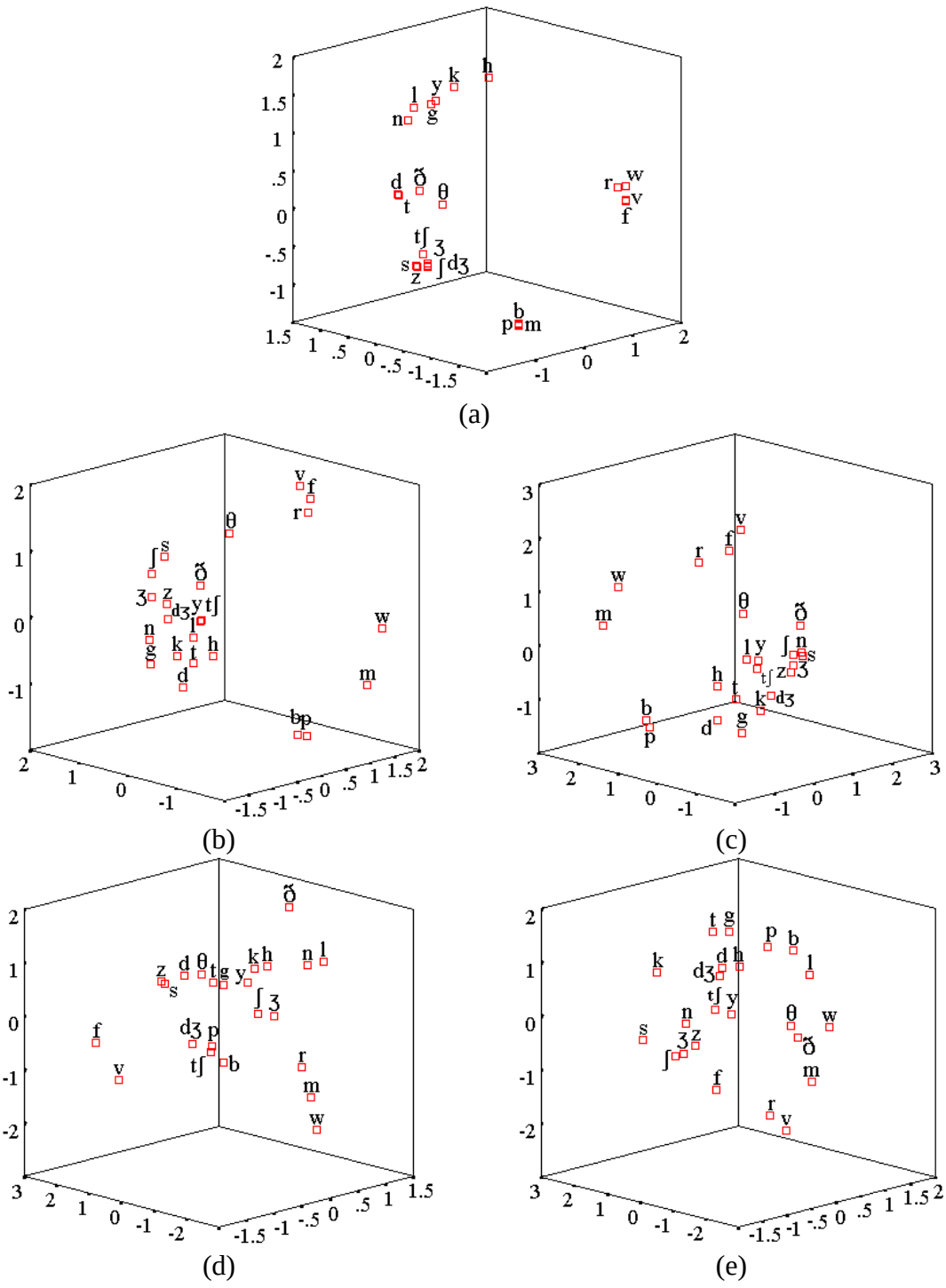


Figure 5.7 3-D MDS analysis of perceptual consonant confusion from (a) lipreading. Predictions from (b) all face retro-reflectors, (c) lip area, (d) cheek area, and (e) chin area.

In Figure 5.7, the 3-D MDS analysis of the confusion matrix $\boldsymbol{C}_{\mathrm{ALL,aiu}}$ and the corresponding matrices predicted from the whole face, lips, cheeks, and chin are displayed (see Appendix C for 3-D MDS plots for other confusion matrices). The figure helps us visualize the relative positions of the 23 consonants in a visual perceptual space. Figure 5.7(a) demonstrated that clusters from lipreading had internal structure (e.g., the members of the group /t, d, s, z, ʃ, ʒ, tʃ, dʒ/ do not have identical coordinates) as well as different distances to other clusters (e.g., /t, d, s, z/ is closer to /θ, ð/ than to /r, f, v/). Such internal structure demonstrated that perceivers can differentiate among consonants, but have trouble categorizing them. For the confusion dissimilarity matrices predicted from the whole face and lips, their 3-D MDS structures [Figures 5.7(b, c)] were similar to those from perceptual data. For the confusion dissimilarity matrices predicted from the cheeks and chin, their 3-D MDS structures [Figures 5.7(d, e)] were far different from those from perceptual data, where the 23 consonants were spread out rather than clustered. Therefore, the cheeks and chin were not sufficient to produce acceptable lipreading results, but combined with the lips, they helped to differentiate these consonants, and thus made them easier to lipread.

The advantage of MDS is that important dimensions in the perceptual space can be examined. For example, we can address questions such as: Do the three most important dimensions in the data have a phonetic interpretation? Is there a dimension corresponding to voicing, or manner of articulation, or place of articulation? Such relationships were examined in as shown Figure 5.8. For each perceptual or predicted dissimilarity matrix, an M-dimensional MDS analysis was applied to obtain a

constellation of the 23 consonants in a visual perceptual space. From this constellation, it was possible to find some underlying dimensions, which were associated with articulatory features or physical measures. Towards this end, multilinear regression was first used to define a best direction, upon which the projections (one-dimensional positions) of the 23 consonants produced the best place of articulation classification. A Pearson correlation was then computed between the 23 consonants' projections (one-dimensional positions) and their places (of articulation). The same procedure was also applied to voicing and manner of articulation. As shown in Figure 5.8, the correlation between perceptual data and place of articulation was very high. Visual perception was modestly related to manner of articulation. These correlations emphasize the fact that visual perception was very good for place, modest for manner, and poor for voicing in pre-vocalic consonants. Also, different numbers of dimensions of MDS yielded different correlation levels. In general, the higher the number of dimensions, the better the correlations. This, again, showed that low dimensional MDS analysis lost information to some degree. But among voicing, manner, and place of articulation, manner depended on the number of dimensions the most, and place the least. Figure 5.8 also showed that the physical measures contained more information about voicing and manner of articulation than was seen in lipreading [Figures 5.8(a, b)]. However, lipreading can uncover more information about place of articulation than these physical measures provide. This can be explained in part by the fact that 17 retro-reflectors were not enough to capture all the information, and physical measures were obtained by averaging across 34 frames. In Figures 5.8(a, b), the combination of the lips, cheeks, and chin did not result in higher

correlation between predicted perceptual data and voicing or manner of articulation, suggesting that the lips, cheeks, and chin may give conflicting cues to voicing and manner of articulation. On the contrary, Figure 5.8(c) showed that place of articulation cues were consistent among the lips, cheeks, and chin. In Table 5.3, talker F2 yielded the highest lipreading accuracy based on the traditional 12 visemes, and this may be due to the fact that, for talker F2, the place of articulation effect was the strongest as shown in Figure 5.8(c).

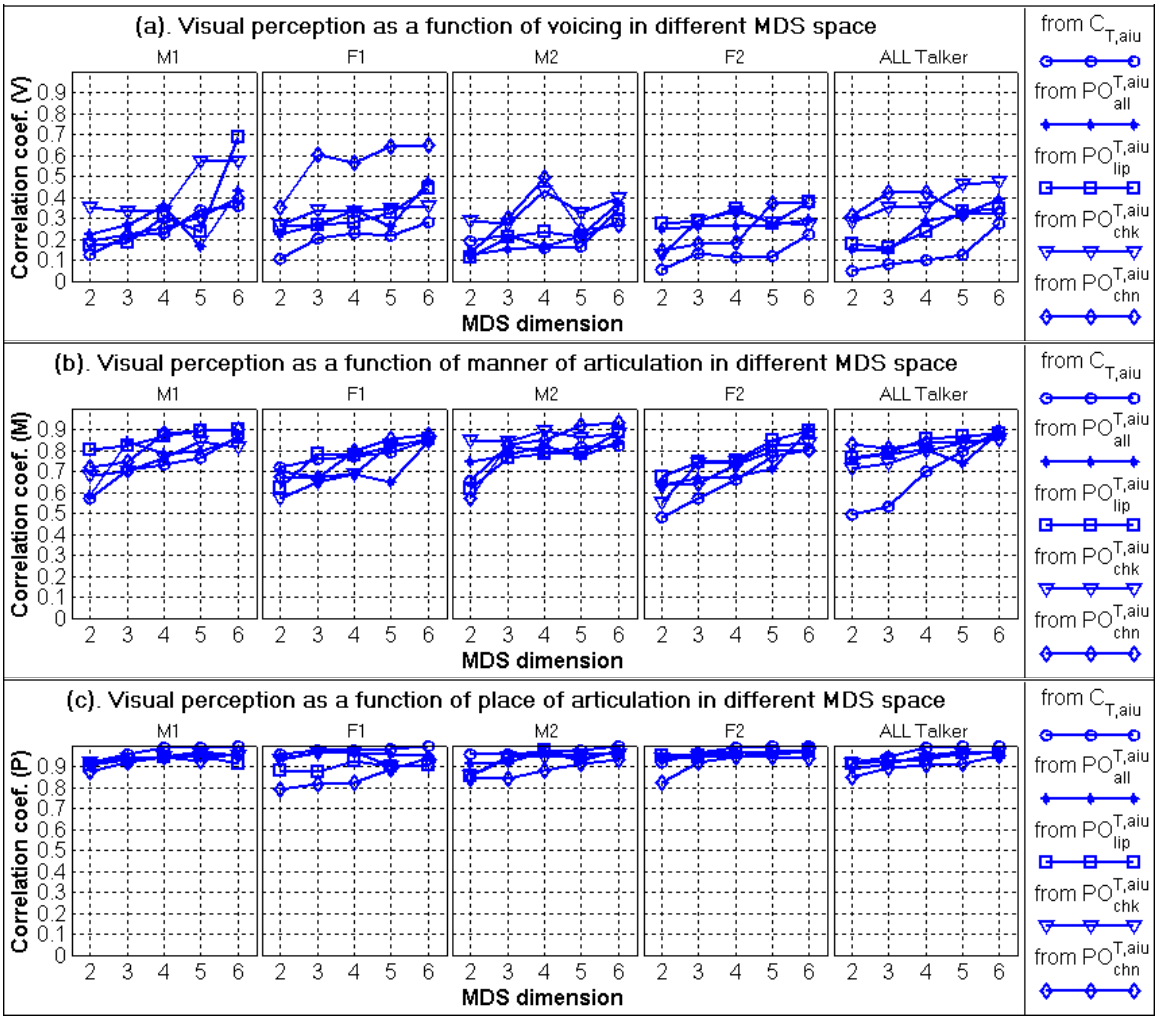


Figure 5.8 Visual perception as a function of voicing, manner, and place of articulation.

5.6. Phoneme Equivalence Class (PEC) Analysis

Table 5.4 PECs across vowel context for different talkers.

Sources	PECs
Talker M1	
Perception	{w} {m p b} {r f v} {θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{ALL}	{w} {m p b} {r f v} {θ ð y l n k g h}{t d s z ʃ ʒ tʃ dʒ}
Predicted^{lip}	{w} {m p b} {f v} {r θ ð y l n k g h} {t d s z ʃ ʒ tʃ dʒ}
Predicted^{chk}	{w} {m p b g} {f v} {θ ð r y l n k h t d} {s z ʃ ʒ tʃ dʒ}
Predicted^{chn}	{r m f v} {w p b θ y l n k g h} {ð t d s z ʃ ʒ tʃ dʒ}
Talker F1	
Perception	{w r} {m p b} {f v} {θ ð} {y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{ALL}	{w r m p b} {f v} {θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{lip}	{w r m p b} {f v} {θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{chk}	{f v} {w r m p b θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{chn}	{w r m p b f v θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Talker M2	
Perception	{w r} {m p b} {f v} {θ ð} {y l n k g h} {t d s z ʃ ʒ tʃ dʒ}
Predicted^{ALL}	{w r} {m p b} {f v θ ð} {y l n t k d g h s z ʃ ʒ tʃ dʒ}
Predicted^{lip}	{w r} {f v} {m p b θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{chk}	{w r m p b f v θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{chn}	{w r m p b f v} {θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Talker F2	
Perception	{w} {m p b} {r f v} {θ ð} {y l n k g h} {t d s z} {ʃ ʒ tʃ dʒ}
Predicted^{ALL}	{w} {m p b} {r f v} {θ ð y l k g h} {n t d s z} {ʃ ʒ tʃ dʒ}
Predicted^{lip}	{w} {m p b} {r f v} {θ ð y l n k g h t d s z} {ʃ ʒ tʃ dʒ}
Predicted^{chk}	{w m p b} {θ ð y l n k g h t d z} {r f v s ʃ ʒ tʃ dʒ}
Predicted^{chn}	{w} {m r f v} {p b θ ð y l n k g h t d} {s z ʃ ʒ tʃ dʒ}
Traditional visemes	*{w} {m p b} {r} {f v} {θ ð} {y} {l} {k g} {h} {t d n} {s z} {ʃ ʒ tʃ dʒ}*

Table 5.5 PECs across talkers for different vowels.

Sources	PECs
C/a/	
Perception	{w} {m p b} {r f v} {θ ð} {y l n k g h} {t d s z ʃ ʒ tʃ dʒ}
Predicted^{ALL}	{w} {m p b} {r f v} {θ ð} {y l n k g h} {t d s z ʃ ʒ tʃ dʒ}
Predicted^{lip}	{w m p b} {r f v} {θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{chk}	{w m p b} {r θ ð y l n k g h t d } {f v s z ʃ ʒ tʃ dʒ}
Predicted^{chn}	{w m p b r f v θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
C/i/	
Perception	{w r} {m p b} {f v} {θ ð} {y l n k g h t d } {s z ʃ ʒ tʃ dʒ}
Predicted^{ALL}	{w} {m p b} {r f v} {θ ð y l n k g h} {t d s z ʃ ʒ tʃ dʒ}
Predicted^{lip}	{w} {m p b t} {r f v} {θ ð y l n k g h d} {s z ʃ ʒ tʃ dʒ}
Predicted^{chk}	{w r m p b f v θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{chn}	{w r m f v}{p b θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
C/u/	
Perception	{w} {m p b} {r f v} {θ ð} {y l n k g h} {t d s z ʃ ʒ tʃ dʒ}
Predicted^{ALL}	{w r m p b} {f v} {θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{lip}	{w r m p b θ ð} {f v} {y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{chk}	{w m p b h} {f v} {r θ ð y l n k g t d s z ʃ ʒ tʃ dʒ}
Predicted^{chn}	{w m p b r f v θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
C/aiu/	
Perception	{w} {m p b} {r f v} {θ ð} {y l n k g h} {t d s z ʃ ʒ tʃ dʒ}
Predicted^{ALL}	{m p b} {w r f v} {θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{lip}	{w m r f v} {p b θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{chk}	{w m p b r f v θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{chn}	{w m r f v} {p b θ ð y l n k g h t d} {s z ʃ ʒ tʃ dʒ}
Traditional visemes	*{w} {m p b} {r} {f v} {θ ð} {y} {l} {k g} {h} {t d n} {s z} {ʃ ʒ tʃ dʒ}*

Tables 5.4 and 5.5 (additional tables are listed in Appendix B) list the PECs for each lipreading result (different talkers and vowel context) and the corresponding results predicted from physical measures. Table 5.4 lists PECs across vowels for each talker. For each talker, the first row lists PECs derived from lipreading results, and the following four rows list PECs predicted from physical measures using all retro-reflectors, lip retro-reflectors, cheek retro-reflectors, and chin retro-reflectors, respectively. Table 5.5 was

similar to Table 5.4 except that PECs were derived across talkers for each vowel. The PECs were also examined against the traditional 12 visemes. In all cases (from lipreading or prediction from physical measures), the palatoalveolars /ʃ, ʒ, tʃ, dʒ/ formed a group, and so did the labiodentals /f, v/.

In the lipreading results, bilabials /m, p, b/ always formed a separate group; dentals /θ, ð/ formed a group and in most cases, formed a separate group; velars /k, g/ formed a group, and so did alveolars /s, z/ and /t, d/ (but not /t, d, n/); labial-velar /w/ formed a separate group or was grouped with /r/; labiodentals /f, v/ formed a separate group or were grouped with palatoalveolar /r/. The resulting PECs, however, were in general agreement with the traditional definition of 12 visemes: velars /k, g/, palatoalveolars /ʃ, ʒ, tʃ, dʒ/, alveolars /s, z/, dentals /θ, ð/, labiodentals /f, v/, and bilabials /p, b, m/ all formed groups except that alveolar /n/ sometimes was not grouped with alveolars /t, d/. For each talker, the main difference lied in the distinction among /θ, ð, y, l, n, k, g, h, t, d, s, z, ʃ, ʒ, tʃ, dʒ/. Talker M1 did not produce much difference among this set of consonants; talker F1 singled /θ, ð/ out from the rest; talker M2 further singled out /y, l, n, k, g/; talker F2 further singled out /t, d, s, z/. Across talkers, PECs for confusion matrices $\boldsymbol{C}_{\mathrm{ALL,a}}$, $\boldsymbol{C}_{\mathrm{ALL,i}}$, and $\boldsymbol{C}_{\mathrm{ALL,u}}$ were in general similar to that for confusion matrix $\boldsymbol{C}_{\mathrm{ALL,aiu}}$. The main difference is that alveolars /t, d/ and alveolars /s, z/ were not grouped for $\boldsymbol{C}_{\mathrm{ALL,i}}$. As shown in Table 5.2, C/u/ syllables were more difficult to lipread than C/a/ and C/i/ syllables, and this was reflected in Table 5.5 because in general fewer PECs were obtained for the confusion matrices $\boldsymbol{C}_{T,\mathrm{u}}$ than for $\boldsymbol{C}_{T,\mathrm{a}}$ and $\boldsymbol{C}_{T,\mathrm{i}}$. Tables

5.4 and 5.5 clearly show that the number of PECs participants can identify was much smaller than 12. For example, talker F2, with the highest sentence intelligibility rating, had the highest number of PECs, seven (four C/a/ and C/u/, and seven for C/i/; see Appendix B). Talker F1, with the lowest sentence intelligibility rating, had only five PECs (five for C/a/ and C/i/, and four for C/u/; see Appendix B). Talker M2, with a medium-high sentence intelligibility rating, had six PECs (six for C/a/, C/i/, and C/u/; see Appendix B). Talker M1, with a medium-high sentence intelligibility rating, had only four PECs (six for C/a/ and C/i/, and five for C/u/; see Appendix B). The number of PECs for each talker was in general agreement with the talker's intelligibility. The visual intelligibility ratings for the four talkers were based on lipreading results from sentences. It appears that visual sentence intelligibility does not solely depend on the number of PECs. It may also depend on many other things (e.g., vowels, prosody, and which consonants are perceived correctly, etc.). For example, the frequencies of /w/, /θ/, and /ð/ are usually high in words so that it is important to correctly identify these consonants (Auer and Bernstein, 1997).

For results predicted from physical measures, Tables 5.4 and 5.5 showed that PECs predicted from all face retro-reflectors agreed the best with the lipreading results, followed by the lip-area-only retro-reflectors. For low intelligibility talkers (M1 and F1; as perceived by deaf lipreaders), there was not much difference between PECs predicted from the whole face vs. just the lip area. For F1, it seems that the talker did not move her cheeks and chin enough, because these areas did not provide much information; for talker M1, the cheek and chin areas provided some information, but it seems redundant (the

talker exaggerated when talking). For high intelligibility talkers (M2 and F2; as perceived by deaf lipreaders), there was a significant difference between PECs predicted from the whole face and those from just the lip area, suggesting that cheek and chin movements provided complementary information to achieve high visual intelligibility. For talker M2, the cheek movements were not informative, while talker F2 produced useful cheek and chin movements. Table 5.5 showed another interesting pattern: For $\boldsymbol{C}_{\mathrm{ALL,a}}$, chin movements were not informative; for $\boldsymbol{C}_{\mathrm{ALL,i}}$, cheek movements were not informative; for $\boldsymbol{C}_{\mathrm{ALL,u}}$, chin movements were not informative; and for $\boldsymbol{C}_{\mathrm{ALL,aiu}}$, cheek movements were not informative. For PECs predicted from the whole face, $\boldsymbol{C}_{\mathrm{ALL,a}}$ yielded the best results, followed by $\boldsymbol{C}_{\mathrm{ALL,i}}$, and $\boldsymbol{C}_{\mathrm{ALL,u}}$ the least. This agreed well with the lipreading performance, but was the opposite result from the correlation analysis, where $\boldsymbol{C}_{\mathrm{ALL,u}}$ yielded the highest correlations between visual perceptual distances and physical distances. It can be concluded that the cheek and chin areas can provide useful information for lipreading, but those areas alone are not sufficient for lipreading.

5.7. General Discussion

Visual perception results showed that the order of the talkers in terms of syllable identification accuracy partially replicated their order in terms of sentence intelligibility reported in Section 2.3.1. These accuracy levels were somewhat similar to those reported by Owens and Blazek (1985; 40% correct for /aCa/, 33% for /iCi/, and 24% for /uCu/)

but lower than those reported by Iverson et al. (1998; 48% for CVs). Note that talker F2's speech had the highest lipreading accuracy based on the 12 viseme categories, and F2 indeed had the highest visual sentence intelligibility rating. This could be because the perceived visual categories for talker F2 tended to fall into the traditionally defined 12 visemes, while they did not for the other talkers. For example, talker F2 may have more prominent facial place of articulation cues.

The relationship between perceptual and optical (physical) measures for visual consonants was examined using a variety of methods. The results indicated that high overall correlations (0.77, 0.74, 0.85, and 0.78 for talkers M1, F1, M2, and F2, respectively) between visual perceptual distances and physical distances could be achieved, depending on the analysis on the data. An MDS analysis of perceptual and physical data showed that information about place of articulation was recovered well by perceivers, while information about voicing or manner of articulation was not. PEC analysis of perceptual and physical data revealed that each talker had his/her own idiosyncratic visual information from different parts of the face, and such differences were somewhat related to the talker's intelligibility. The phoneme-based lipreading accuracies were in general agreement with the talkers' sentence intelligibility ratings. Traditional viseme-based lipreading accuracies showed a slightly different pattern, but they agreed well with results from PEC analysis, supporting the view that, for nonsense syllables, traditionally defined visemes are a better basis for measuring intelligibility as in (Kricos and Lesner, 1982).

The high overall correlations between perceptual and physical measures were

related to visual intelligibility ratings and gender. Talkers with high intelligibility ratings yielded more detailed information, and so did the larger faces of the male talkers (Ostberg et al., 1988). At the same distance from perceivers, a larger face would result in more visual information processed by human brains and thus better lipreading results. The better lipreading results come from stronger relationship between lipreading and visual stimulus. Recall that multilinear regression and Pearson correlation were used to assess the relationship between two data sets. During lipreading, participants had to guess from time to time. If a talker's sentence intelligibility was low (or his face is small in size), then lipreading may be more vulnerable to noise (guessing), and this would result in lower correlations. As for vowel context, C/u/ syllables yielded higher correlations between visual perceptual distances and physical distances than C/a/ and C/i/ syllables even though the movements for /u/ were smaller. This could be due to the fact that the movements for /u/ were stronger in the back and forth direction than those of /a/ and /i/. Of the face movements, the lips (55%) and cheeks (36%) were somewhat more informative for visual perception than the chin (32%). But the chin was also informative. When combining all physical measures, about 66% of the variance in visual consonant confusions was accounted for. As anticipated, combining the three sets of physical measures (from the lips, cheeks, and chin) resulted in higher correlation than using only the lips.

Montgomery and Jackson (1983) used 2-D maps to represent the visual vowel perceptual space. Results in this chapter showed that a 3-D MDS was acceptable, while a 2-D MDS space was not sufficient to represent the perceptual data based on their

correlations with physical measures. From 3-D to 6-D MDS, the correlations did not differ significantly, and one speculation is that lipreading consisted mainly of recovering the information about place of articulation, which did not depend too much on the number of dimensions. As for the physical measures, they were obtained in a 3-D space, and thus the 3-D representations are sufficient.

An MDS analysis showed that clusters from lipreading had an internal structure as well as different distances to other clusters (Auer and Bernstein, 1997; Iverson et al., 1998). For the confusion dissimilarity matrices predicted from the cheeks and chin, their 3-D MDS structures were far different from those from perceptual data, but combined with the lips, they helped to differentiate these consonants and thus make them easier to lipread. The 3-D MDS representations of the perceptual data quantitatively confirmed that lipreading recovers the places of articulation the best. The correlation between perceptual data (represented in MDS space) and manner of articulation decreased rapidly when the number of MDS dimensions decreased, suggesting that manner of articulation by itself was a multidimensional structure in a visual perceptual space. The correlation between perceptual data and place of articulation did not depend too much on the number of dimensions, suggesting that place of articulation can be approximated with only one dimension in the perceptual space. The strong correlation between perceptual data and manner of articulation suggests that perceivers also got manner of articulation information from visual speech, not just place of articulation. This contradicts the general view that manner of articulation is mainly an auditory feature (Binnie et al., 1974; Fisher, 1968; Green and Kuhl, 1991; Ladefoged, 2001; Kricos and Lesner, 1982). It is interesting

to note that the physical measures contained more information about voicing and manner of articulation than was recovered by lipreading, while lipreading can uncover more information about place of articulation than our physical measures provide.

Visual perception experiments revealed that PECs were fewer than 12, and the PECs were a function of vowel context and talkers' sentence intelligibility ratings. As expected, higher sentence intelligibility for a talker corresponded to more PECs in the perception of that talker's CV stimuli (with fewer phonemes in each PEC), and lipreading results showed a strong effect of place of articulation. Across talkers, the PECs were in general similar for visual perceptual distance vectors $\boldsymbol{VD}_{\mathrm{ALL,a}}$, $\boldsymbol{VD}_{\mathrm{ALL,i}}$, and $\boldsymbol{VD}_{\mathrm{ALL,u}}$. For results predicted from physical measures, PECs obtained from all face retro-reflectors agreed the best with the lipreading results, followed by those obtained only from the lip area. However, the cheek and chin areas were also contributing significantly to lipreading. PEC analysis further showed that, for low intelligibility talkers (M1 and F1; as perceived by deaf lipreaders), the lip area contributed almost all the information, and the cheeks and chin did not move much (talker F1) or their movements were redundant with the lips (talker M1). On the contrary, high intelligibility talkers (M2 and F2; as perceived by deaf lipreaders) exhibited complementary cheek and chin movements to achieve high visual intelligibility. For the intelligibility testing on the four talkers, results from deaf lipreaders indicated that talker M2 was more intelligible than talker M1, while results from hearing lipreaders indicated that talker M1 was more intelligible than talker M2. The reason may be that deaf lipreaders look at more than the lips for phonemic distinctions, while hearing subjects used the cheek and chin movements only to confirm

what they saw around the lips. PEC analysis also showed that when producing C/a/ or C/u/, the chin movements were low dimensional while C/i/ syllables had low dimensional cheek movements.

5.8. Summary

In this chapter, the relationship between visual speech perception and physical measures was examined. The correlations between visual perceptual distances and physical distances were related to visual intelligibility ratings, vowel context, areas on the face, and gender. In general, about 66% of the variance in visual consonant confusions was accounted for by physical measures. Of the face movements, the lips (55%) and cheeks (36%) were more informative for visual perception, than the chin (32%). An MDS analysis showed that the physical measures contained more information about voicing and manner of articulation than perceptual data, while lipreading can uncover more information about place of articulation than physical measures. Visual perception experiments revealed that PECs were related to vowel context and talkers' sentence intelligibility ratings. The contribution of different parts of the face to lipreading differed from talker to talker. Interestingly, PECs obtained from all face retro-reflectors agreed the best with the lipreading results, followed by those obtained only from the lip area.

CHAPTER 6. EXAMINING THE CORRELATIONS BETWEEN FACE MOVEMENTS AND SPEECH ACOUSTICS USING MUTUAL INFORMATION FACES

6.1. Introduction

In the summer of 2003, the author had an internship at IBM T.J. Watson Center and had the opportunity to look at the correlation between face movements and speech acoustics from a different perspective. For the correlation analyses in Chapter 4, the face movements were represented by the 3-D trajectories of a few face markers. In this chapter, the correlation between face movements and speech acoustics was examined using a database of recorded videos and a mutual-information-face technique. A mutual information face is a plot/map of the mutual information between audio and video pixels. For comparison, the mutual-information algorithm was also applied to the *DcorSENT* (see Chapter 2). The results from these experiments could help understand the findings in Chapter 4. Note that the videos were prepared at IBM (Potamianos and Neti, 2003) and the mutual-information-faces algorithm was implemented by Nock et al. (2002, 2003) also at IBM. Any opinions or conclusions expressed in this chapter do not necessarily reflect the views of IBM Audio-Visual Speech Technologies group.

6.2. Background

In Chapter 4, face movements were represented using the 3-D trajectories of a few face markers (using the Qualisys^{TM} motion capture system). Such data can be used to explore the basic relationship between face movements and speech acoustics, while the distracting factors are ignored. However, recording face movements using motion caption systems is not feasible for real applications. Instead, researchers often need to process video archives, where faces are in 2-D and they may have different poses. Therefore, the correlation between face movements and speech acoustics was re-examined using a video database recorded in an office environment (Potamianos and Neti, 2003). This correlation can be used for such tasks as assessing lip synchronization quality in movie editing and computer animation, and for finding who is talking in a video (Nock et al., 2002, 2003). For comparison, the mutual-information algorithm was also applied to the audio-visual speech database *DcorSENT* (see Chapter 2).

6.3. Method

6.3.1. Database

The database (named "OFF") was described in detail by Potamianos and Neti (2003). The database was recorded in the participants' offices at IBM without the use of a teleprompter. Therefore, lighting, background, and head-pose varied greatly. The videos were captured using a laptop-based audio-visual data collection system. The built-in

laptop microphone was used to record wideband audio and an inexpensive web-cam was used to record uncompressed video through the USB 2.0 interface. The videos were recorded at 30 Hz and at a 320x280 pixel size. The original database was very large, and for the analysis in this chapter, only three video clips (different individuals) were chosen (see Figures 6.1-6.3).

Figure 6.1 Video clip 1: “nine seven nine nine one eight six two eight zero”.

Figure 6.2 Video clip 2: “Six one six nine three five eight”.

Figure 6.3 Video clip 3: “Two nine six four one one nine”.

The mutual-information algorithm was also applied to the database *DcorSENT*, which is described in Chapter 2. Note that only Sentence 3 (spoken by four talkers with four repetitions) was used for the mutual-information analysis.

The main difference between these two types of databases is that for the IBM video database, the temporary intensity changes for each pixel represent how each pixel interacts with its neighboring pixels. These changes can reflect how face muscles are moving. However, the temporary intensity changes for one pixel do not reflect the movements of one specific point on the face, and they depend on the lighting and intensity of neighboring pixels. For example, if one pixel in the image has the same intensity as its neighboring pixels, then the intensity for this pixel does not change when the talker is talking. On the contrary, in the database *DcorSENT*, temporary changes for each retro-reflector represent the movements of a specific point on the face. Therefore, it does not depend on the lighting and other factors.

The video database can be of higher resolution and is easy to record. The *DcorSENT* has robust movements, but its resolution and availability are limited.

6.3.2. Mutual Information Faces

The mutual information $I(A,V)$ of audio (A) and video (V) indicates how much information is shared between A and V (or how much information is learned about A when V is given, and vice versa). Suppose that there are two flashing lights: One is controlled by sound, and the other one flashes randomly. Then there is a large amount of mutual information between the sound and the controlled light. The mutual-information-faces algorithm was described in (Nock et al., 2002). This implementation differs from that used by other people (Fisher and Darrell, 2002; Hershey and Movellan, 1999; Slaney and Covell, 2001). In this algorithm, the feature vectors derived from audio (A) and video (V) are considered as being samples from a multivariate Gaussian probability distribution $p(A,V)$. Gaussian assumptions are also made for $P(A)$ and $P(V)$. Then the mutual information $I(A, V)$ between the random variable A and V can be obtained as:

$$I(A,V) = \frac{1}{2}\log\frac{|\Sigma_A| \cdot |\Sigma_V|}{|\Sigma_{A,V}|} \tag{6.1}$$

where Σ_A, Σ_V, and $\Sigma_{A,V}$ denote empirical estimates of covariance matrices for A, V, and joint A-V distributions, respectively. To produce mutual information faces, 24 mel-frequency cepstral coefficients were extracted from the audio signal, at a rate of 100 Hz. Videos were linearly interpolated to give a 100 Hz frame rate, which matches the audio processing. Then mutual information was computed between the sequence of acoustic feature vectors and the sequence of pixels at each location. Video clips 1, 2, and 3 have 410, 410, and 320 frames, respectively.

For the sentences in *DcorSENT*, the acoustic feature vector was a 17-dimensional vector (16 LSPs plus RMS energy) at a rate of 120 Hz. The visual feature vector consisted of the 3-D coordinates of each retro-reflector at a rate of 120 Hz. Sentence durations are listed in Table 4.3.

In mutual-information plots, pixel brightness was used to indicate whether the mutual information is high (white) or low (black). In other words, a bright pixel in a mutual-information plot indicated more shared information between the audio and this video pixel than a dark pixel does.

6.3.3. Results

The corresponding mutual-information faces for the three video clips (Figures 6.1-6.3) are shown in Figures 6.4-6.6, respectively. The mutual information faces for database *DcorSENT* are shown in Figure 6.7, where a mutual information face was generated for each utterance. Note that in Figure 6.7, the lower lip center was missing for talker F2 due to dropouts in the 3-D motion reconstruction.

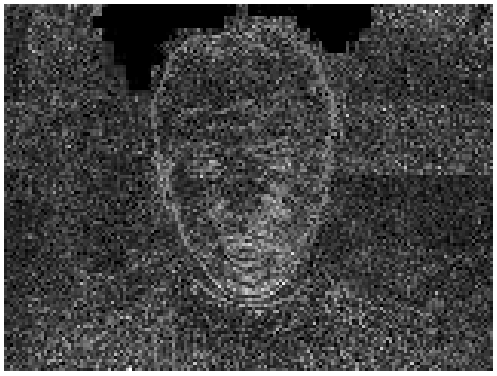


Figure 6.4 Mutual information face for video clip 1.

Figure 6.5 Mutual information face for video clip 2.

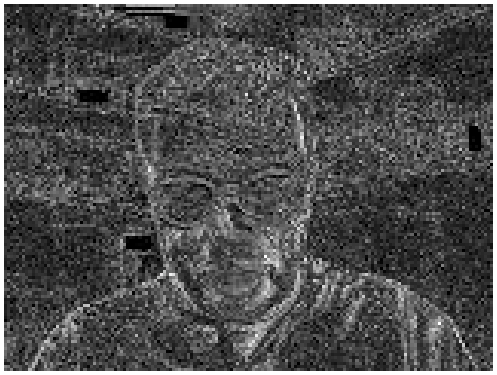


Figure 6.6 Mutual information face for video clip 3.

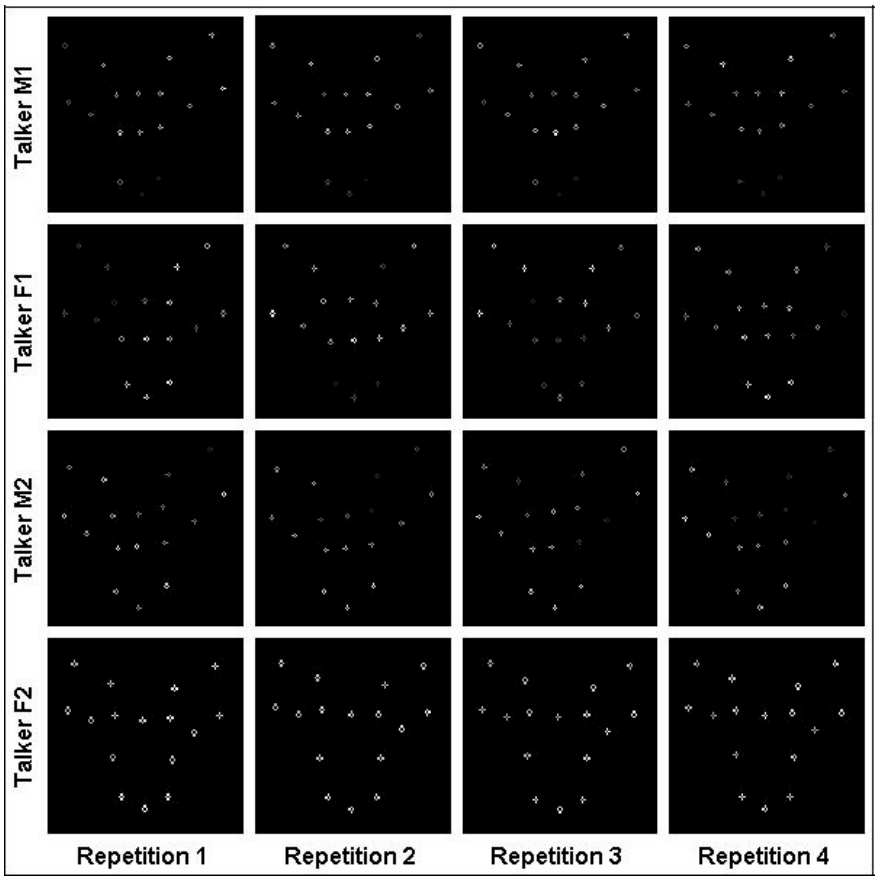


Figure 6.7 Mutual information faces for database *DcorSENT* (Sentence 3, four talkers, and four repetitions).

6.3.4. Discussion

Figures 6.4-6.6 indicated there was a significant amount of mutual information between the audio and the background (areas surrounding faces), which was not a good result. This may be due to noise from the camera. Note that mutual information faces in (Nock et al., 2002) were generated using a database recorded in a studio environment and the camera was very stable. Therefore, the background showed no mutual information with the audio (Figure 6.8). Also, Nock et al. (2003) used Butz-style features based on pixel

intensity changes in the local area, and the Butz features get rid of this problem to some extent.

(a) (b)

Figure 6.8 Mutual information faces in (Nock et al., 2002).

Note that in Figures 6.4-6.6, low mutual information does not necessarily mean low correlation between speech acoustics and face movements. For example, the cheek area does not show high mutual information, but only because the cheeks tend to have uniform intensity. In other words, high mutual information appears at the regions that are both related to speech events and have enough variability in intensity. This is especially true for boundary regions as shown in Figures 6.4-6.6.

Nevertheless, the mutual information "faces" are still clearly shown in Figures 6.4-6.6. The mutual information, however, also depends on the content of speech. Figures 6.4-6.6 show different patterns of mutual information. Mutual information faces can be used to limit the search space in face detection or detect who is talking in one video clip.

The mutual information faces (Figure 6.7) for database *DcorSENT* indicated how strongly each point on the face was coupled with acoustic signals. For three (M1, M2, and F2) of the four talkers, the mutual information pattern was consistent across repetitions. The inconsistency for talker F1 may contribute to her low intelligibility and

thus low correlations (see Chapter 4). The mutual information faces for talkers M1, F1, and M2 showed clear signs of non-symmetry of face movements (left vs. right). Results in Table 4.2 indicated that the predictions of face movements from speech acoustics were better for talker M2 than for talker M1, and in Figure 6.7, talker M2 seems to have a stronger mutual information face. Similarly, the difference in mutual information faces for talkers F1 and F2 can be used to explain the difference in the correlations for talkers F1 and F2 reported in Table 4.2. Figure 4.8 showed that the lip movements were better predicted from speech acoustics than were the chin movements for talker M1, and this coincides with the low mutual information on the chin for him, as shown in Figure 6.7.

6.4. Summary

In this chapter, the correlation between face movements and speech acoustics was examined from a different point of view (compared to Chapter 4). First, a mutual-information-faces algorithm (Nock et al., 2002, 2003) was applied to three video clips (Potamianos and Neti, 2003). Although the algorithm has some limitations (it depends on lighting conditions, intensity variability, etc.), results indicated significant amounts of mutual information between speech acoustics and face movements. For comparison, the mutual-information algorithm was also applied to the *DcorSENT*, and results were consistent with the results reported in Chapter 4, that is, lip movements were well predicted from speech acoustics and thus had strong mutual information with speech acoustics.

CHAPTER 7. SUMMARY, IMPLICATIONS, AND FUTURE DIRECTIONS

7.1. Summary

In this dissertation, the relationship between speech acoustics and optical (and articulatory) movements, and the relationship between visual consonant perception and optical measures were examined in an effort to improve future visual speech synthesis and audio-visual speech recognition. A database of 69 consonant-vowel (CV) syllables and three sentences spoken by four talkers was recorded with four simultaneous data streams: audio, video, face movements, and tongue movements. The relationships between face movements, tongue movements, and speech acoustics were examined using multilinear regression. To examine the relationship between visual consonant perception and optical measures, another database of 69 CV syllables was recorded with three simultaneous data streams: audio, video, and face movements. Each talker's syllable productions were presented for identification in a visual-only condition to six perceivers. Physical measures of the talkers' productions were analyzed using multilinear regression, and perceptual measures of the perceivers' responses were analyzed using multidimensional scaling (MDS) and phoneme equivalence classes (PECs).

7.2. Major Results

7.2.1. Chapter 4 – On the Relationship between Face Movements, Tongue Movements, and Speech Acoustics

Results showed that the relationship between speech acoustics and optical movements was not uniform across talkers, vowel context, place of articulation, and individual optical, articulatory, or acoustic channels. For CV syllables, the correlations between face movements and tongue movements were high (r = 0.70 - 0.88), and optical and articulatory data could be well predicted from speech acoustics (r = 0.74 - 0.82). LSP data were better predicted from EMA than from OPT (r = 0.54 - 0.61 vs. r = 0.37 - 0.55). Another interesting fact about these correlations was the asymmetry of the predictions. In general, optical and articulatory movements were easier to predict from speech acoustics than the reverse. The data from the two males gave better predictions than those from the two females. Therefore, face size may be an effect in the predictions. The prediction of C/a/ syllables was better than C/i/ and C/u/. Furthermore, vowel-dependent predictions produced much better correlations than syllable-independent predictions. In regard to the acoustic features, the 2nd LSP pair, which is around the 2nd formant frequency, and RMS energy, which is related to mouth aperture, were better predicted than other LSP pairs. The tongue and face movements are more related to the front cavity of the vocal tract and thus correlated well with the 2nd formant. Across different places of articulation, lingual places in general resulted in better predictions of one data stream from another compared

to bilabial and glottal places. Among the manners of articulation, plosive consonants yielded lower correlations than others, while voicing had no influence on the correlations. The chin movements were the best predicted, followed by the lips, and then the cheeks. A large level of redundancy among TB, TM, and TT and among chin, cheek, and lip movements was found.

In general, predictions for syllables yielded higher correlations than those for sentences, which suggests a strong context effect in the relationships. The internal tongue movements cannot predict the RMS energy and LSP well over long periods (sentences), while they were predicted reasonably well for short periods (CVs). When predicted from LSP, TB and TM were better predicted than TT. When predicted from OPT, TT was the best predicted for sentences, while the worst for CVs. For sentences, using all channels usually resulted in better prediction; lip movements were the most informative when predicting LSP or EMA; when predicting LSP or OPT, TT was the most informative channel. Due to the non-uniformity of the relationships for CV syllables, a dynamical model (a backward and forward filter) was proposed to enhance the relationship between speech acoustics and optical movements for sentences, and an improvement of about 17% was reported.

7.2.2. Chapter 5 - The Relationship between Visual Speech Perception and Physical Measures

Results showed that physical measures accounted for about 66% of the variance of visual consonant perception. Among the different facial regions, the lip area (55%) was the

most informative, although the cheeks (36%) and the chin (32%) also contributed significantly to visual intelligibility. The variance accounted for in the visual perceptual results by the physical measures demonstrated that visual speech stimulus structure drives visual speech perception. The correlations were not uniform across vowel context and talkers. Talkers with high intelligibility ratings yielded more detailed information, and so did the larger faces of the male talkers. As for vowel context, C/u/ syllables yielded higher correlations between visual perceptual distances and physical distances than C/a/ and C/i/ syllables even though the movements for /u/ were smaller.

The correlations between visual perceptual distances and physical distances varied with different MDS spaces. Results in the current study showed that a 3-D MDS was acceptable (from 3-D to 6-D MDS, the correlations did not differ much). An MDS analysis showed that clusters from lipreading had an internal structure as well as different distances to other clusters. For the confusion dissimilarity matrices predicted from the whole face and the lips, their 3-D MDS structures were similar to those from perceptual data. For the confusion dissimilarity matrices predicted from the cheeks and chin, their 3-D MDS structures were far different from those from perceptual data, where the 23 consonants were spread out rather than clustered. Therefore, the cheeks and chin were not sufficient to produce acceptable lipreading results, but combined with the lips, they helped to differentiate these consonants, and thus made them easier to lipread. The correlation between perceptual data and manner of articulation decreased rapidly when the number of MDS dimensions decreased, suggesting that manner of articulation by itself was a multidimensional structure in a visual perceptual space. The strong

correlation between perceptual data and manner of articulation suggests that perceivers also got manner of articulation information from visual speech, not just place of articulation. This contradicts the general view that manner of articulation is mainly auditory. Lipreading participants tended to overestimate information about place of articulation, while underestimating information about voicing or manner of articulation on the talkers' face. That is, lipreaders did better than predicted from physical measures when identifying place of articulation, while they did worse when identifying voicing and manner of articulation.

Visual perception experiments revealed that PECs were fewer than 12, and that the PECs were a function of vowel context and talkers' sentence intelligibility ratings. In general, as expected, higher sentence intelligibility corresponded to more PECs in the perception of CV stimuli (with fewer phonemes in each PEC), and the obtained PECs tended to agree with the places of articulation. For results predicted from physical measures, PECs obtained from all face retro-reflectors agreed the best with the lipreading results, followed by those obtained only from the lip area. As expected, the lip area was the most informative for lipreading. However, the cheek and chin areas were also contributing significantly to lipreading. PEC analysis further showed that, for low intelligibility talkers (M1 and F1; as perceived by deaf lipreaders), the lip area contributed almost all the information, and the cheeks and chin did not move much (talker F1) or their movements were redundant with the lips (talker M1). On the contrary, high intelligibility talkers (M2 and F2; as perceived by deaf lipreaders) exhibited complementary cheek and chin movements to achieve high visual intelligibility. In their

cases, the cheek and chin areas provided useful supplementary information for lipreading, even though those areas alone were not sufficient for lipreading.

7.2.3. Chapter 6 - Examining the Correlations between Face Movements and Speech Acoustics Using Mutual Information Faces

The mutual information faces were clearly shown in video clips. This information can be used for such tasks as limiting the search space in face detection and detecting who is talking in a video clip (Nock et al., 2002, 2003). The mutual information, however, also depended on the content of speech.

The mutual information faces for the database *DcorSENT* indicated how each point on the face was coupled with the acoustic signals. For three (M1, M2, and F2) of the four talkers, the mutual information pattern was consistent across repetitions. The inconsistency for talker F1 may contribute to her low intelligibility and thus low correlations. The mutual information faces for talkers M1, F1, and M2 showed clear signs of non-symmetry of face movements. Differences across mutual information faces can be used to explain differences across the correlations reported in Chapter 4.

7.3. Implications for Visual Speech Synthesis

7.3.1. Visual Speech Synthesis: A Promising Approach

The most emergent application for visual speech synthesis is to build research tools. These tools can be used to benefit the hearing-impaired, to design visual perception experiments (without human talkers), to predict intelligibility in various environments, to improve audio-visual speech recognition/coding, etc. Visual speech synthesis systems can be used in many real applications such as second-language learning, video communication, clinical evaluations, remote-education, internet agents, animation films, games, virtual reality, ATM, etc.

Animating faces on computers began in the early 1970s, and research on various aspects of computer facial animations has been growing ever since (Bailly, 2001, 2002; Bergeron and Lachapelle, 1985; Bregler et al., 1997; DiPaola, 1989, 1991; Ezzat and Poggio, 1998; King, 2001; Lee et al, 1993, 1995; Magnenat-Thalmann et al., 1988; Platt and Badler, 1981; Parke, 1974, 1982; Parke and Waters, 1996; Pearce et al., 1986; Rydfalk, 1987; Saintourens et al., 1990; Terzopoulos and Waters, 1990, 1991; Waters, 1987, 1992; Waters and Levergood, 1993). Recently, research on facial animation has attracted more attention: This trend is demonstrated by the appearance of Audio-Visual Speech Processing Workshops (1997 in Rhodes, Greece; 1998 in Terrigal, Australia; 1999 in Santa Cruz; 2001 in Scheelsminde, Denmark; 2003 in Saint Jorioz, France). In

2002, several special sessions on audio-visual speech processing were held at ICSLP'02 in Denver, Colorado.

7.3.2. Challenges in Visual Speech Synthesis

Speech articulators cannot move from one gesture to another instantly, and thus at any moment, articulators may appear to be simultaneously adjusted to two or more gestures (Benguerel and Pichora-Fuller, 1982; Miller, 1999). Coarticulation and animation are the two most challenging problems in visual speech synthesis. Synthesizing without any coarticulation effects would result in unnatural stimuli. However, how to drive visual speech synthesis to get more natural coarticulation patterns has not been extensively studied. Analogous to locus equations for acoustic speech synthesis, Cohen and Massaro (1993) used a dominance coarticulation model to approximate coarticulation patterns, but such an approximation may result in loss of naturalness.

Another challenge is to evaluate the visual intelligibility of synthetic visual speech. Current research systems have used human subjects to evaluate the quality of visual speech synthesis (Cohen and Massaro, 1993; Cohen et al., 1996; Le Goff et al., 1994). This method takes a lot of manpower and is subject to bias. An objective measurement of the quality of synthetic visual speech may become possible, if visual speech perception can be predicted from the corresponding optical measures.

7.3.3. Implications from the Current Study

A theoretical ideal driving source for facial animation is speech acoustics, because optical signals are simultaneous by-products of acoustic speech production (Lavagetto et al., 1994; Massaro et al., 1999; Rao et al., 1997; Williams et al., 2000). For the current study, optical and articulatory data could be well predicted from speech acoustics, this demonstrated that speech acoustics contained most of the information about coarticulation and movements of articulators. Therefore, when using speech acoustics as a driving source for visual speech synthesis, coarticulation patterns can be recovered well (if not perfectly). Moreover, such a method does not need additional consideration of synchronization between acoustic and optical speech. However, the prediction of optical speech from speech acoustics appears to be phoneme-dependent. Therefore, when predicting optical speech from speech acoustics, the context effect should be considered, or one can exploit the dynamical information as discussed in Chapter 4. Furthermore, because the 2nd LSP pair correlated better with physical movements than did other LSP pairs, when predicting face or tongue movements from speech acoustics, more resolution could be placed around the 2nd LSP pair.

Lip, chin, and cheek movements are redundant to some degree. However, cheek and chin movements still contribute significantly to visual speech perception. Therefore, for synthetic talking faces, the requirement for a highly intelligible face is that the cheek and chin movements should also be carefully created. “Speaking clearly” has been explored by Picheny et al. (1985, 1986, 1989), while how to “speak clearly using facial gestures” is still an ongoing and a less-explored question. For synthetic faces,

understanding this issue is crucial. Cheek movements should not only convey redundant information but also additional information. The issue of how to make a talker more intelligible is also about how to make a talker's face better sampled. In this regard, larger moving regions (e.g., more pixels) or high-dimensional movements may be proposed.

Currently, visual speech synthesis development goes through a long cycle. Human perceivers are needed to judge the performance of any visual speech synthesizer. However, if the relationship between visual speech perception and physical measures is known, then it would be easier to assess the performance of a visual speech synthesizer. This will, of course, reduce the design cycle and its cost. Results in this dissertation showed that visual speech perception could be predicted successfully from physical measures, especially if more face movements were captured.

7.4. Future Directions

The current study showed that a significant amount of optical information could be recovered from speech acoustics, and such predictions were context-dependent. Therefore, our future acoustics-driven visual speech synthesis system will probably be a diphone-based concatenation system (stored small segments of speech are retrieved when they are needed). [Peterson et al. (1958) proposed the *diphone*, which is delimited in the middle of each phoneme, and thus the boundaries are located in the most stable region.]

For the experiments on optical phonetics, we will examine more facial markers and thus attempt to adequately account for visual speech perception by physical

measures. Then we need to find out which facial regions are important for visual speech perception, especially for the perception of place of articulation, because lipreaders appear to overestimate the place of articulation information from visual speech. Subsequently, we will focus on movements for these "important" regions.

The estimation of optical speech from speech acoustics is not 100% successful. Some optical movements cannot be estimated from speech acoustics. We refer to these movements orthogonal optical movements (to speech acoustics). Once we have a visual speech synthesizer driven by speech acoustics, we will examine whether these orthogonal optical movements have a significant effect on visual speech perception.

A limitation of the current study is that correlation analysis was carried out uniformly across time without taking into account important gestures or facial landmarks. For example, some specific face gestures or movements may be very important for visual speech perception, such as mouth closure for a bilabial sound. In the future, physiological and perceptual experiments should be conducted to define which face movements are of importance to visual speech perception, so that those movements can be better predicted.

For the relationship between visual speech perception and optical measures, the dynamical characteristics of the face movements were included in physical distances, but the averaging effect in some cases may not have been desirable. In the future, the effects of consonant duration and velocity of the movements will be examined. Also, for physical distances, a simple average was taken across the talkers. One question is that, because different talkers have different styles of speaking and different face sizes, do we need some normalization across talkers, and if so, how? In the current study, the

coordinates of retro-reflectors were used. Visual speech perception, however, may use just relative information. Therefore, we will examine the correlation as a function of facial configurations, such as lip width, lip height, lip protrusion, lip opening area, stress on the face, etc.

APPENDIX A. CONFUSION MATRICES

In this appendix, we present the confusion matrices for the 23-consonant identification experiments described in Chapter 2. Each row represents a distribution of responses to one stimulus. $\boldsymbol{C}_{T,V}$ represents one confusion matrix, where T is the talker and V is the vowel context. Results are shown in confusion matrices. The results were first pooled in 12 stimulus-response (23x23) confusion matrices for each talker and vowel context. In each of these confusion matrices, each stimulus has 120 responses (2 repetitions x 6 participants x 10 trials). These confusion matrices were then pooled across talkers or vowels (or both).

Table A.1 Perceptual confusion matrix $\boldsymbol{C}_{\mathrm{M1,a}}$.

	y	w	r	l	m	n	p	t	k	b	d	g	h	θ	ð	s	z	f	v	ʃ	ʒ	tʃ	dʒ
y	23			25		20		1	11		12	12	13	1	1						1		
w	1	117	1					1															
r			3															107	10				
l				104		2		2		1	3			5	3								
m					29		24			67													
n				101		15		1			2		1										
p		1			33		23			63													
t				5		7		41	1		33	3	2	3	2	18	3				1	1	
k	9			18		9		5	35		9	13	22										
b					29		20			71													
d				15		14		24			44	1		3	1	14	2		1	1			
g	8			11				3	42		3	12	39		1	1							
h											1		118									1	
θ	1			16		11		4	3		5	5		40	35								
ð	1			3		11		4			6			33	59	2				1			
s						1		10			18					77	11				1		2
z								13			22				1	70	13					1	
f											1							102	17				
v																		99	21				
ʃ				2		1		5		1	15					57	6			8	2	15	8
ʒ					1	1		6			8					38	2			21	2	24	17
tʃ							1	2			3		1			38	6		1	29	5	20	14
dʒ						1		16			19			1		52	7			6	1	12	5

Table A.2 Perceptual confusion matrix $\boldsymbol{C}_{\mathrm{M2,a}}$.

	y	w	r	l	m	n	p	t	k	b	d	g	h	θ	ð	s	z	f	v	ʃ	ʒ	tʃ	dʒ
y	5			36		16		5	23		10	7	18										
w	2	117			1																		
r		72	43					1	1				1			1				1			
l				115		1		1			2	1											
m					62		8			50													
n				86		18			4		7	2	2	1									
p					52		25			43													
t				1		2		39	1	1	25	1	2	1		39	7			1			
k				35		2		2	41		4	14	21									1	
b				1	68		11			40													
d				8				47	2		38	3	2		1	12	3	1		1			2
g	1			29		12		3	33		17	18	7										
h	1			5				4	11			8	91										
θ														51	69								
ð														55	64			1					
s								21	2		22					68	7						
z				1				20			20					65	9					4	1
f																		102	18				
v								1										97	22				
ʃ								10			5					11	1			22	3	39	29
ʒ	1							19			16	1	1			3	1			6		37	35
tʃ	2							12	1		5			1		1				10	1	50	37
dʒ	1			3		1		26	1		12	1				1				2	1	42	29

Table A.3 Perceptual confusion matrix $\boldsymbol{C}_{\mathrm{F1,a}}$.

	y	w	r	l	m	n	p	t	K	b	d	g	h	θ	ð	s	z	f	v	ʃ	ʒ	tʃ	dʒ
y	6			11		16		6	37		14	14	4	1	1	10							
w	2	107	9							1	1												
r	2	23	74	1	5	2	1		2	4	1		1					4					
l	3			17		14		20	19		18	10	8	1	1	6	1			1			1
m					46		28			46													
n	3					10		22	22		31	9	4	2		17							
p					46		43			29	1								1				
t						2		20	2		29	1	3	2		52	9						
k	1			2		9		3	33		11	7	48	2		4							
b			1		43		38			37													1
d								21			23				1	59	7					5	4
g	5			12		13		3	26		12	16	23	2		8							
h	1					7		2	16		3	8	82			1							
θ					1			2						28	88	1							
ð								2			2		1	48	67								
s	1					1		26			27			1	3	54	7						
z					1			18			19	1		1	2	57	8			1	1	3	8
f								1					1					105	13				
v																		103	17				
ʃ	1			1		2	1	3	2		3					17				15	4	26	45
ʒ	1			5		5		2			7		1	1	4	16	2			11	2	31	32
tʃ	3					3	1	8	2		10	9			2	32	2			6	2	21	19
dʒ															1	11	1			24	9	32	42

Table A.4 Perceptual confusion matrix $\boldsymbol{C}_{\mathrm{F2,a}}$.

	y	w	r	l	m	n	p	t	k	b	d	g	h	θ	ð	s	z	f	v	ʃ	ʒ	tʃ	dʒ
y	35			17		27		3	10		9	8	2	2	7								
w	3	117																					
r			2															99	19				
l				108		5					4	2			1								
m			1		29		18			72													
n	2			1		13		36	1		42	7	6			9	2				1		
p					18	1	30			68						3							
t	2			2		3		26	4		33	3	3			34	1			6		2	1
k				5		1		1	34			13	66										
b			1		11		41			67													
d						1		23			26			1	1	53	9					1	5
g	3			14		7		2	26		2	15	48	1								1	1
h	1								9			5	103	1		1							
θ								1						44	75								
ð	4			4		18		10	6		20	10	17	11	19	1							
s								3			14					77	24			2			
z								4			14					92	10						
f																		99	21				
v		1	4										1					97	17				
ʃ								1			6					13	1			30	7	32	30
ʒ				1							1					2	1			29	7	34	45
tʃ									1							5				28	3	56	27
dʒ								1		1										31	2	40	45

Table A.5 Perceptual confusion matrix $\boldsymbol{C}_{\mathrm{M1,i}}$.

	y	w	r	l	m	n	p	t	K	b	d	g	h	θ	ð	s	z	f	v	ʃ	ʒ	tʃ	dʒ
y	18					5		2	21	1	5	17	7					30	11		2		1
w		117	2							1													
r		2	21				1											81	15				
l				87		5		3						19	6								
m					17		32			70											1		
n				94		19		2	1		1			3									
p					10		35			75													
t				7		19		20			45	2		13	4	6	2				1	1	
k	28				1	6		1	40		2	15	24	1	1			1					
b					15		44			61													
d				41		64		1	2		6			4	1	1							
g	11		1	2		22		3	38		7	16	16	3	1								
h	2				1	1			11		1	2	99			1	1					1	
θ				6		21					2	1		56	34								
ð				2		6		5		1	3			45	58								
s								7			7					81	22			1	1		1
z			1					12	1		20			1		59	18			4	1	1	2
f				1									1					79	39				
v																		78	42				
ʃ									1		2	8	1			42	5			23	11	17	10
ʒ				1				2			3	18				33	4			23	7	13	16
tʃ				1		1		17			21	6				55	17			2			
dʒ	1							1			3	15	1	1		49	9			15	2	14	9

Table A.6 Perceptual confusion matrix $\boldsymbol{C}_{\mathrm{M2,i}}$.

	y	w	r	l	m	n	p	t	k	b	d	g	h	θ	ð	s	z	f	v	ʃ	ʒ	tʃ	dʒ
y	2			5		25		1	42		6	18	20	1									
w		119	1																				
r		100	19															1					
l				66		20			7		2	8	17										
m					60		17			42						1							
n	1			30		44		1	22		1	8	13										
p			1		62		17			40													
t				1	1	21		16	13		30	2	30			6							
k				1		7			17		3	3	87	2									
b					39		29			52													
d						15		31	5		27	8	13			18	1	1			1		
g	1			1		24		1	28		2	9	54										
h						12			17			10	81										
θ														60	60								
ð				3	1	2		1	4		3	3	3	60	39		1						
s								5			20	2				78	14					1	
z				1		3		9			25		1			64	16			1			
f								1										87	32				
v																		89	31				
ʃ						2		2	2		8	12	4			10				40	8	20	12
ʒ									1		1	12				12				55	4	16	19
tʃ				2		1		5			8	8	4			24	3			26	2	24	13
dʒ						1					6	10				20	6			40	2	17	18

Table A.7 Perceptual confusion matrix $\boldsymbol{C}_{\mathrm{F1,i}}$.

	y	w	r	l	m	n	p	t	k	b	d	g	h	θ	ð	s	z	f	v	ʃ	ʒ	tʃ	dʒ
y	4			1		9		22	10		37	4	4	7		18	2			1			1
w		78	35		2					3									1				1
r		52	66		1											1							
l	2			3		7		29			33	2	5			34	5						
m					84		7			28	1												
n	2			20		35		10	2		17	5	4			23		1				1	
p		3	1		48		29			39													
t				1		4		22	2		25	3	4	7	12	35	3			1	1		
k	3				1	12		13	22		18	5	37	2	1	6							
b					26		36			58													
d						1		21	3		30	2				48	14			1			
g						2		22	1		18	4	1			40	5			11	2	2	12
h				2		21		2	15		13	1	56	4		6							
θ				1										77	42								
ð														68	52								
s						1		14		1	24		1	6	3	54	12		2	2			
z	2					1	5	5		14	14	3	1	1		47	6			2	5	3	11
f																		84	36				
v						1												74	45				
ʃ	2					3		1			5		1	2		33	3			18	9	9	34
ʒ								2			2	20	2			18	2			41	13	7	13
tʃ								1				20	1			7				50	3	13	25
dʒ												22				9	1			51	4	12	21

Table A.8 Perceptual confusion matrix $\boldsymbol{C}_{\mathrm{F2,i}}$.

	y	w	r	l	m	n	p	t	k	b	d	g	h	θ	ð	s	z	f	v	ʃ	ʒ	tʃ	dʒ
y				16		31			18		8	4	41	1					1				
w		116	3				1																
r		6	47		1											1		59	6				
l	2			75	1	29		1	3			1	6	2									
m				1	47		30			42													
n				3		15		10	13		15	6	53			4	1						
p					15		57			48													
t						8		20	5		30		4			41	12						
k	1		1	1		10		3	18		1	14	70	1									
b					18		43			59													
d						11		29	1		27	1	2			45	4						
g						4		8	11		18	5	40			21	12					1	
h	1					2			11			5	98	1				1		1			
θ														68	52								
ð				13		6		12	2		6	4	1	43	30	3							
s	1							3			9					90	17						
z								2			10					85	23						
f			1			1								1				84	33				
v																		89	31				
ʃ											1	16				5				55	5	11	27
ʒ												12				4				52	6	19	27
tʃ										1	1	16				3				41	9	27	22
dʒ												10				3				59	8	17	23

Table A.9 Perceptual confusion matrix $\boldsymbol{C}_{\mathrm{M1,u}}$.

	y	w	r	l	m	n	p	t	k	b	d	g	h	θ	ð	s	z	f	v	ʃ	ʒ	tʃ	dʒ
y	37	24	1	7					1			2	48										
w		119											1										
r	1	1	1	1		1		5			4			2		7	5	68	19			1	4
l				110		1							1	8									
m					49		20			50	1												
n	1			115		1								2							1		
p					31		26			63													
t	7	1		17	1	6		14	4		13	1		14	2	15	21					1	3
k	6	26		3		2			12	1		6	64										
b				1	38		21			60													
d	5	1		52		5		11	1		22			11	2	3	6				1		
g	12	19		17		2		3	23			5	39										
h	10	32	1	2		1			1		1	1	70						1				
θ				47		1								66	5		1						
ð	6	2		23		6		15			17		1	27	10	7	5	1					
s				1		1		8			9			5	1	48	41			1	1	2	2
z				1		2		9			5			1		57	37			4		3	1
f															1			93	26				
v																		96	24				
ʃ	4					1		6	2		11					44	35			4	1	9	3
ʒ								13			8				1	43	26			14	1	6	8
tʃ	9			1		1		13	2		6			2		35	29			5	4	4	9
dʒ	2							5	1		4	1				53	35			10	3	3	3

Table A.10 Perceptual confusion matrix $\boldsymbol{C}_{\mathrm{M2,u}}$.

	y	w	r	l	m	n	p	t	k	b	d	g	h	θ	ð	s	z	f	v	ʃ	ʒ	tʃ	dʒ
y	31	30		20		2			3		1		29	4									
w		119			1																		
r	1	99	15		1						1			1		2							
l		9		102					1				2	4			2						
m					111		4			4			1										
n	3	12	2	70		5		2	4			1	17	3	1								
p		1			68		27			24													
t	22	14		3				25	11		9	4	15			6	6						5
k		33	1	1					1				83				1						
b					71		20			28			1										
d	21	18	1	10		5		17	4		12	4	7	9		4	7					1	
g	23	23	1	16		1			6			2	43	4	1								
h		41		1					2				76										
θ	2	9		7	2								1	79	18				1				1
ð				1										99	20								
s	1	2				1		14			8			1		47	34			5	2	1	4
z	1	1						8	1		7			1		53	33			6	1	3	5
f																	1	90	29				
v														4				87	29				
ʃ	5							21			7			2		41	17			8	3	9	7
ʒ	2							16			13					37	22	1		10	3	5	11
tʃ	17	3						33	2		15		2			15	9			8	3	7	6
dʒ	19	7		1		1		18	2		7	3	2			16	14			10		10	10

Table A.11 Perceptual confusion matrix $\boldsymbol{C}_{\mathrm{F1,u}}$.

	y	w	r	l	m	n	p	t	k	b	d	g	h	θ	ð	s	z	f	v	ʃ	ʒ	tʃ	dʒ
y	56			1		2		9	12		17	3	2			1	2		1	7	2	1	4
w		77	12		31																		
r		19	74	1	22			1					2					1					
l	57	1		9		1		10	9		8	12	3	2		1	2			1		3	1
m					103		9			7			1										
n	11			1				39	2		19	2	1			3	16			8		9	9
p			2		72		14			31										1			
t	21	1				3		32	3		14		1			10	9			7		5	14
k	30	5		2		6		5	18		13	19	14			1	4	1		2			
b					70		22			27								1					
d	15			1				21			9	1	1			6	15			14	2	9	26
g	48			5		2		3	23		3	18	16	1		1							
h	7	22		1					5			4	79	1									1
θ														109	11								
ð				3										96	21								
s	10							15			19			1		8	11		1	20		7	28
z	11							31			17			2		9	14			17	2	1	16
f										1						1		93	25				
v																		86	33	1			
ʃ	16					1		9			7					3	2			33	4	14	31
ʒ	17			1				12			7	1				1	3		1	24	4	20	29
tʃ	30			1		1		14	4		7	2				8	10			14	1	3	25
dʒ	17							6			12	1				1	4			28	11	17	23

Table A.12 Perceptual confusion matrix $\boldsymbol{C}_{\mathrm{F2,u}}$.

	y	w	r	l	m	n	p	t	k	b	d	g	h	θ	ð	s	z	f	v	ʃ	ʒ	tʃ	dʒ
y	84							4	6		2		1			1	1	2	1	8		5	5
w		81	30		4					3									2				
r			7															84	29				
l	47			59		4			4		2	3											1
m					46		24			50													
n	57			13		1		9	2		17	3	1	5	4		1			3		2	2
p					51		23			46													
t	16							25	8		12	4	6			4	8			11	1	7	18
k	22	17		3				4	22		2	15	34			1							
b					59		26			35													
d	39					2		20	3		22	6	6				3		1	10	1	3	4
g	35	8		2		1			34		1	22	16					1					
h	6	29			3		2		7	1	2	5	65										
θ				3		1		1						102	13								
ð	3			3		1		3				1		95	13						1		
s	4	1						23			14		2			29	30			8		2	7
z	6							14	1		18					25	33			11	2	5	5
f																		86	34				
v	2										1							88	29				
ʃ	1							8			11					6	3			38	10	15	28
ʒ	1							6			12					14	6			38	5	13	25
tʃ	2							11		1	14			1		1	4			37	5	17	27
dʒ	1							6	1		9					11	6			30	3	17	36

Table A.13 Perceptual confusion matrix $\boldsymbol{C}_{\mathrm{M1,aiu}}$.

	y	w	r	l	m	n	p	t	k	b	d	g	h	θ	ð	s	z	f	v	ʃ	ʒ	tʃ	dʒ
y	78	24	1	32		25		3	33	1	17	31	68	1	1			30	11		3		1
w	1	353	3					1		1			1										
r	1	3	25	1		1	1	5			4			2		7	5	256	44			1	4
l				301		8		5		1	3		1	32	9								
m					95		76			187	1										1		
n	1			310		35		3	1		3		1	5							1		
p		1			74		84			201													
t	7	1		29	1	32		75	5		91	6	2	30	8	39	26				2	3	3
k	43	26		21	1	17		6	87	1	11	34	110	1	1			1					
b				1	82		85			192													
d	5	1		108		83		36	3		72	1		18	4	18	8		1	1	1		
g	31	19	1	30		24		9	103		10	33	94	3	2	1							
h	12	32	1	2	1	2			12		3	3	287			1	1		1			2	
θ	1			69		33		4	3		7	6		162	74		1						
ð	7	2		28		23		24		1	26		1	105	127	9	5	1		1			
s				1		2		25			34			5	1	206	74			2	3	2	5
z			1	1		2		34	1		47			2	1	186	68			8	1	5	3
f				1							1		1		1			274	82				
v																		273	87				
ʃ	4			2		2		11	3	1	28	8	1			143	46			35	14	41	21
ʒ				1	1	1		21			19	18			1	114	32			58	10	43	41
tʃ	9			2		2	1	32	2		30	6	1	2		128	52		1	36	9	24	23
dʒ	3					1		22	1		26	16	1	2		154	51			31	6	29	17

Table A.14 Perceptual confusion matrix $\boldsymbol{C}_{\mathrm{M2,aiu}}$.

	y	w	r	l	m	n	p	t	k	b	d	g	h	θ	ð	s	z	f	v	ʃ	ʒ	tʃ	dʒ
y	38	30		61		43		6	68		17	25	67	5									
w	2	355	1		2																		
r	1	271	77		1			1	1		1		1	1		3		1		1			
l		9		283		21		1	8		4	9	19	4			2						
m					233		29			96			1			1							
n	4	12	2	186		67		3	30		8	11	32	4	1								
p		1	1		182		69			107													
t	22	14		5	1	23		80	25	1	64	7	47	1		51	13			1			5
k		33	1	37		9		2	59		7	17	191	2			1					1	
b				1	178		60			120			1										
d	21	18	1	18		20		95	11		77	15	22	9	1	34	11	2		1	1	1	2
g	25	23	1	46		37		4	67		19	29	104	4	1								
h	1	41		6		12		4	30			18	248										
θ	2	9		7	2								1	190	147				1				1
ð				4	1	2		1	4		3	3	3	214	123		1	1					
s	1	2				1		40	2		50	2		1		193	55			5	2	2	4
z	1	1		2		3		37	1		52		1	1		182	58			7	1	7	6
f								1									1	279	79				
v								1						4				273	82				
ʃ	5					2		33	2		20	12	4	2		62	18			70	14	68	48
ʒ	3							35	1		30	13	1			52	23	1		71	7	58	65
tʃ	19	3		2		1		50	3		28	8	6	1		40	12			44	6	81	56
dʒ	20	7		4		3		44	3		25	14	2			37	20			52	3	69	57

Table A.15 Perceptual confusion matrix $\boldsymbol{C}_{\mathrm{F1,aiu}}$.

	y	w	r	l	m	n	p	t	k	b	d	g	h	θ	ð	s	z	f	v	ʃ	ʒ	tʃ	dʒ
y	66			13		27		37	59		68	21	10	8	1	29	4		1	8	2	1	5
w	2	262	56		33					4	1								1				1
r	2	94	214	2	28	2	1	1	2	4	1		3			1		5					
l	62	1		29		22		59	28		59	24	16	3	1	41	8			2		3	2
m					233		44			81	1		1										
n	16			21		45		71	26		67	16	9	2		43	16	1		8		10	9
p		3	3		166		86			99	1								1	1			
t	21	1		1		9		74	7		68	4	8	9	12	97	21			8	1	5	14
k	34	5		4	1	27		21	73		42	31	99	4	1	11	4	1		2			
b			1		139		96			122								1					1
d	15			1		1		63	3		62	3	1		1	113	36			15	2	14	30
g	53			17		17		28	50		33	38	40	3		49	5			11	2	2	12
h	8	22		3		28		4	36		16	13	217	5		7							1
θ				1	1			2						214	141	1							
ð				3				2			2		1	212	140								
s	11					2		55		1	70		1	8	6	116	30		3	22		7	28
z	13				1	1	5	54		14	50	4	1	4	2	113	28			20	8	7	35
f								1		1			1			1		282	74				
v						1												263	95	1			
ʃ	19			1		6	1	13	2		15		1	2		53	5			66	17	49	110
ʒ	18			6		5		16			16	21	3	1	4	35	7		1	76	19	58	74
tʃ	33			1		4	1	23	6		17	31	1		2	47	12			70	6	37	69
dʒ	17							6			12	23			1	21	6			103	24	61	86

Table A.16 Perceptual confusion matrix $\boldsymbol{C}_{\mathrm{F2,aiu}}$.

	y	w	r	l	m	n	p	t	k	b	d	g	h	θ	ð	s	z	f	v	ʃ	ʒ	tʃ	dʒ
y	119			33		58		7	34		19	12	44	3	7	1	1	2	2	8		5	5
w	3	314	33		4		1			3									2				
r		6	56		1											1		242	54				
l	49			242	1	38		1	7		6	6	6	2	1								1
m			1	1	122		72			164													
n	59			17		29		55	16		74	16	60	5	4	13	4			3	1	2	2
p					84	1	110			162						3							
t	18			2		11		71	17		75	7	13			79	21			17	1	9	19
k	23	17	1	9		11		8	74		3	42	170	1		1							
b			1		88		110			161													
d	39					14		72	4		75	7	8	1	1	98	16		1	10	1	4	9
g	38	8		16		12		10	71		21	42	104	1		21	12	1			1	1	1
h	8	29			3	2	2		27	1	2	15	266	2		1		1		1			
θ				3		1		2						214	140								
ð	7			20		25		25	8		26	15	18	149	62	4					1		
s	5	1						29			37		2			196	71			10		2	7
z	6							20	1		42					202	66			11	2	5	5
f			1			1								1				269	88				
v	2	1	4								1		1					274	77				
ʃ	1							9			18	16				24	4			123	22	58	85
ʒ	1			1				6			13	12				20	7			119	18	66	97
tʃ	2							11	1	2	15	16		1		9	4			106	17	100	76
dʒ	1							7	1	1	9	10				14	6			120	13	74	104

Table A.17 Perceptual confusion matrix $\boldsymbol{C}_{\mathrm{ALL,a}}$.

	y	w	r	l	m	n	p	t	k	b	d	g	h	θ	ð	s	z	f	v	ʃ	ʒ	tʃ	dʒ
y	69			89		79		15	81		45	41	37	4	9	10					1		
w	8	458	10		1			1		1	1												
r	2	95	122	1	5	2	1	1	3	4	1		2			1		210	29	1			
l	3			344		22		23	19	1	27	13	8	6	5	6	1			1			1
m			1		166		78			235													
n	5			188		56		59	27		82	18	13	3		26	2				1		
p		1			149	1	121			203	1					3			1				
t	2			8		14		126	8	1	120	8	10	6	2	143	20			7	1	3	1
k	10			60		21		11	143		24	47	157	2		4						1	
b			2	1	151		110			215													1
d				23		15		115	2		131	4	2	4	4	138	21	1	1	2		6	11
g	17			66		32		11	127		34	61	117	3	1	9						1	1
h	3			5		7		6	36		4	21	394	1		2						1	
θ	1			16	1	11		7	3		5	5		163	267	1							
ð	5			7		29		16	6		28	10	18	147	209	3		1		1			
s	1					2		60	2		81			1	3	276	49			2	1		2
z				1	1			55			75	1		1	3	284	40			1	1	8	9
f								1			1		1					408	69				
v		1	4					1					1					396	77				
ʃ	1			3		3	1	19	2	1	29					98	8			75	16	112	112
ʒ	2			6	1	6		27			32	1	2	1	4	59	6			67	11	126	129
tʃ	5					3	2	22	4		18	9	1	1	2	76	8		1	73	11	147	97
dʒ	1			3		2		43	1	1	31	1		1	1	64	8			63	13	126	121

Table A.18 Perceptual confusion matrix $\boldsymbol{C}_{\mathrm{ALL,i}}$.

	y	w	r	l	m	n	p	t	k	b	d	g	h	θ	ð	s	z	f	v	ʃ	ʒ	tʃ	dʒ
y	24			22		70		25	91	1	56	43	72	9		18	2	30	12	1	2		2
w		430	41		2		1			4									1				1
r		160	153		2		1									2		141	21				
l	4			231	1	61		33	10		35	11	28	21	6	34	5						
m				1	208		86			182	1					1					1		
n	3			147		113		23	38		34	19	70	3		27	1	1				1	
p		3	2		135		138			202													
t				9	1	52		78	20		130	7	38	20	16	88	17			1	2	1	
k	32		1	2	2	35		17	97		24	37	218	6	2	6		1					
b					98		152			230													
d				41		91		82	11		90	11	15	4	1	112	19	1		1	1		
g	12		1	3		52		34	78		45	34	111	3	1	61	17			11	3	2	12
h	3			2	1	36		2	54		14	18	334	5		7	1	1		1		1	
θ				7		21					2	1		261	188								
ð				18	1	14		18	6	1	12	7	4	216	179	3	1						
s	1					1		29		1	60	2	1	6	3	303	65		2	3	1	1	1
z	2		1	1		4	5	28	1	14	69	3	2	2		255	63			7	6	4	13
f			1	1		1		1					1	1				334	140				
v						1												330	149				
ʃ	2					5		3	3		16	36	6	2		90	8			136	33	57	83
ʒ				1				4	1		6	62	2			67	6			171	30	55	75
tʃ				3		2		23		1	30	50	5			89	20			119	14	64	60
dʒ	1					1		1			9	57	1	1		81	16			165	16	60	71

Table A.19 Perceptual confusion matrix $\boldsymbol{C}_{\mathrm{ALL,u}}$.

	y	w	r	l	m	n	p	t	k	b	d	g	h	θ	ð	s	z	f	v	ʃ	ʒ	tʃ	dʒ
y	208	54	1	28		4		13	22		20	5	80	4		2	3	2	2	15	2	6	9
w		396	42		36					3			1						2				
r	2	119	97	2	23	1		6			5		2	3		9	5	153	48			1	4
l	104	10		280		6		10	14		10	15	6	14		1	4			1		3	2
m					309		57			111	1		2										
n	72	12	2	199		7		50	8		36	6	19	10	5	3	17			11	1	11	11
p		1	2		222		90			164										1			
t	66	16		20	1	9		96	26		48	9	22	14	2	35	44			18	1	13	40
k	58	81	1	9		8		9	53	1	15	40	195			2	5	1		2			
b				1	238		89			150			1					1					
d	80	19	1	63		12		69	8		65	11	14	20	2	13	31		1	24	4	13	30
g	118	50	1	40		6		6	86		4	47	114	5	1	1		1					
h	23	124	1	4	3	1	2		15	1	3	10	290	1					1				1
θ	2	9		57	2	2		1					1	356	47		1		1				1
ð	9	2		30		7		18			17	1	1	317	64	7	5	1			1		
s	15	3		1		2		60			50		2	7	1	132	116		1	34	3	12	41
z	18	1		1		2		62	2		47			4		144	117			38	5	12	27
f										1					1	1	1	362	114				
v	2										1			4				357	115	1			
ʃ	26					2		44	2		36			2		94	57			83	18	47	69
ʒ	20			1				47			40	1			1	95	57	1	1	86	13	44	73
tʃ	58	3		2		2		71	8	1	42	2	2	3		59	52			64	13	31	67
dʒ	39	7		1		1		35	4		32	5	2			81	59			78	17	47	72

Table A.20 Perceptual confusion matrix $\boldsymbol{C}_{\mathrm{ALL,aiu}}$.

	y	w	r	l	m	n	p	t	k	b	d	g	h	θ	ð	s	z	f	v	ʃ	ʒ	tʃ	dʒ
y	301	54	1	139		153		53	194	1	121	89	189	17	9	30	5	32	14	16	5	6	11
w	8	1284	93		39		1	1		8	1		1						3				1
r	4	374	372	3	30	3	2	7	3	4	6		4	3		12	5	504	98	1		1	4
l	111	10		855	1	89		66	43	1	72	39	42	41	11	41	10			2		3	3
m			1	1	683		221			528	2		2			1					1		
n	80	12	2	534		176		132	73		152	43	102	16	5	56	20	1		11	2	12	11
p		5	4		506	1	349			569	1					3			1	1			
t	68	16		37	2	75		300	54	1	298	24	70	40	20	266	81			26	4	17	41
k	100	81	2	71	2	64		37	293	1	63	124	570	8	2	12	5	2		2		1	
b			2	2	487		351			595			1					1					1
d	80	19	1	127		118		266	21		286	26	31	28	7	263	71	2	2	27	5	19	41
g	147	50	2	109		90		51	291		83	142	342	11	3	71	17	1		11	3	3	13
h	29	124	1	11	4	44	2	8	105	1	21	49	1018	7		9	1	1	1	1		2	1
θ	3	9		80	3	34		8	3		7	6	1	780	502	1	1		1				1
ð	14	2		55	1	50		52	12	1	57	18	23	680	452	13	6	2		1	1		
s	17	3		1		5		149	2	1	191	2	3	14	7	711	230		3	39	5	13	44
z	20	1	1	3	1	6	5	145	3	14	191	4	2	7	3	683	220			46	12	24	49
f			1	1		1		2		1	1		2	1	1	1	1	1104	323				
v	2	1	4			1		1			1		1	4				1083	341	1			
ʃ	29			3		10	1	66	7	1	81	36	6	4		282	73			294	67	216	264
ʒ	22			8	1	6		78	1		78	64	4	1	5	221	69	1	1	324	54	225	277
tʃ	63	3		5		7	2	116	12	2	90	61	8	4	2	224	80		1	256	38	242	224
dʒ	41	7		4		4		79	5	1	72	63	3	2	1	226	83			306	46	233	264

APPENDIX B. PHONEME EQUIVALENCE CLASSES

In this appendix, we present tables that list the phoneme equivalence classes (PECs) for each lipreading result (different talkers and vowel context) and the corresponding results predicted from physical measures. For each table, the first row lists PECs derived from lipreading results and the following four rows list PECs predicted from physical measures using all retro-reflectors, lip retro-reflectors, cheek retro-reflectors, and chin retro-reflectors, respectively.

Table B.1 PECs across vowel context for different talkers.

Sources	PECs
Talker M1	
Perception	{w} {m p b} {r f v} {θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{ALL}	{w} {m p b} {r f v} {θ ð y l n k g h}{t d s z ʃ ʒ tʃ dʒ}
Predicted^{lip}	{w} {m p b} {f v} {r θ ð y l n k g h} {t d s z ʃ ʒ tʃ dʒ}
Predicted^{chk}	{w} {m p b g} {f v} {θ ð r y l n k h t d} {s z ʃ ʒ tʃ dʒ}
Predicted^{chn}	{r m f v} {w p b θ y l n k g h} {ð t d s z ʃ ʒ tʃ dʒ}
Talker F1	
Perception	{w r} {m p b} {f v} {θ ð} {y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{ALL}	{w r m p b} {f v} {θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{lip}	{w r m p b} {f v} {θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{chk}	{f v} {w r m p b θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{chn}	{w r m p b f v θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Talker M2	
Perception	{w r} {m p b} {f v} {θ ð} {y l n k g h} {t d s z ʃ ʒ tʃ dʒ}
Predicted^{ALL}	{w r} {m p b} {f v θ ð} {y l n t k d g h s z ʃ ʒ tʃ dʒ}
Predicted^{lip}	{w r} {f v} {m p b θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{chk}	{w r m p b f v θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{chn}	{w r m p b f v} {θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Talker F2	
Perception	{w} {m p b} {r f v} {θ ð} {y l n k g h} {t d s z} {ʃ ʒ tʃ dʒ}
Predicted^{ALL}	{w} {m p b} {r f v} {θ ð y l k g h} {n t d s z} {ʃ ʒ tʃ dʒ}
Predicted^{lip}	{w} {m p b} {r f v} {θ ð y l n k g h t d s z} {ʃ ʒ tʃ dʒ}
Predicted^{chk}	{w m p b} {θ ð y l n k g h t d z} {r f v s ʃ ʒ tʃ dʒ}
Predicted^{chn}	{w} {m r f v} {p b θ ð y l n k g h t d} {s z ʃ ʒ tʃ dʒ}

Table B.2 PECs across talkers for different vowels.

Sources	PECs
Vowel /a/	
Perception	{w} {m p b} {r f v} {θ ð} {y l n k g h} {t d s z ʃ ʒ tʃ dʒ}
Predicted^{ALL}	{w} {m p b} {r f v} {θ ð} {y l n k g h} {t d s z ʃ ʒ tʃ dʒ}
Predicted^{lip}	{w m p b} {r f v} {θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{chk}	{w m p b} {r θ ð y l n k g h t d } {f v s z ʃ ʒ tʃ dʒ}
Predicted^{chn}	{w m p b r f v θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Vowel /i/	
Perception	{w r} {m p b} {f v} {θ ð} {y l n k g h t d } {s z ʃ ʒ tʃ dʒ}
Predicted^{ALL}	{w} {m p b} {r f v} {θ ð y l n k g h} {t d s z ʃ ʒ tʃ dʒ}
Predicted^{lip}	{w} {m p b t} {r f v} {θ ð y l n k g h d} {s z ʃ ʒ tʃ dʒ}
Predicted^{chk}	{w r m p b f v θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{chn}	{w r m f v}{p b θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Vowel /u/	
Perception	{w} {m p b} {r f v} {θ ð} {y l n k g h} {t d s z ʃ ʒ tʃ dʒ}
Predicted^{ALL}	{w r m p b} {f v} {θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{lip}	{w r m p b θ ð} {f v} {y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{chk}	{w m p b h} {f v} {r θ ð y l n k g t d s z ʃ ʒ tʃ dʒ}
Predicted^{chn}	{w m p b r f v θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Vowel /aiu/	
Perception	{w} {m p b} {r f v} {θ ð} {y l n k g h} {t d s z ʃ ʒ tʃ dʒ}
Predicted^{ALL}	{m p b} {w r f v} {θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{lip}	{w m r f v} {p b θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{chk}	{w m p b r f v θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{chn}	{w m r f v} {p b θ ð y l n k g h t d} {s z ʃ ʒ tʃ dʒ}

Table B.3 PECs of C/a/ for different talkers.

Sources	PECs
Talker M1	
Perception	{w} {m p b} {r f v} {h} {θ ð y l n k g} {t d s z ʃ ʒ tʃ dʒ}
Predicted^{ALL}	{w} {m p b} {r f v} {h θ ð y l n k g} {t d s z ʃ ʒ tʃ dʒ}
Predicted^{lip}	{w} {m p b} {r f v} {h θ ð y l n k g } {t d s z ʃ ʒ tʃ dʒ}
Predicted^{chk}	{w} {m p b y n k g h} {r f v θ ð l t d s z ʃ ʒ tʃ dʒ}
Predicted^{chn}	{w m p b r f v h y n k g θ ð} {l t d s z ʃ ʒ tʃ dʒ}
Talker F1	
Perception	{w r} {m p b} {f v} {θ ð} {y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{ALL}	{w m p b} {f v} {r θ ð y r l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{lip}	{w r m p b f v θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{chk}	{w m p b} {r f v θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{chn}	{w r m p b f v θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Talker M2	
Perception	{w r} {m p b} {f v} {θ ð} {y l n k g h} {t d s z ʃ ʒ tʃ dʒ}
Predicted^{ALL}	{w r} {m p b} {f v} {θ ð} {y l n t k d g h s z ʃ ʒ tʃ dʒ}
Predicted^{lip}	{w r} {m p b} {f v} {θ ð} {y l n t k d g h s z ʃ ʒ tʃ dʒ}
Predicted^{chk}	{w r} {m p b n} {f v θ ð} {l k g h} {y t d s z ʃ ʒ tʃ dʒ}
Predicted^{chn}	{w m p b r f v} {θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Talker F2	
Perception	{w} {m p b} {r f v} {y l n k g h θ ð t d s z ʃ ʒ tʃ dʒ}
Predicted^{ALL}	{w} {m p b} {r f v} {l} {y k g h θ ð} {n t d s z} {ʃ ʒ tʃ dʒ}
Predicted^{lip}	{w} {r m f v} {p b} {n k g h θ ð} {y l t d s z} {ʃ ʒ tʃ dʒ}
Predicted^{chk}	{w} {p b k g h θ ð} {m r f v y l n t d s z ʃ ʒ tʃ dʒ}
Predicted^{chn}	{w} {b k g h θ ð} {r m f v} {p y l n t d s z ʃ ʒ tʃ dʒ}

Table B.4 PECs of C/i/ for different talkers.

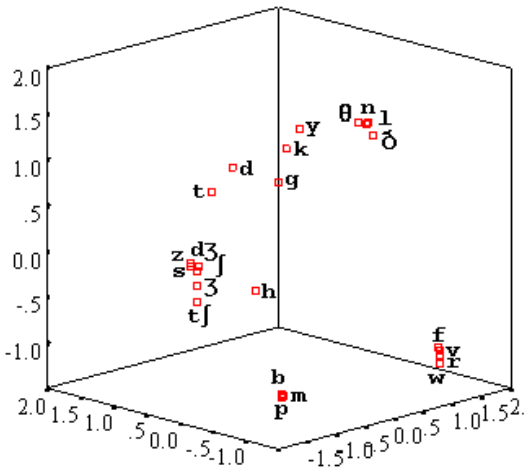
Sources	PECs
Talker M1	
Perception	{w} {m p b} {r f v} {y k g h} {θ ð l n t d} {s z ʃ ʒ tʃ dʒ}
Predicted^{ALL}	{w} {m p b y k g h l n t d} {r f v θ ð} {s z ʃ ʒ tʃ dʒ}
Predicted^{lip}	{w m p b y k g h θ ð l n t d} {r f v} {s z ʃ ʒ tʃ dʒ}
Predicted^{chk}	{w} {m p b r f v y k g h θ ð l n t d} {s z ʃ ʒ tʃ dʒ}
Predicted^{chn}	{w m p b k} {y g h l n t d} {r f v θ ð} {s z ʃ ʒ tʃ dʒ}
Talker F1	
Perception	{w r} {m p b} {f v} {θ ð} {y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{ALL}	{w r} {m f v} {p b θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{lip}	{w r} {m f v} {p b θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{chk}	{w r m p b f v θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{chn}	{w r m f v} {p b θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Talker M2	
Perception	{w r} {m p b} {f v} {θ ð} {y l n t k d g h} {s z ʃ ʒ tʃ dʒ}
Predicted^{ALL}	{w r} {m p b} {f v} {θ ð} {y l n t k d g h} {s z ʃ ʒ tʃ dʒ}
Predicted^{lip}	{w r} {m p b} {f v} {θ ð} {y l n t k d g h} {s z ʃ ʒ tʃ dʒ}
Predicted^{chk}	{w r} {m p b} {f v} {θ ð}{y l n t k d g h} {s z ʃ ʒ tʃ dʒ}
Predicted^{chn}	{w r m p b f v θ ð y l n t k d g h s z ʃ ʒ tʃ dʒ}
Talker F2	
Perception	{w} {m p b} {r f v} {θ ð} {y l n k g h} {t d s z} {ʃ ʒ tʃ dʒ}
Predicted^{ALL}	{w m p b} {r f v θ} {ð y l n k g h t d s z} {ʃ ʒ tʃ dʒ}
Predicted^{lip}	{w} {m p b} {r f v} {θ ð y l n k g h t d} {s z} {ʃ ʒ tʃ dʒ}
Predicted^{chk}	{w m p b r} {ð y l n k g h t d s z} {f v θ ʃ ʒ tʃ dʒ}
Predicted^{chn}	{w m p b θ ð y l n k g h t d} {r f v s z ʃ ʒ tʃ dʒ}

Table B.5 PECs of C/u/ for different talkers.

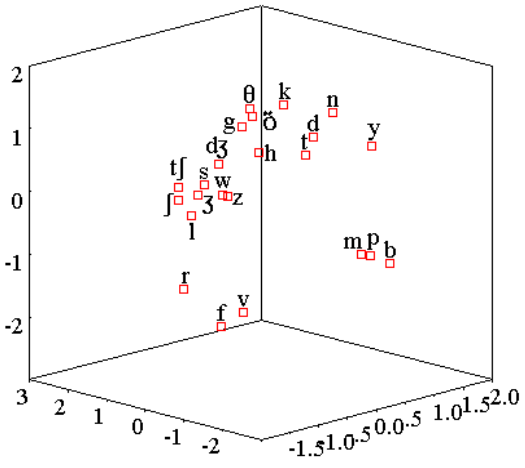
Sources	PECs
Talker M1	
Perception	{w y k g h} {m p b} {r f v} {θ ð l n t d} {s z ʃ ʒ tʃ dʒ}
Predicted^{ALL}	{w y k g h} {m p b} {r f v} {θ l n d} {ð t s z ʃ ʒ tʃ dʒ}
Predicted^{lip}	{w m p b y l n k g h θ} {r f v ð t d s z ʃ ʒ tʃ dʒ}
Predicted^{chk}	{w m p b y l n k g h θ} {r f v ð t d s z ʃ ʒ tʃ dʒ}
Predicted^{chn}	{w p b} {r f v} {θ y l m n k g h} {ð t d s z ʃ ʒ tʃ dʒ}
Talker F1	
Perception	{w r m p b} {f v} {θ ð} {y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{ALL}	{w r m h} {f v} {p b θ ð y l n k g t d s z ʃ ʒ tʃ dʒ}
Predicted^{lip}	{w r m} {f v} {p b θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{chk}	{f v}{w r m p b θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{chn}	{r f v}{w m p b θ ð y l n k g h t d s z ʃ ʒ tʃ dʒ}
Talker M2	
Perception	{w r} {m p b} {f v} {θ ð} {y l n k g h} {t d s z ʃ ʒ tʃ dʒ}
Predicted^{ALL}	{m p b} {f v θ ð l} {w r y n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{lip}	{w r m p b θ ð y l n k g h d} {f v t s z ʃ ʒ tʃ dʒ}
Predicted^{chk}	{w r m p b y n k g h t d s z ʃ ʒ tʃ dʒ} {f v θ ð l}
Predicted^{chn}	{w m p b y n k g h} {r f v θ ð l} {t d s z ʃ ʒ tʃ dʒ}
Talker F2	
Perception	{m p b} {r f v} {θ ð} {w y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{ALL}	{m p b} {r f v} {θ ð} {w y l n k g h} {t d s z ʃ ʒ tʃ dʒ}
Predicted^{lip}	{w m p b θ ð} {r f v} {y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{chk}	{m p b r f v θ ð w y l n k g h t d s z ʃ ʒ tʃ dʒ}
Predicted^{chn}	{m p b r f v θ ð w y l n k g h t d s z ʃ ʒ tʃ dʒ}

APPENDIX C. 3-D MULTIDIMENSIONAL SCALING ANALYSES

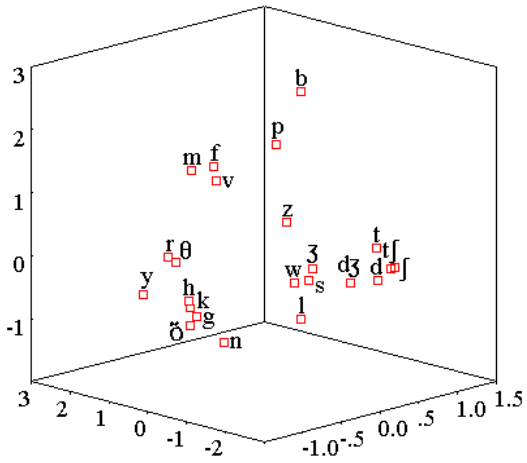
In this appendix, we present figures of a 3-D multidimensional scaling (MDS) analysis of the confusion dissimilarity matrices from Appendix A and the corresponding matrices predicted from all, lips, cheeks, and chin retro-reflectors. Each figure is labeled in the format of *V_T_src*, where *V* can be /a/, /i/, /u/, or /aiu/; *T* can be M1, M2, F1, or F2; *src* can be perception, ALL (all retro-reflectors), lip, chk (cheeks), or chn (chin). If *T* is absent, then the results were pooled across all talkers.



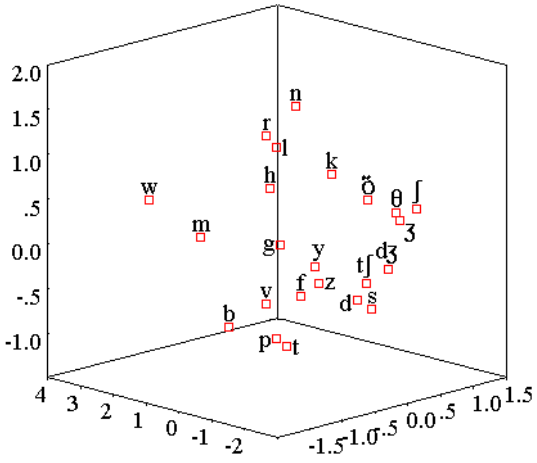
/a/_M1_perception



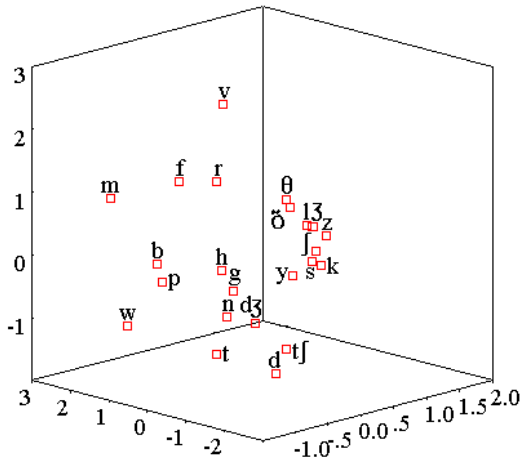
/a/_M1_ALL



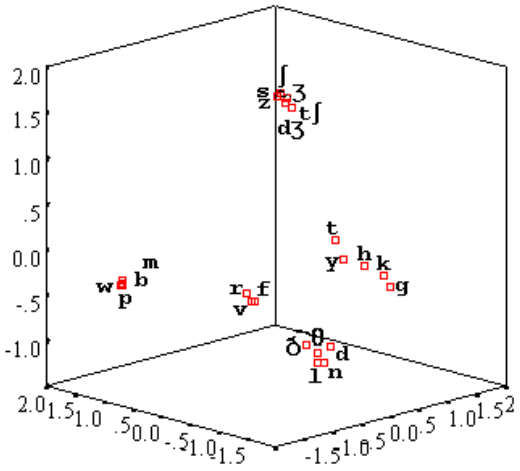
/a/_M1_lip



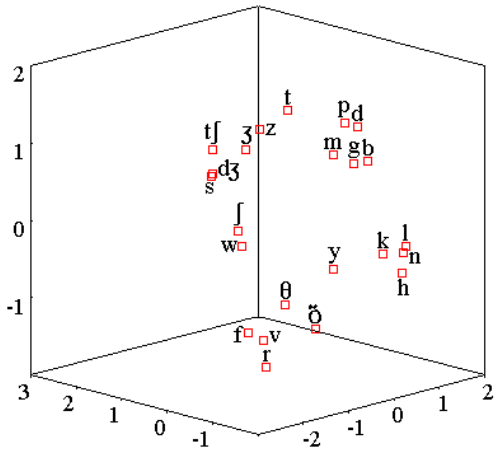
/a/_M1_chk



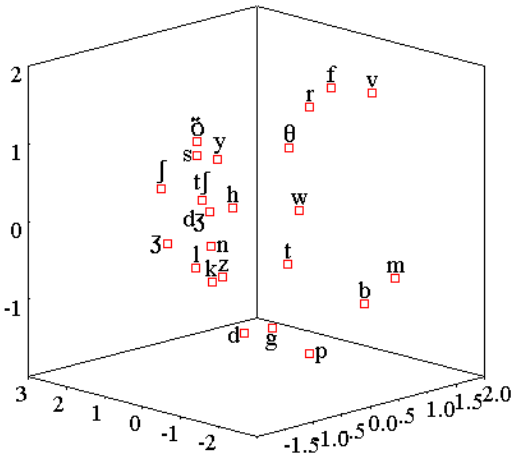
/a/_M1_chn



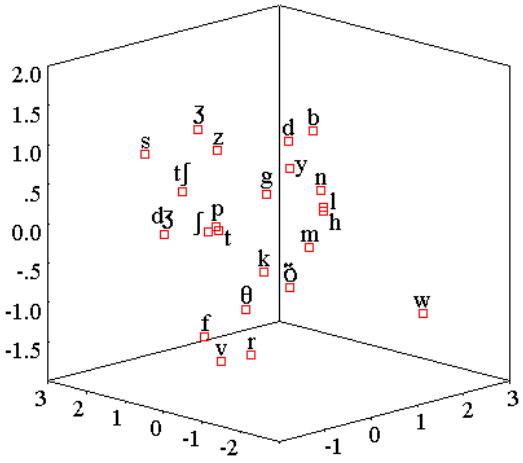
/i/_M1_perception



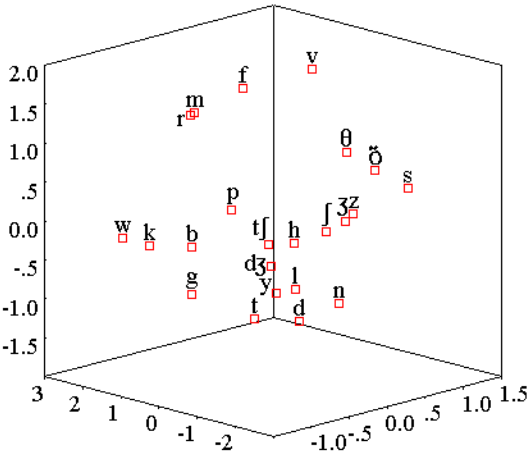
/i/_M1_ALL



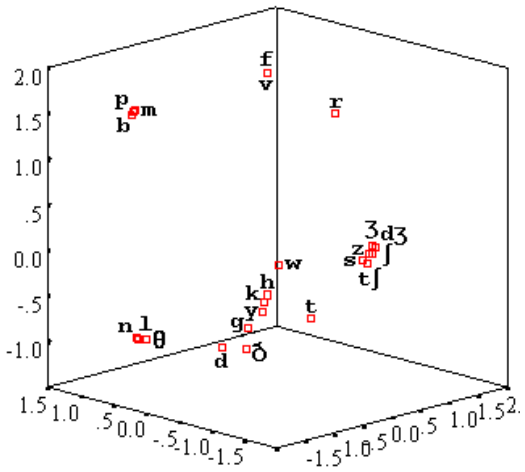
/i/_M1_lip



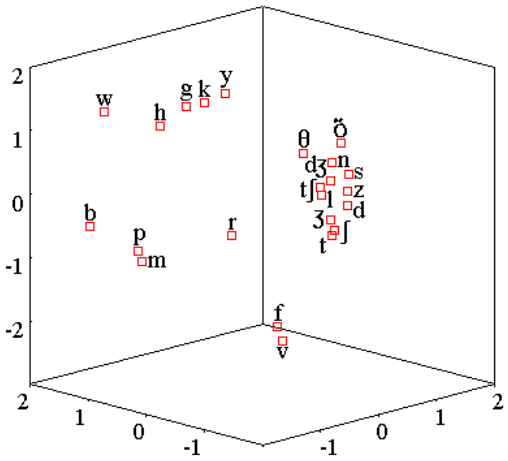
/i/_M1_chk



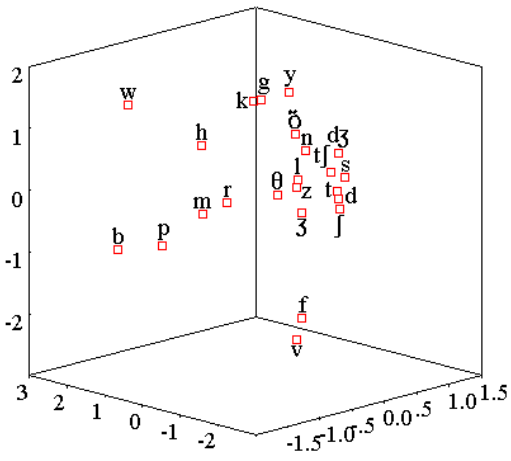
/i/_M1_chn



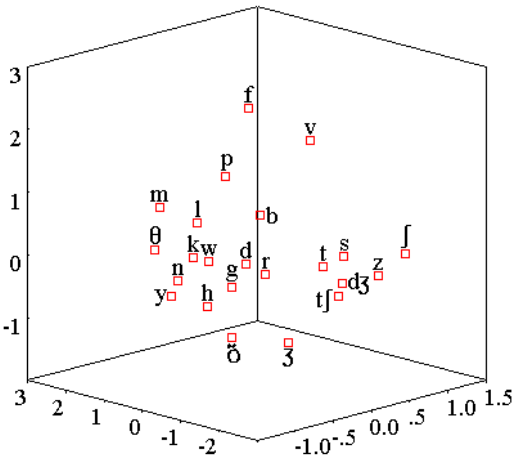
/u/_M1_perception



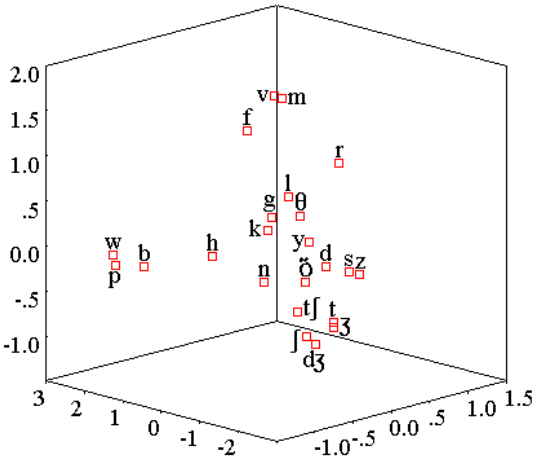
/u/_M1_ALL



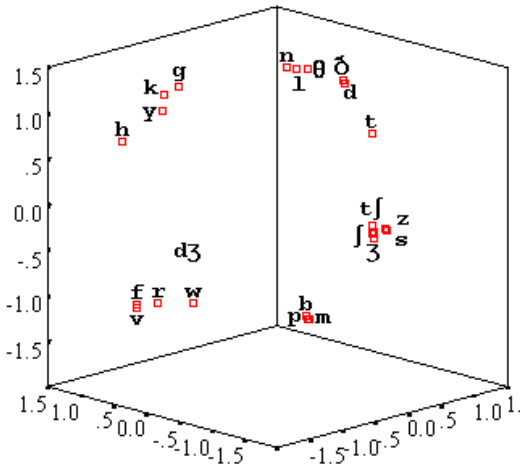
/u/_M1_lip



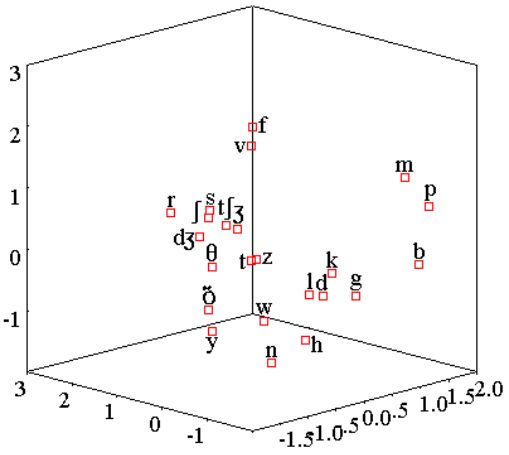
/u/_M1_chk



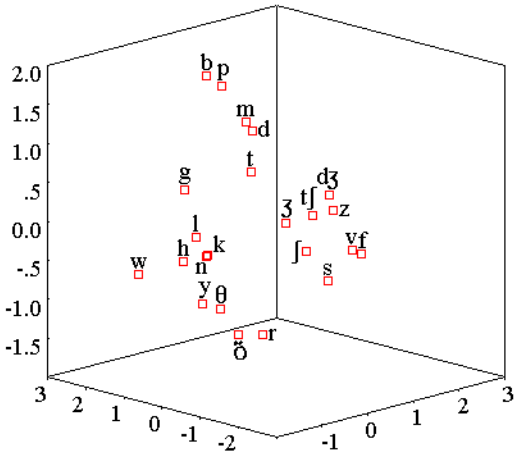
/u/_M1_chn



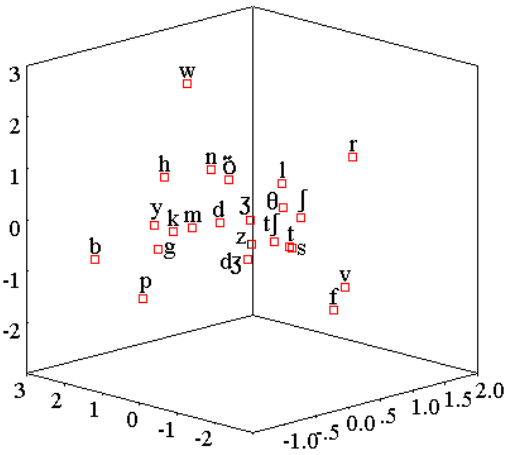
/aiu/_M1_perception



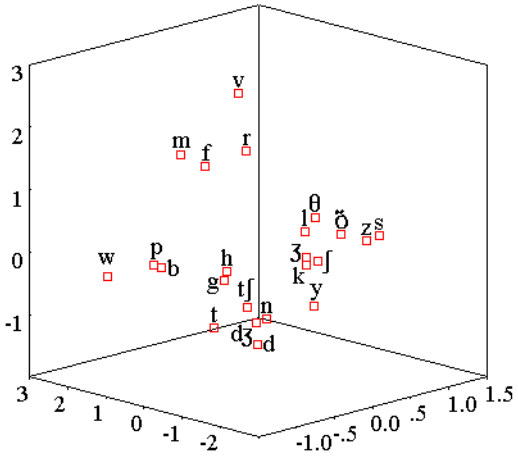
/aiu/_M1_ALL



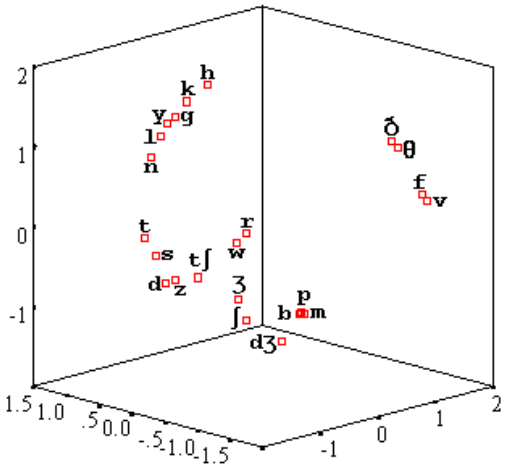
/aiu/_M1_lip



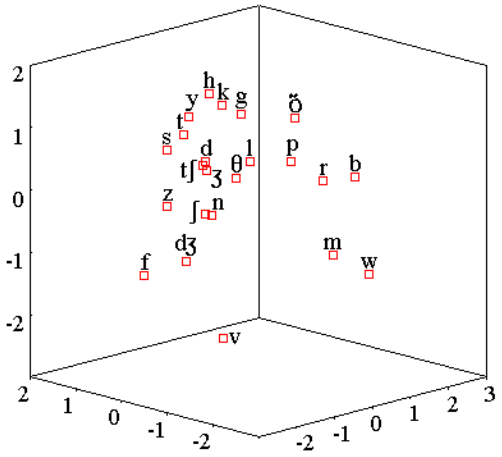
/aiu/_M1_chk



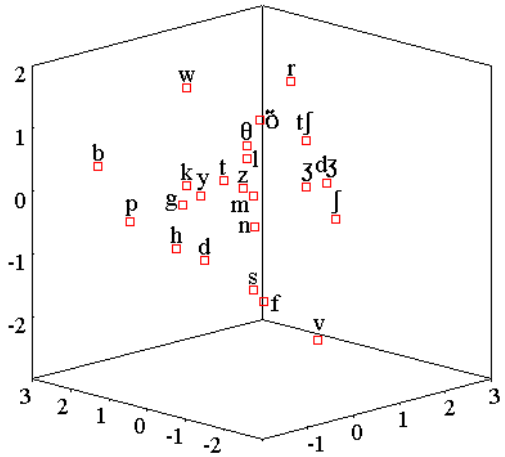
/aiu/_M1_chn



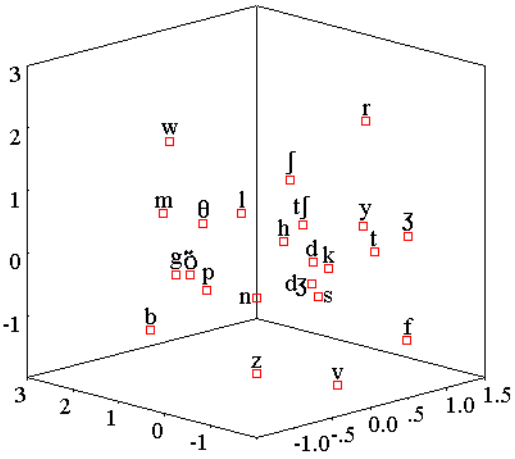
/a/_F1_perception



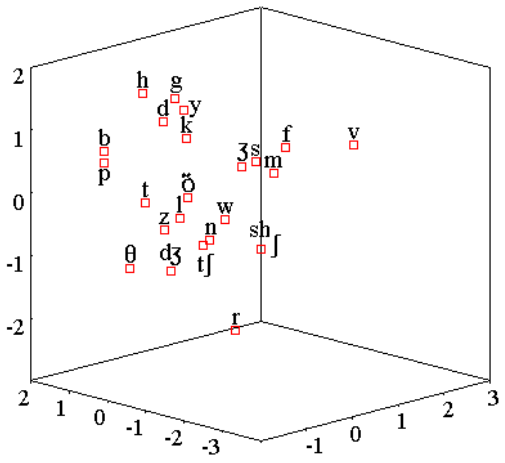
/a/_F1_ALL



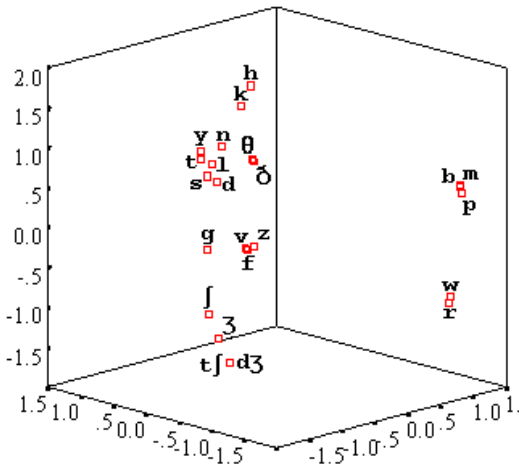
/a/_F1_lip



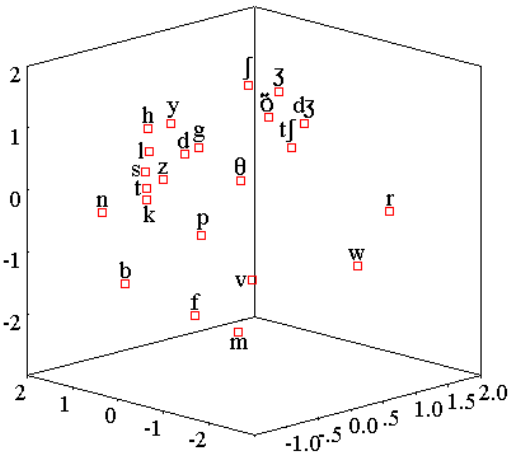
/a/_F1_chk



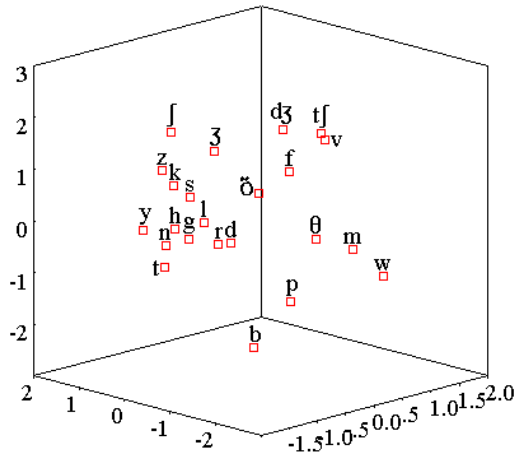
/a/_F1_chn



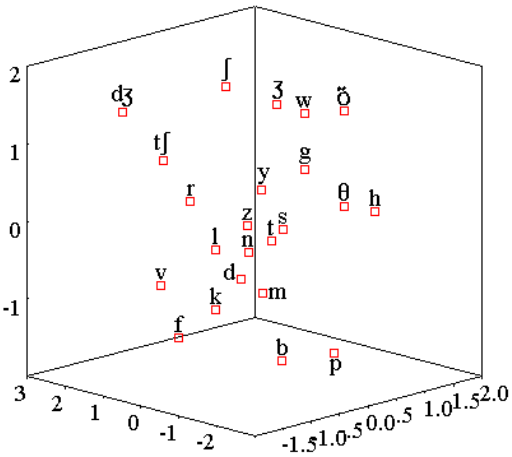
/i/_F1_perception



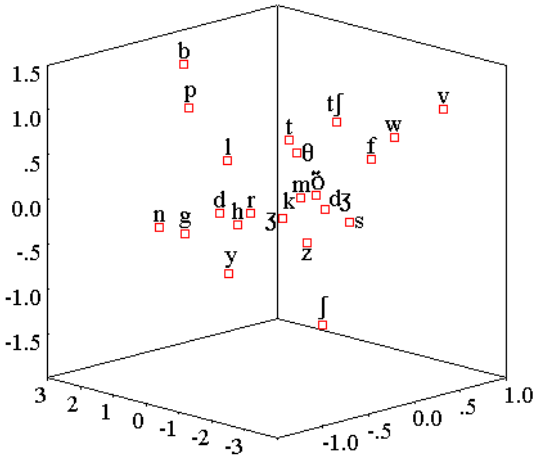
/i/_F1_ALL



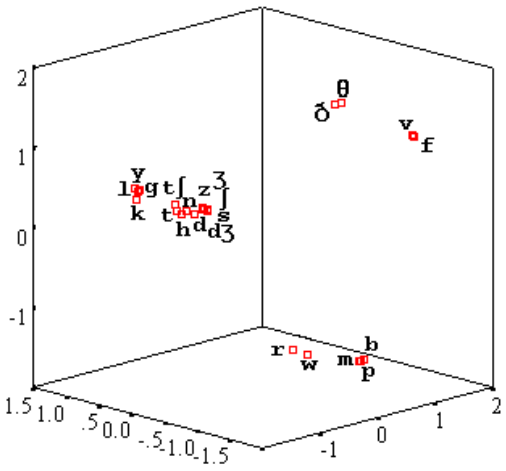
/i/_F1_lip



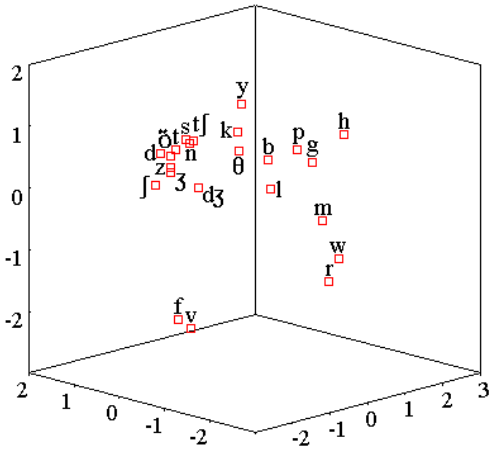
/i/_F1_chk



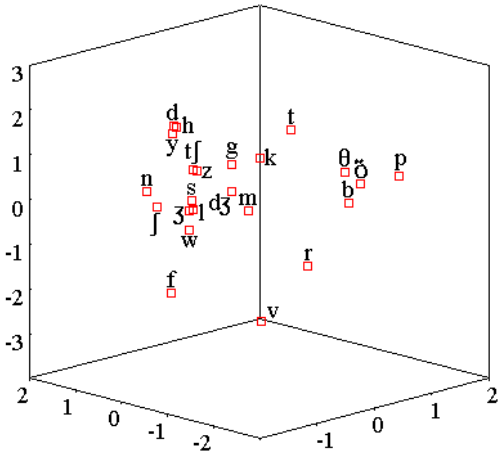
/i/_F1_chn



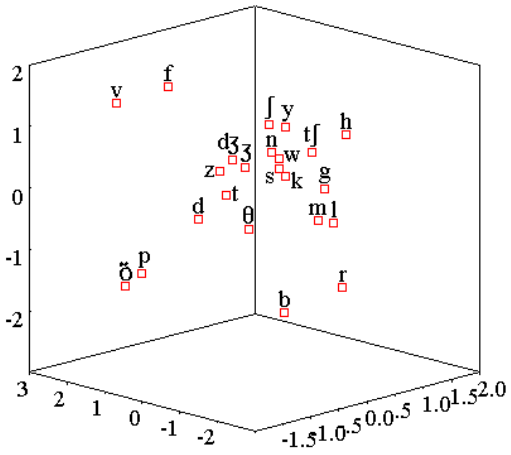
/u/_F1_perception



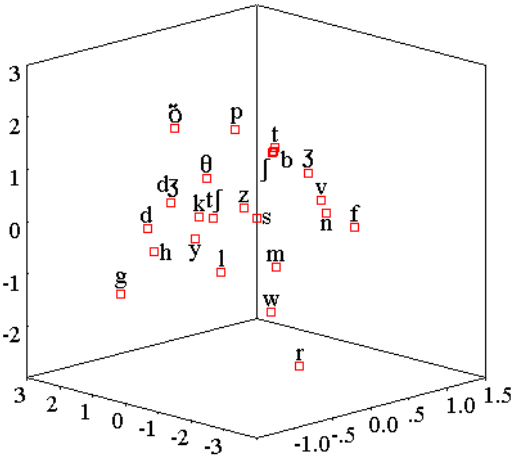
/u/_F1_ALL



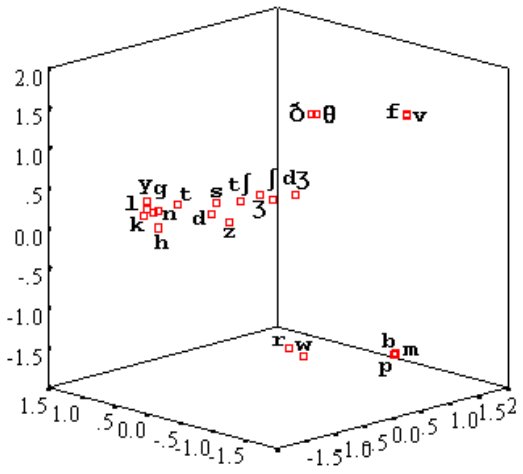
/u/_F1_lip



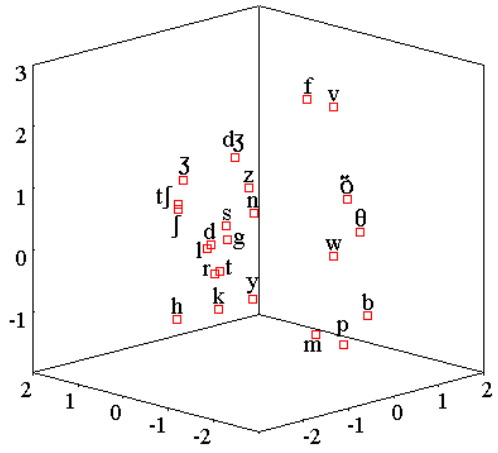
/u/_F1_chk



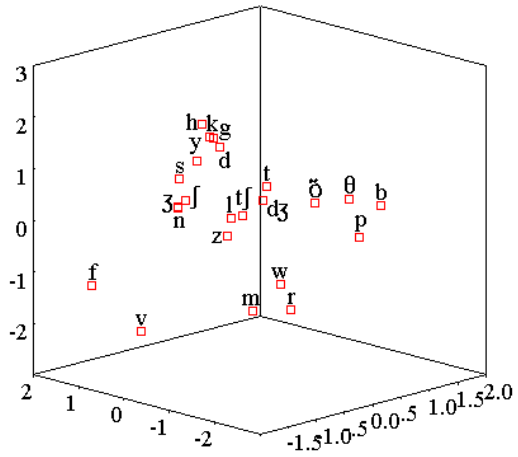
/u/_F1_chn



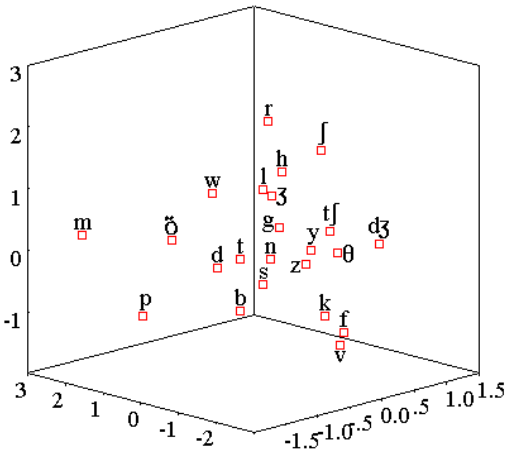
/aiu/_F1_perception



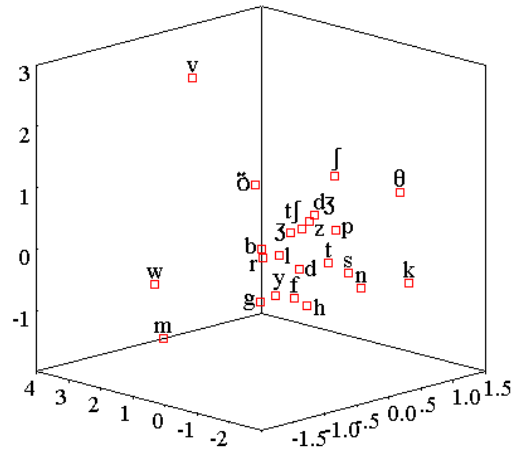
/aiu/_F1_ALL



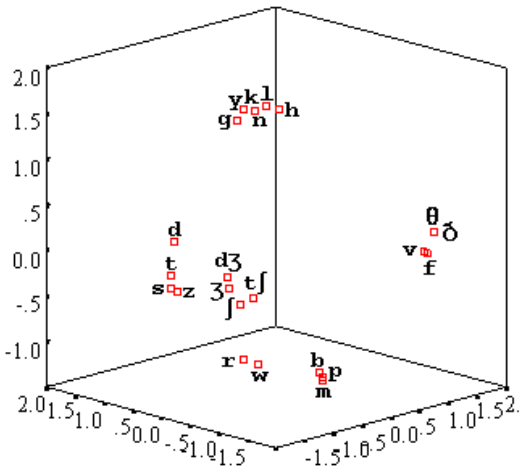
/aiu/_F1_lip



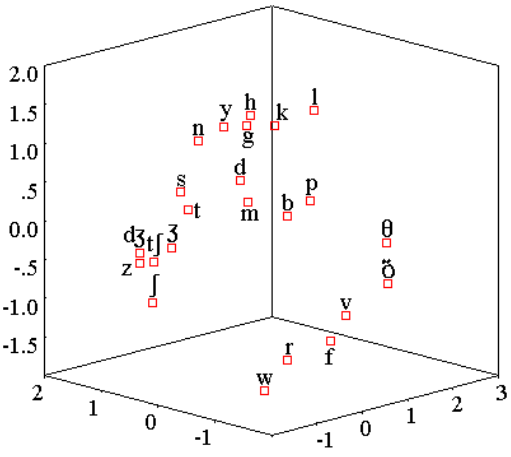
/aiu/_F1_chk



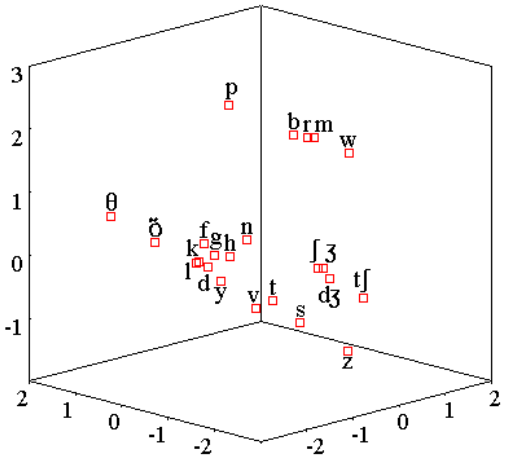
/aiu/_F1_chn



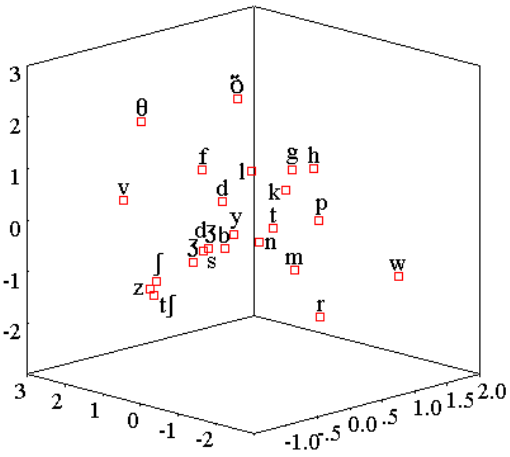
/a/_M2_perception



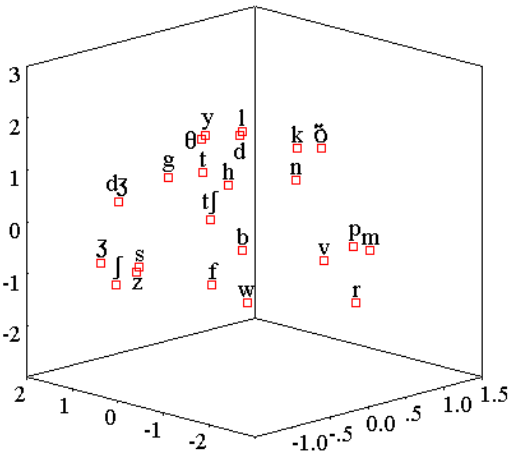
/a/_M2_ALL



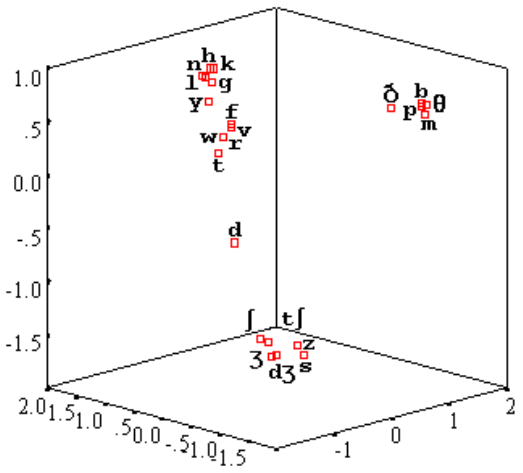
/a/_M2_lip



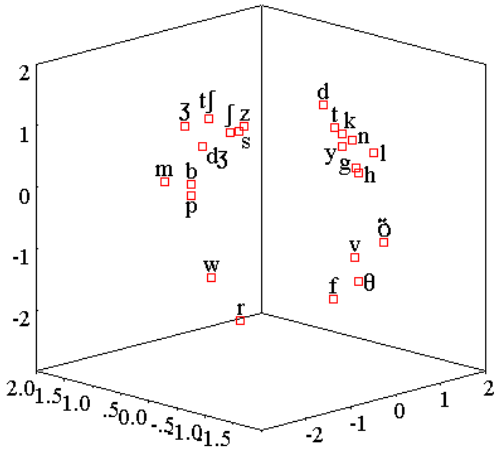
/a/_M2_chk



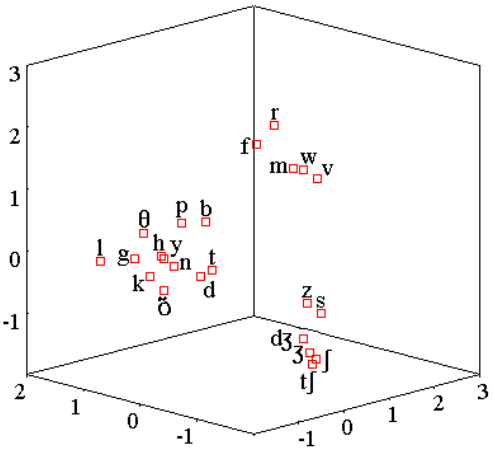
/a/_M2_chn



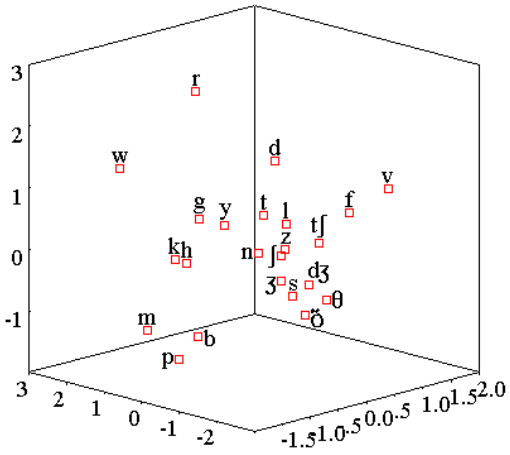
/i/_M2_perception



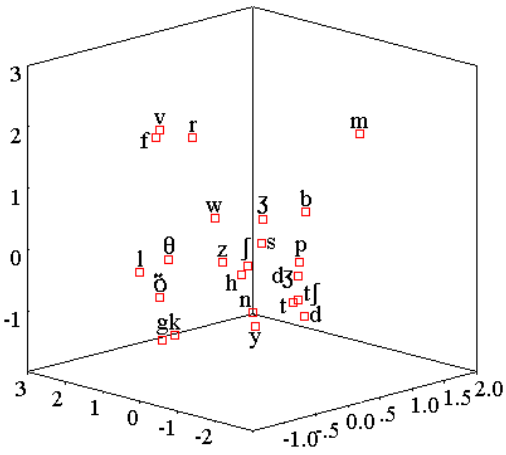
/i/_M2_ALL



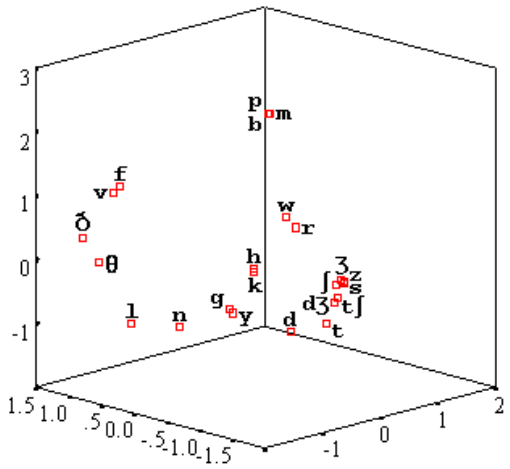
/i/_M2_lip



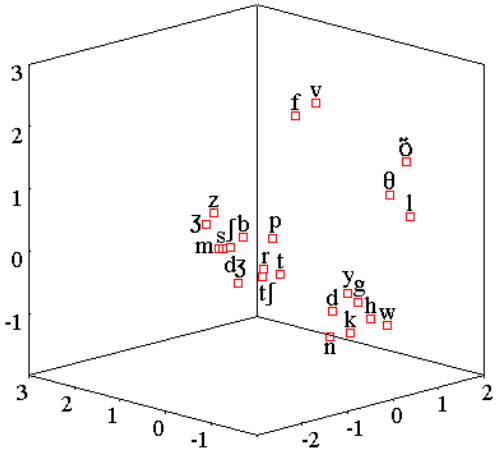
/i/_M2_chk



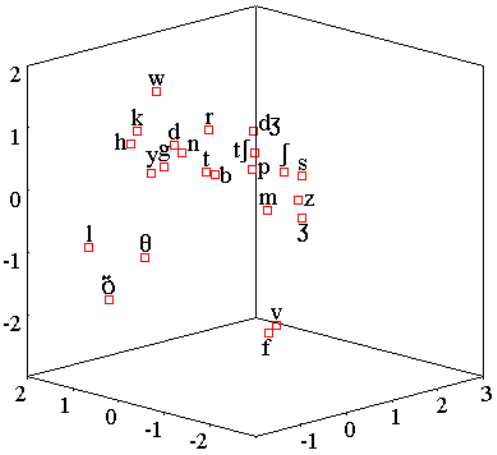
/i/_M2_chn



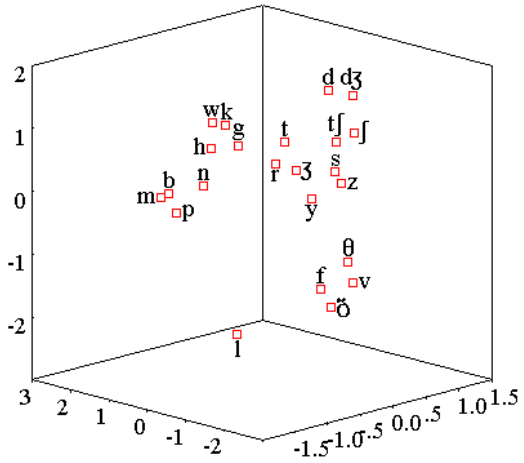
/u/_M2_perception



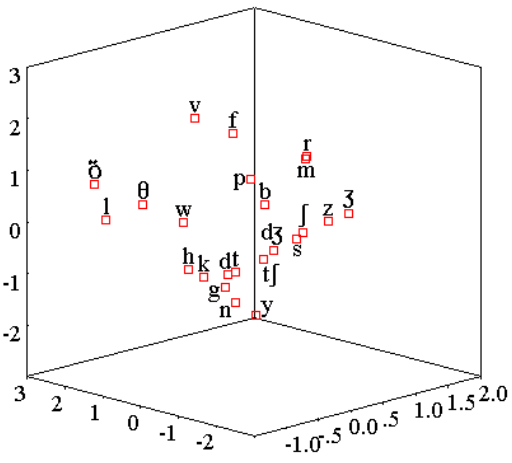
/u/_M2_ALL



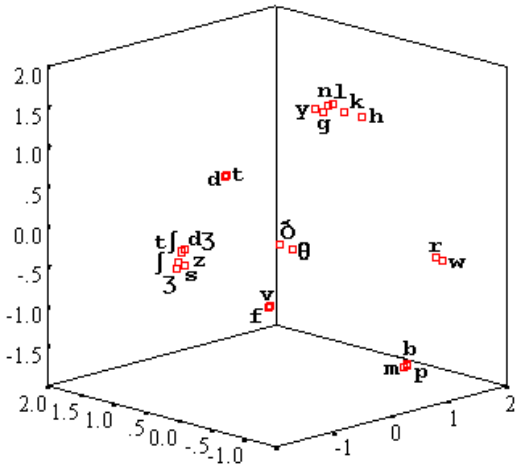
/u/_M2_lip



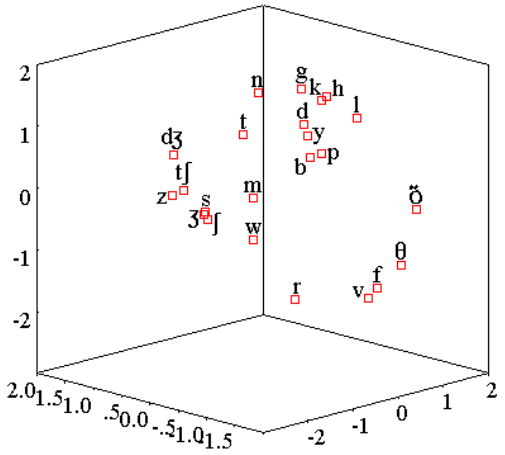
/u/_M2_chk



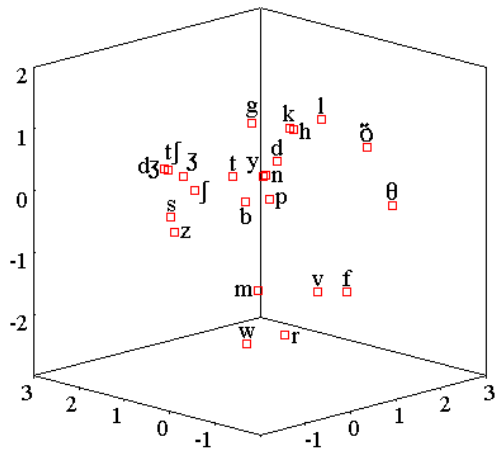
/u/_M2_chn



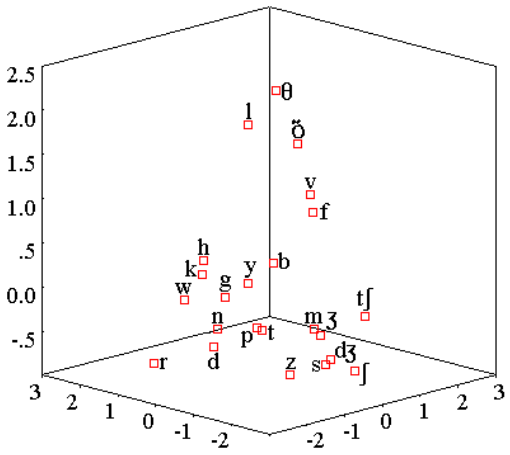
/aiu/_M2_perception



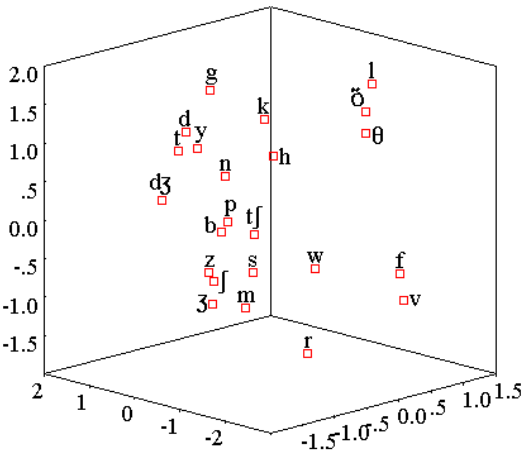
/aiu/_M2_ALL



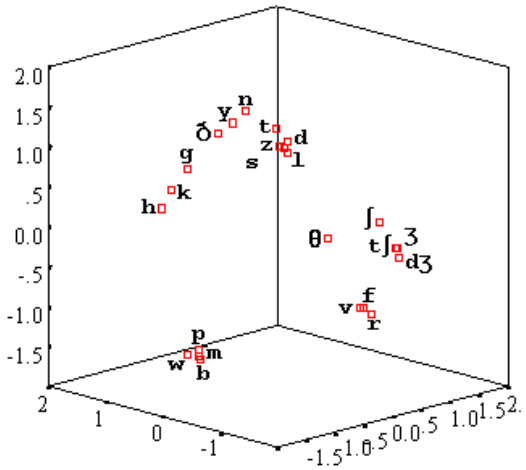
/aiu/_M2_lip



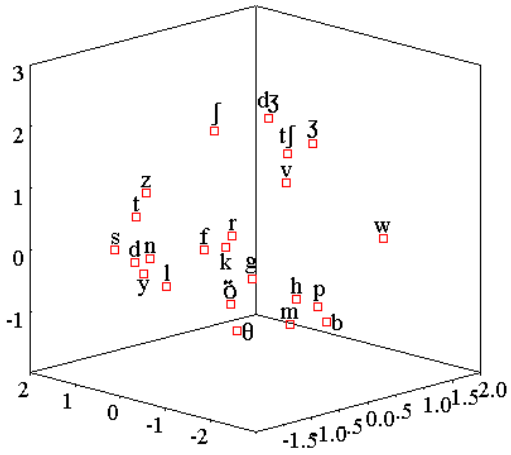
/aiu/_M2_chk



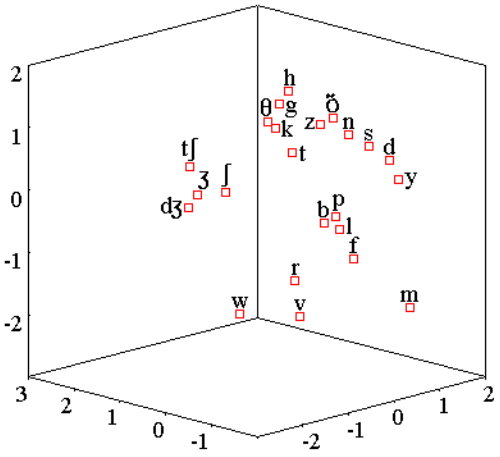
/aiu/_M2_chn



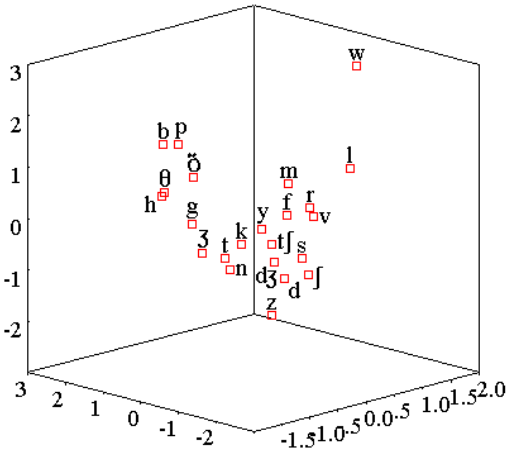
/a/_F2_perception



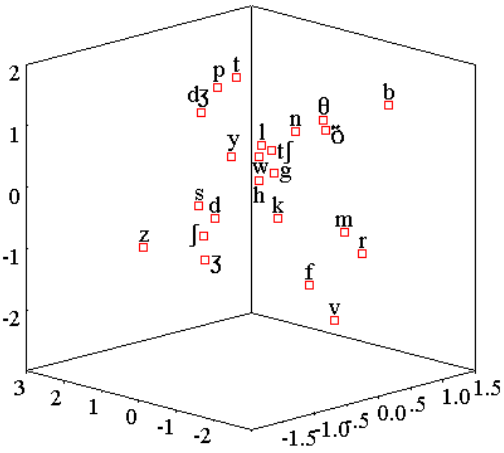
/a/_F2_ALL



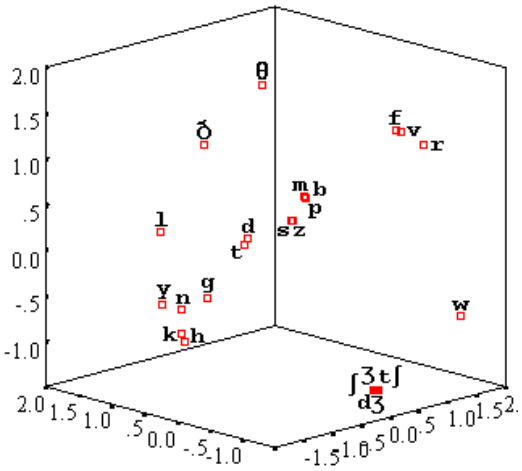
/a/_F2_lip



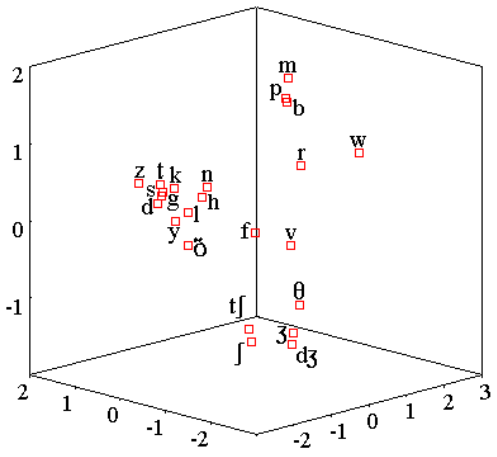
/a/_F2_chk



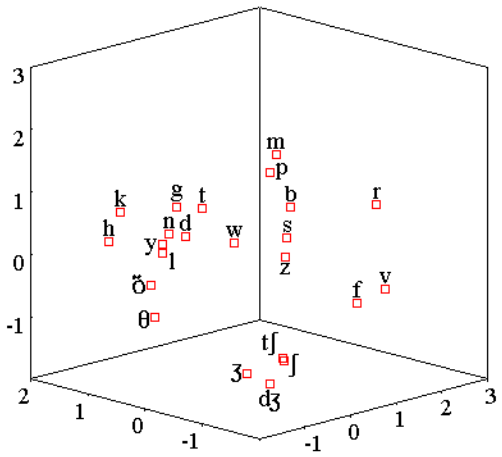
/a/_F2_chn



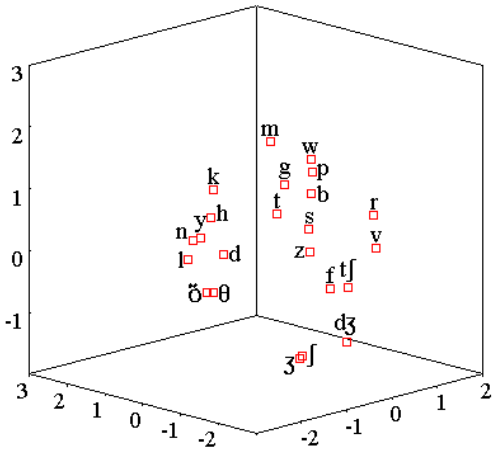
/i/_F2_perception



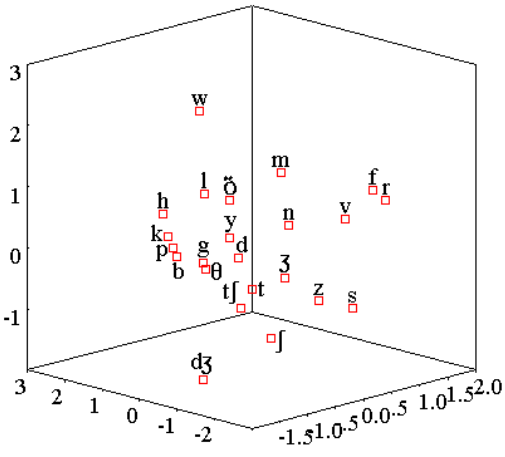
/i/_F2_ALL



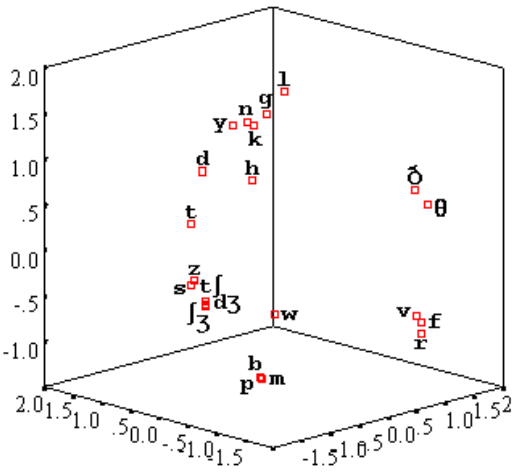
/i/_F2_lip



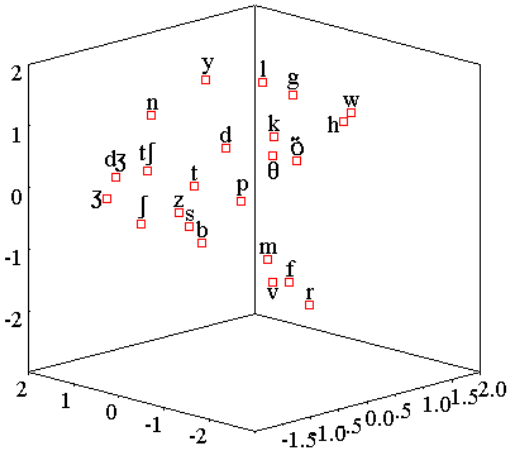
/i/_F2_chk



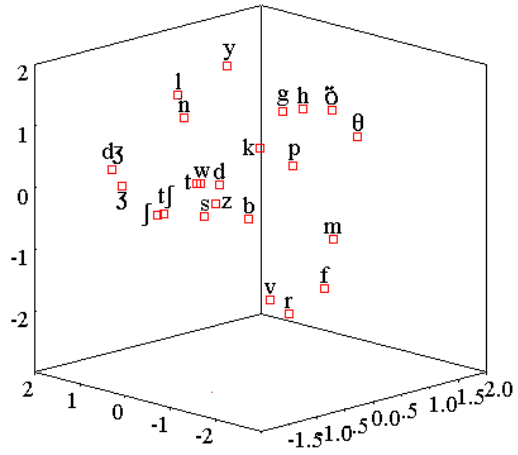
/i/_F2_chn



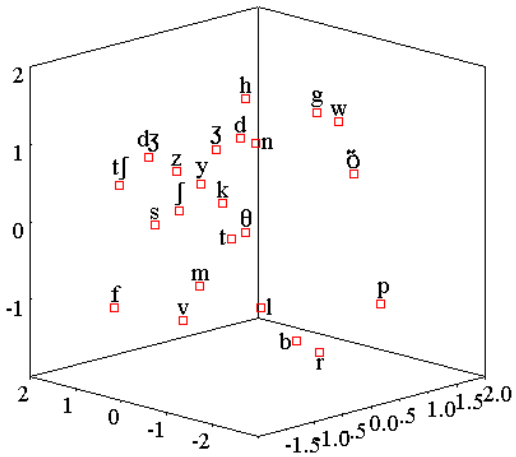
/u/_F2_perception



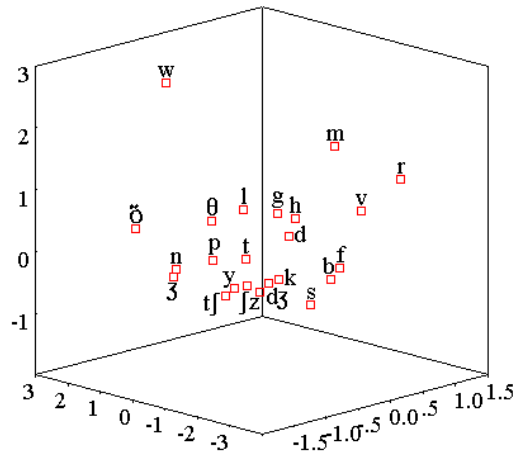
/u/_F2_ALL



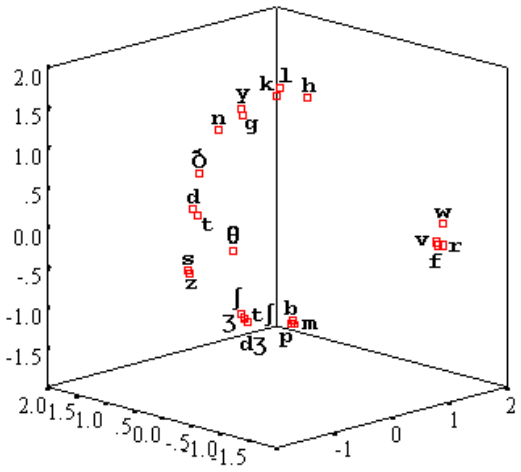
/u/_F2_lip



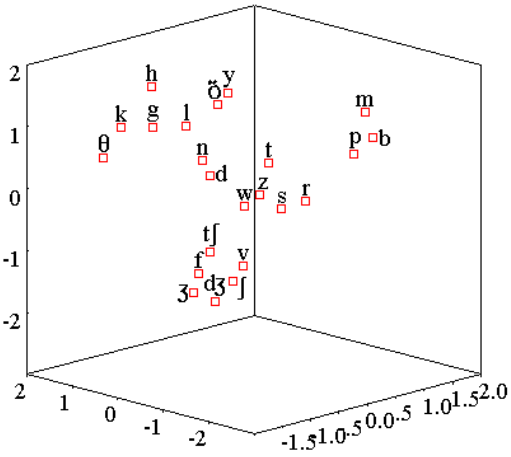
/u/_F2_chk



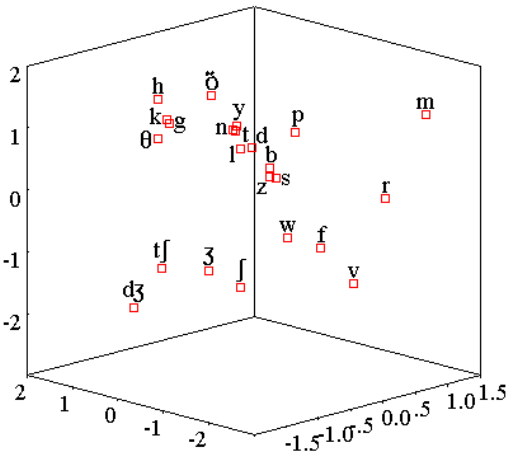
/u/_F2_chn



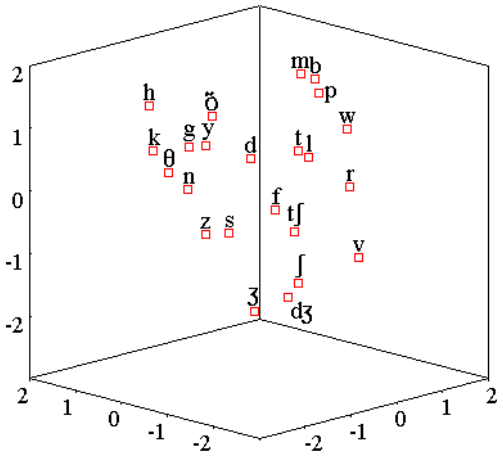
/aiu/_F2_perception



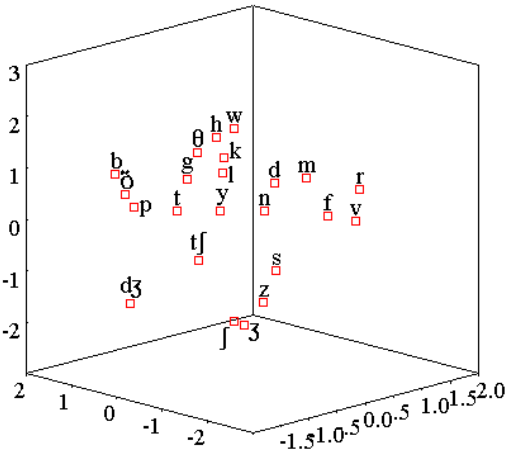
/aiu/_F2_ALL



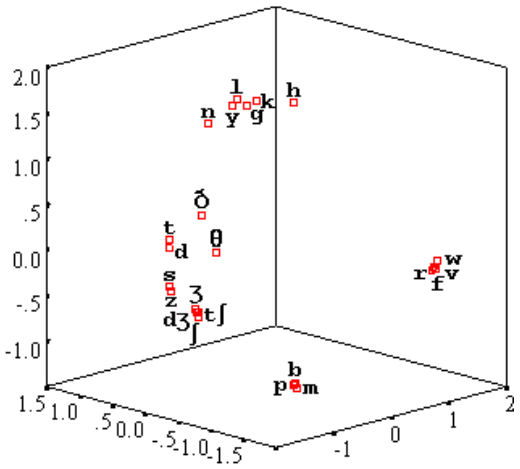
/aiu/_F2_lip



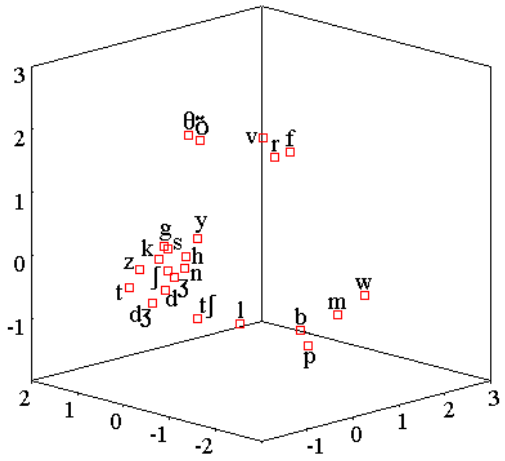
/aiu/_F2_chk



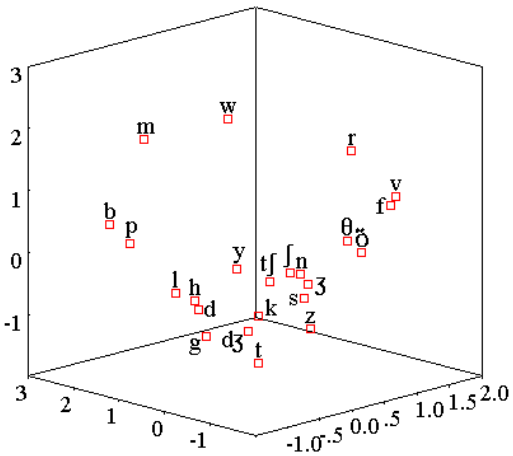
/aiu/_F2_chn



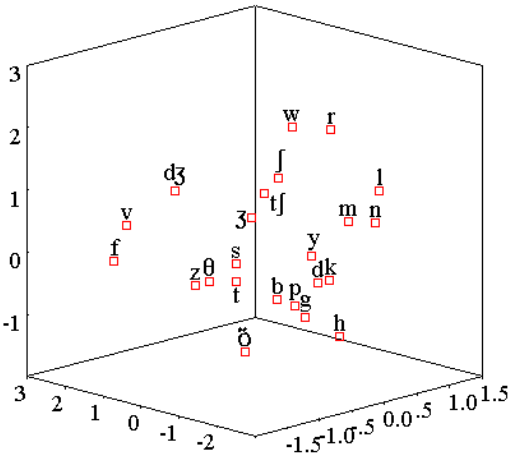
/a/_perception



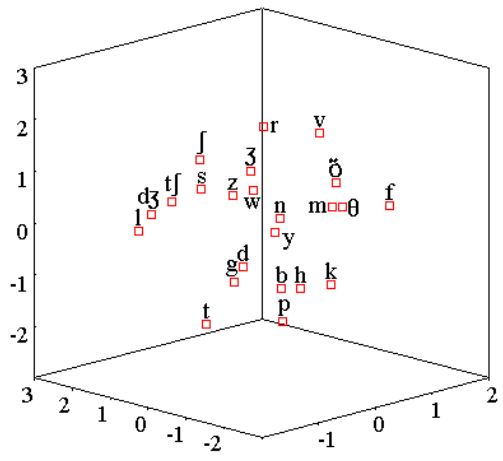
/a/_ALL



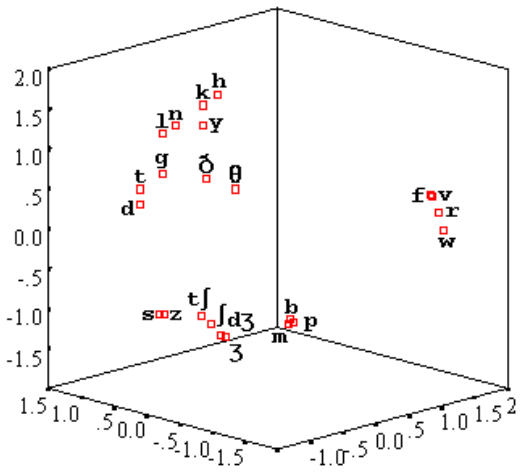
/a/_lip



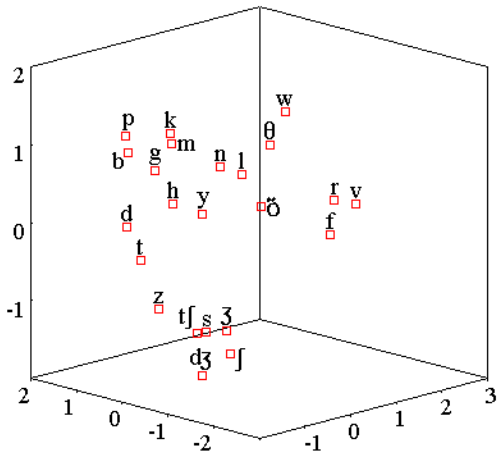
/a/_chk



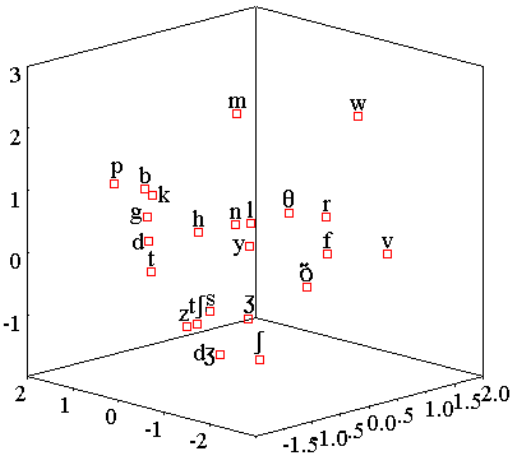
/a/_chn



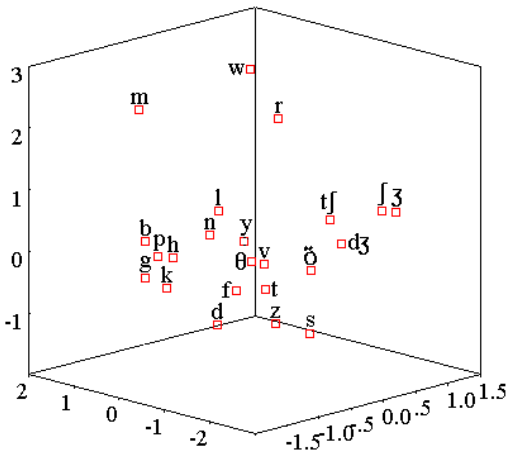
/i/_perception



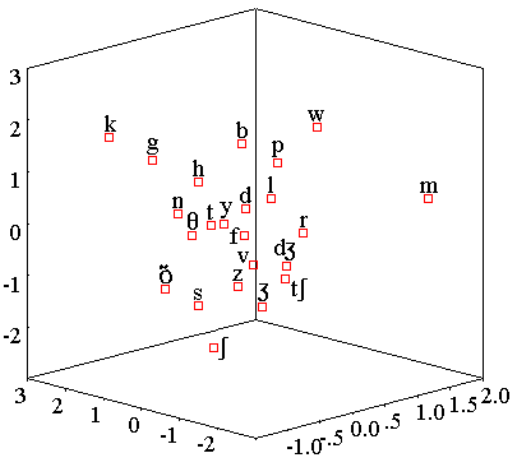
/i/_ALL



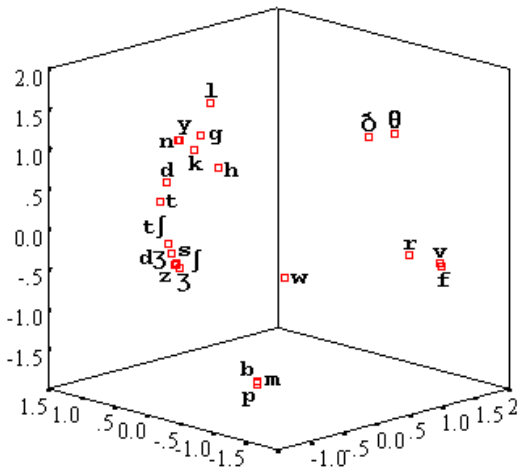
/i/_lip



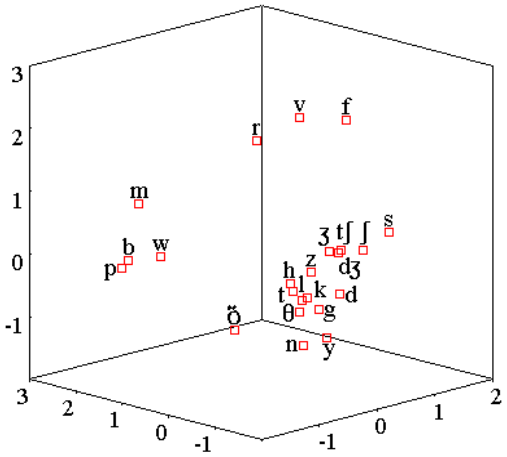
/i/_chk



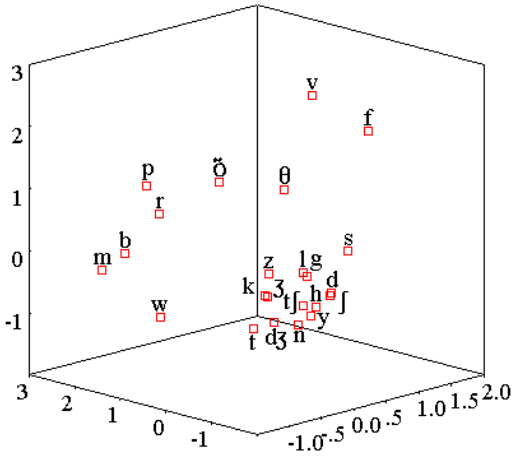
/i/_chn



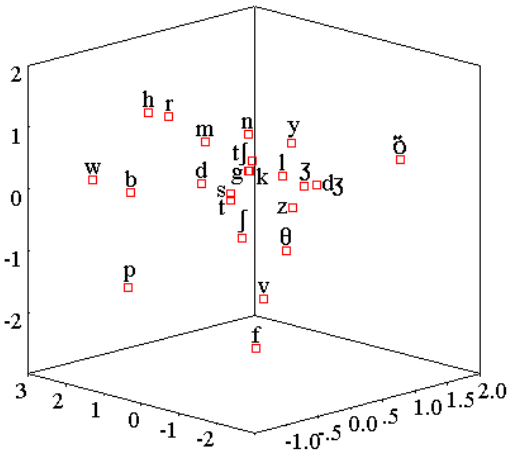
/u/_perception



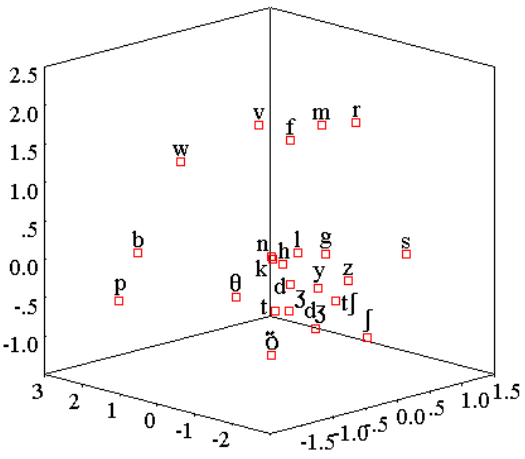
/u/_ALL



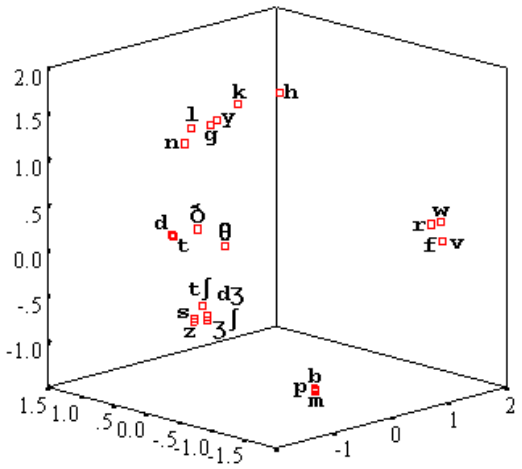
/u/_lip



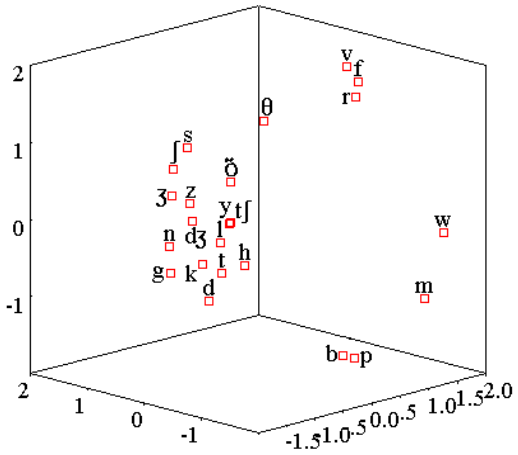
/u/_chk



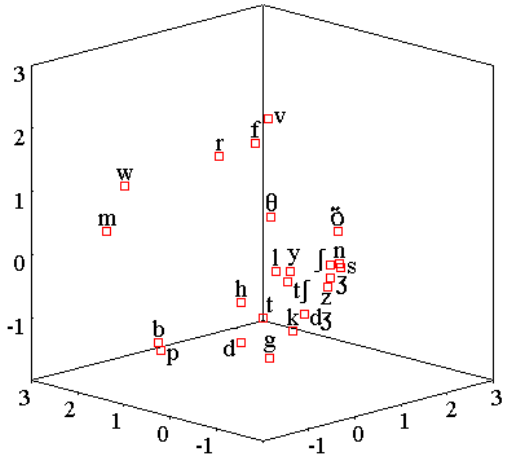
/u/_chn



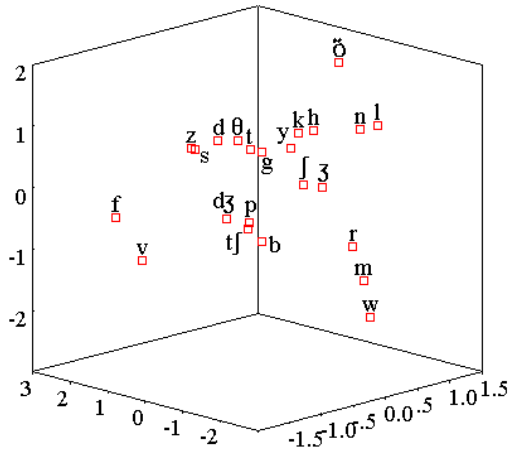
/aiu/_perception



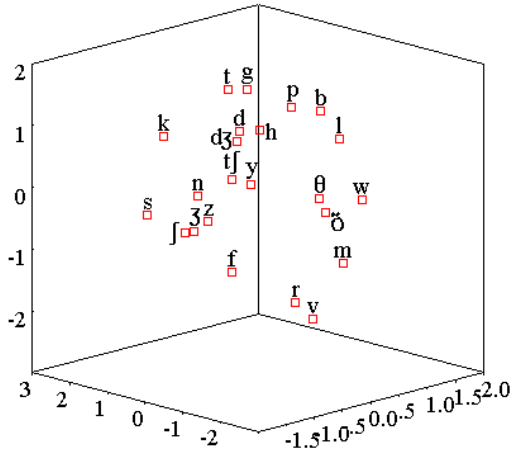
/aiu/_ALL



/aiu/_lip



/aiu/_chk



/aiu/_chn

BIBLIOGRAPHY

Agelfors, E., Beskow, J., Granström, B., Lundeberg, M., and Salvi, G. (1999). "Synthetic visual speech driven from auditory speech," *Proc. AVSP*, Santa Cruz, CA, 123-127.

Aldenderfer, M.S. and Blashfield, R.K. (1984). *Cluster analysis*. Beverly Hills and London: Sage Pubns.

Alwan A., Narayanan, S., and Haker, K. (1997). "Towards articulatory-acoustic models of liquid consonants. Part II: the rhotics," *J. Acoust. Soc. Am.*, 101(2): 1078-1089.

Atal, B.S., Chang, J.J., Mathews, M.V., and Tukey, J.W. (1978). "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique," *J. Acoust. Soc. Am.*, 63: 1535-1556.

Atal, B.S. and Rioul, O. (1989). "Neural networks for estimating articulatory positions from speech," *J. Acoust. Soc. Am.*, 86: 123-131.

Auer, E.T. and Bernstein, L.E. (1997). "Speechreading and the structure of the lexicon: computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness," *J. Acoust. Soc. Am.*, 102(6): 3704-3710.

Badin, P., Beautemps, D., Laboissiere, R., and Schwartz, J.L. (1995). "Recovery of vocal tract geometry from formants for vowels and fricative consonants using a midsagittal-to-area function conversion model," *J. Phonetics*, 23: 221-229.

Bailly, G. (2001). "Audiovisual speech synthesis," *Proc. 4th ETRW on Speech Synthesis*, Perthshire, Scotland.

Bailly, G. (2002). "Audiovisual speech synthesis. From ground truth to models," *Proc. ICSLP*, Denver, CO, 1453-1456.

Bangayan, P., Alwan, A., and Narayanan, S. (1996). "From MRI and acoustic data to articulatory synthesis: a case study of the laterals," *Proc. ICSLP*, Philadelphia, PA, 793-796.

Barbosa, A.V. and Yehia, H.C. (2001). "Measuring the relation between speech acoustics and 2D facial motion," *Proc. ICASSP*, Salt Lake City, UT.

Barker, J.P. and Berthommier, F. (1999). "Estimation of speech acoustics from visual speech features: a comparison of linear and non-linear models," *Proc. AVSP*, Santa Cruz, CA, 112-117.

Benguerel, A.P. and Pichora-Fuller, M.K. (1982). "Coarticulation effects in lipreading," *J. Speech Hear. Res.*, 25: 600-607.

Benkí, J.R. (2001). "Place of articulation and first formant transition pattern both affect perception of voicing in English," *J. Phonetics*, 29: 1-22.

Benoît, C., Guiard-Marigny, T., Le Goff, B., and Adjoudani, A. (1996). "Which components of the face do humans and machines best speechread?" in D.G. Stork and M.E. Hennecke (Eds.), *Speechreading by Humans and Machines* (pp. 315-328), New York: Springer-Verlag.

Benoît, C., Lallouache, M.T., Mohamadi, T., and Abry, C. (1992). "A set of French visemes for visual speech synthesis," in G. Bailly and C. Benoît (Eds.), *Talking Machines: Theories, Models, and Designs* (pp. 485-504), Elsevier Science.

Bergeron, P. and Lachapelle, P. (1985). "Techniques for Animating Characters," *SIGGRAPH Tutorial: Advanced Computer Graphics Animation*, 2: 61-79.

Bernstein, L.E., Auer, E.T., Chaney, B., Alwan, A., and Keating, P.A. (2000a). "Development of a facility for simultaneous recordings of acoustic, optical (3-D motion and video), and physiological speech data," *J. Acoust. Soc. Am.*, 107: 2887.

Bernstein, L.E., Demorest, M.E., and Tucker, P.E. (2000b). "Speech perception without hearing," *Perception and Psychophysics*, 62(2): 233-252.

Bertsimas, D. and Tsitsiklis, J.N. (1997). *Introduction to linear optimization*. Athena Scientific.

Binnie, C.A., Montgomery, A.A., and Jackson, P.L. (1974). "Auditory and visual contributions to the perception of consonants," *J. Speech Hear. Res.*, 17: 619-630.

Blough, D.S. (2001). "The Perception of Similarity," in R.G. Cook (Ed.), *Avian visual cognition* [On-line]. Available: http://www.pigeon.psy.tufts.edu/avc/.

Bregler, C., Covell, M., and Slaney, M. (1997). "Video rewrite: visual speech synthesis from video," *Proc. AVSP*, Rhodes, Greece, 153-156.

Carstens Medizinelektronik GmbH (1993). *Articulograph AG100 User's Handbook*.

Cattell, R.B. (1966). "The scree test for the number of factors," *Multivariate Behavioral Research*, 1: 245-276.

Chan, M.T., Zhang, Y., and Huang, T.S. (1998). "Real-time lip tracking and bimodal continuous speech recognition," *Proc. IEEE 2nd Workshop on Multimedia Signal Processing*, Los Angeles, CA, 65-70.

Chen, M. and Alwan, A. (2001). "On the perception of voicing for plosives in noise," *Proc. EUROSPEECH*, Aalborg, Denmark, 1: 175-178.

Chen, W. and Alwan, A. (2000). "Place of articulation cues for voiced and voiceless plosives and fricatives in syllable-initial position," *Proc. ICSLP*, Beijing, China, 4: 113-116.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English*. New York: Harper and Row.

Cohen, M.M. and Massaro, D.W. (1993). "Modeling coarticulation in synthetic visual speech," in N.M. Thalmann and D. Thalmann (Eds.), *Models and Techniques in Computer Animation* (pp. 141-155), Tokyo: Springer-Verlag.

Cohen, M.M., Walker, R.L., and Massaro, D.W. (1996). "Perception of synthetic visual speech," in D.G. Stork and M.E. Hennecke (Eds.), *Speechreading by Humans and Machines*, Springer-Verlag.

Cooper, F.S., Delattre, P.C., Liberman, A.M., Borst, J.M., and Gerstman, L.J. (1952). "Some experiments on the perception of synthetic speech sounds," *J. Acoust. Soc. Am.*, 24: 597-606.

Cox, R.M., Matthews, I., and Bangham, A. (1997). "Combining noise compensation with visual information in speech recognition," *Proc. AVSP*, Rhodes, Greece, 53-56.

DiPaola, S. (1989). "Implementation and use of a 3D parameterized facial modeling and animation system," *ACM SIGGRAPH Course Notes*, 22: 18-33.

DiPaola, S. (1991). "Extending the range of facial types," *J. Visualization and Computer Animation*, 2(4): 129-131.

Dusan, S. and Deng, L. (1998). "Recovering vocal tract shapes from MFCC parameters," *Proc. ICSLP*, Sydney, Australia.

Efron, B. (1982). "The Jacknife, the bootstrap, and other resampling plans," *CBMS-NSF Regional Conference Series in Applied Mathematics 38*, Society for Industrial and Applied Mathematics, Philadelphia, PA.

Erber, N.P. (1975). "Auditory-visual perception of speech," *J. Speech Hear. Disorder*, 40: 481-492.

Eriksen, C.W. and Hake, H.W. (1955). "Multidimensional stimulus differences and accuracy of discrimination," *Psychological Review*, 67: 79-300.

Ezzat, T. and Poggio, T. (1998). "MikeTalk: A Talking Facial Display Based on Morphing Visemes," *Proc. Computer Animation Conference*, Philadelphia, PA.

Fant, G. (1960). *The acoustic theory of speech production*. S-Gravenhage: Mouton.

Fisher, C.G. (1968). “Confusion among visually perceived consonants,” *J. Speech Hear. Res.*, 11: 796-804.

Fisher, J.W., III and Darrell, T. (2002). "Informative subspaces for audio-visual processing: high-level function from low-level fusion," *Proc. ICASSP*, 4104-4107

Flanagan, J.L. (1965). *Speech analysis synthesis and perception*. New York: Springer-Verlag.

Franks, J.R. (1972). “The confusion of English consonant clusters in lipreading,” *J. Speech Hear. Res.*, 15: 474-482.

Girin, L., Schwartz, J.L., and Feng, G. (2001). “Audio-visual enhancement of speech in noise,” *J. Acoust. Soc. Am.*, 109(6): 3007-3020.

Goldstone, R.L. (1999). “Similarity,” in R.A. Wilson and F. Keil (Eds.), *The MIT Encyclopedia of the Cognitive Sciences* (MITECS) [On-line]. Available: http://cognet.mit.edu/MITECS/Entry/goldstone.html.

Grant, K.W. and Seitz, P.F. (2000). “The use of visible speech cues for improving auditory detection of spoken sentences,” *J. Acoust. Soc. Am.*, 108(3): 1197-1208.

Green, K.P. and Kuhl, P.K. (1991). “Integral processing of visual place and auditory voicing information during phonetic perception,” *J. Experimental Psychology: Human Perception and Performance*, 17: 278-288.

Hays, W.L. (1994). *Statistics*. 5th edition, Fort Worth, TX: Harcourt Brace College Publishers.

Henton, C. and Litwinowics, P. (1994). “Saying and seeing it with feeling: techniques for synthesizing visual, emotional speech,” *Proc. 2nd ESCA-IEEE Workshop on Speech Synthesis*, New York, NY, 73-76.

Hershey, J. and Movellan, J. (1999). "Using audio-visual synchrony to locate sounds," *Proc. NIPS*, 813-819.

Horn, R.A. and Johnson, C.R. (1985). *Matrix analysis*. Cambridge University Press.

Iverson, P., Bernstein, L.E., and Auer, E.T. (1998). “Modeling the interaction of phonemic intelligibility and lexical structure in the audiovisual word recognition,” *Speech Commun.*, 26: 45-63.

Jakobson, R., Fant, G., and Halle, M. (1952). *Preliminaries to Speech Analysis*. Cambridge, MA: MIT Press.

Kailath, T., Sayed, A.H., and Hassibi, B. (2000). *Linear Estimation*. Prentice Hall.

Keating, P.A., Cho, T., Baroni, M., Mattys, S., Bernstein, L.E., Chaney, B., and Alwan, A. (2000). "Articulation of word and sentence stress," *J. Acoust. Soc. Am.*, 108: 2466.

King, S.A. (2001). *A Facial Model and Animation Techniques for Animated Speech*. Ph.D. dissertation, Ohio State Univ., Dept. of Computer and Information Science.

Kricos, P. and Lesner, S. (1982). "Differences in visual intelligibility across talkers," *The Volta Review*, 84: 219-225.

Kruskal, J.B. and Wish, M. (1978). *Multidimensional scaling*. Beverly Hills, CA and London: Sage Publications.

Ladefoged, P. (2001). *A course in phonetics*. 4th edition, Harcourt College Publishers.

Ladefoged, P., Harshman, R., Goldstein, L., and Rice, L. (1978). "Generating vocal tract shapes from formant frequencies," *J. Acoust. Soc. Am.*, 64: 1027-1035.

Langereis, M.C., Bosman, A.J., Olphen, A.F., and Smoorenburg, G.F. (2000). "Relation between speech perception and speech production in adult cochlear implant users," *The Nature of Speech Perception Workshop*, Utrecht, Netherlands.

Lavagetto, D., Arzarello, M., and Caranzano, M. (1994). "Lipreadable frame animation driven by speech parameters," *Proc. IEEE Int. Symposium on Speech, Image Processing and Neural Networks*, Hong Kong, China, 14-16.

Lee, L.J., Fieguth, P., and Deng, L. (2001). "A functional articulatory dynamic model for speech production," *Proc. ICASSP*, Salt Lake City, UT.

Lee, Y., Terzopoulos, D., and Waters, K. (1993). "Constructing physics-based facial models of individuals," *Proc. Graphics Interface*, Canadian Human-Computer Communications Society, Toronto, Canada, 1-8.

Lee, Y., Terzopoulos, D., and Waters, K. (1995). "Realistic modeling of facial animation," *Proc. ACM SIGGRAPH*, Los Angeles, CA, 55-62.

Le Goff, B., Guiard-Marigny, T., Cohen, M., and Benoît, C. (1994). "Real-time analysis-synthesis and intelligibility of talking faces," *Proc. 2nd ESCA/IEEE Workshop on Speech Synthesis*, New Paltz, NY, 53-56.

Levinson, S.E. and Schmidt, C.E. (1983). "Adaptive computation of articulatory parameters from the speech signal," *J. Acoust. Soc. Am.*, 74(4): 1145-1154.

Liberman, A.M. (1957). “Some results of research on speech perception,” *J. Acoust. Soc. Am.*, 29: 117-123.

Liberman, A.M. and Mattingly, I.G. (1985). “The motor theory of speech perception revised,” *Cognition*, 21(1): 1-36.

Lieberman, P. and Blumstein, S. (1988). *Speech physiology, speech perception, and acoustic phonetics*. Cambridge: Cambridge University Press.

Lindsay, D. (1997). “Talking Head,” *Invention and Technology*, 57-63.

Lisker, L. (1975). “Is it VOT or a first-formant transition detector?” *J. Acoust. Soc. Am.*, 57: 1547-1551.

Lisker, L. and Abramson, A.S. (1964). “A cross-language study of voicing in initial stops: acoustical measurements,” *Word*, 20: 384-422.

Lisker, L. and Abramson, A.S. (1970). “The voicing dimension: some experiments in comparative phonetics,” *Proc. ICPhS*, Prague: Academia.

Luettin, J. and Dupont, S. (1998). “Continuous audio-visual speech recognition,” *Proc. 5th European Conference on Computer Vision*, Freiburg, Germany, 657–673.

Luettin, J., Thacker, N.A., and Beet, S.W. (1996). “Active shape models for visual speech feature extraction,” in D.G. Stork and M.E. Hennecke (Eds.), *Speechreading by Humans and Machines* (pp. 383-390), Springer Verlag.

MacLeod, A. and Summerfield, Q. (1987). “Quantifying the contribution of vision to speech perception in noise,” *British J. Audiology*, 21: 131-141.

Magnenat-Thalmann, N., Primeau, N.E., and Thalmann, D. (1988). “Abstract muscle actions procedures for human face animation,” *The Visual Computer*, 3: 290-297.

Massaro, D.W. (1998). *Perceiving talking faces, from speech perception to behavioral principle*. Cambridge, MA: MIT.

Massaro, D.W., Beskow, J., Cohen, M.M., Fry, C.L., and Rodriguez T. (1999). “Picture my voice: audio to visual speech synthesis using artificial neural networks,” *Proc. AVSP*, Santa Cruz, CA, 133-138.

McGurk, H. and MacDonald, J. (1976). “Hearing lips and seeing voices,” *Nature*, 264: 746-748.

Miller, J.L. (1999). “Speech perception,” in R.A. Wilson and F. Keil (Eds.), *The MIT Encyclopedia of the Cognitive Sciences* (MITECS) [On-line]. Available: http://cognet.mit.edu/MITECS/Entry/miller.html.

Miller, R.G. (1974). “The jackknife-a review,” *Biometrika*, 61: 1-15.

Montgomery, A.A. and Jackson, P.L. (1983). “Physical characteristics of the lips underlying vowel lipreading performance,” *J. Acoust. Soc. Am.*, 76: 2134-2144.

Mori, K. and Sonoda, Y. (1998). “Relationship between lip shapes and acoustical characterics during speech,” *Proc. ICSLP*, Sydney, Australia.

Munhall, K.G. and Vatikiotis-Bateson, E. (1998). “The moving face during speech communication,” in R. Campbell, B. Dodd, and D. Burnham (Eds.), *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-visual Speech* (pp. 123-139), East Sussex, UK: Psychology Press.

Narayanan, S. and Alwan, A. (2000). “Articulatory-acoustic models for fricative consonants,” *IEEE Trans. Speech and Audio Proc.*, 8(3): 328-344.

Narayanan, S., Alwan, A., and Haker, K. (1997). “Towards articulatory-acoustic models of liquid consonants. Part I: the laterals,” *J. Acoust. Soc. Am.*, 101(2): 1064-1077.

Nearey, T.M. (1990). “The segment as a unit of speech perception,” *J. Phonetics*, 18: 347–373.

Nearey, T.M. (1992). “Context effects in a double-weak theory of speech perception,” *Language and Speech*, 35: 153–172.

Nearey, T.M. (1997). “Speech perception as pattern recognition,” *J. Acoust. Soc. Am.*, 101: 3241-3254.

Nock, H.J., Iyengar, G., and Neti, C. (2002). “Assessing face and speech consistency for monologue detection in video,” *Proc. ACM Multimedia*, Juan-les-Pins, France.

Nock, H.J., Iyengar, G., and Neti, C. (2003). “Speaker localisation using audio-visual synchrony: an empirical study,” *Proc. CIVR*, Urbana, IL.

Ostberg, O., Lindstrom, B., and Renhall, P.O. (1988). “Contribution to speech intelligibility by different sizes of videophone display,” *Proc. Workshop on Videophone Terminal Design*, Torino, Italy.

Ouni, S. and Laprie, Y. (2000). “Improving acoustic-to-articulatory inversion by using hypercube codebooks,” *Proc. ICSLP*, Beijing, China.

Owens, E. and Blazek, B. (1985). “Visemes observed by hearing-impaired and normal hearing adult viewers,” *J. Speech Hear. Res.*, 28: 381-393.

Parke, F.I. (1974). *A parametric model for human faces*. Ph.D. dissertation, Dept. of Computer Sciences, Univ. of Utah.

Parke, F.I. (1982). “Parameterized models for facial animation,” *IEEE Computer Graphics*, 2: 61-68.

Parke, F.I. and Waters, K. (1996). *Computer facial animation*. AK Peters.

Pearce, A., Wyvill, B., Wyvill, G., and Hill, D. (1986). “Speech and expression: a computer solution to face animation,” *Proc. Graphics Interface*, Calgary, Canada, 136-140.

Perkell, J.S., Cohen, M., Svirsky, M., Matthies, M., Garabieta, I., and Jackson, M. (1992). “Electromagnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements,” *J. Acoust. Soc. Am.*, 92: 3078–3096.

Peterson, G., Wang, W., and Sivertsen, E. (1958). “Segmentation techniques in speech synthesis,” *J. Acoust. Soc. Am.*, 30: 739-742.

Picheny, M.A., Durlach, N.I., and Braida, L.D. (1985). “Speaking clearly for the hard of hearing I: intelligibility differences between clear and conversational speech,” *J. Speech Hear. Res.*, 28: 96-103.

Picheny, M.A., Durlach, N.I., and Braida, L.D. (1986). “Speaking clearly for the hard of hearing II: acoustic characteristics of clear and conversational speech,” *J. Speech Hear. Res.*, 29: 434-446.

Picheny, M.A., Durlach, N.I., and Braida, L.D. (1989). “Speaking clearly for the hard of hearing III: an attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech,” *J. Speech Hear. Res.*, 32: 600-603.

Platt, S.M. and Badler, N.I. (1981). “Animating facial expressions,” *Computer Graphics*, 15: 245-252.

Potamianos, G., Luettin, J., and Neti, C. (2001). “Hierarchical discriminant features for audio-visual LVCSR,” *Proc. ICASSP*, Salt Lake City, UT.

Potamianos, G. and Neti, C. (2003). “Audio-visual speech recognition in challenging environments,” *Proc. EUROSPEECH*, Geneva, Switzerland.

Rabiner, L.R. and Schafer, R.W. (1978). *Digital processing of speech signals*. Prentice Hall.

Rahim, M.G. and Goodyear, C.C. (1990). “Estimation of vocal tract filter parameter using a neural net,” *Speech Commun.*, 9: 49-55.

Rao, R., Mersereau, R., and Chen, T. (1997). "Using HMM's in audio-to-visual conversion," *IEEE Workshop on Multimedia Signal Processing*, Princeton, NJ, 19-24.

Rubin, P. and Vatikiotis-Bateson, E. (1998). *Talking heads* [On-line]. Available: http://www.haskins.yale.edu/haskins/heads.html.

Rydfalk, M. (1987). "CANDIDE: A parameterized Face," *Technical Report LiTH-ISY-I-0866*, Dept. of Electrical Engineering, Linkoping Univ., Sweden.

Sachs, L. (1984). *Applied Statistics: A Handbook of Techniques*. Springer-Verlag.

Saintourens, M., Tramus, M.H., Huitric, H., and Nahas, M. (1990). "Creation of a synthetic face speaking in real time with a synthetic voice," *Proc. 1st ESCA Workshop on Speech Synthesis*, Autrans, France, 249-252.

Schönle, P.W., Gräbe, K., Wenig, P., Höhne, J., Schrader, J., and Conrad, B. (1987). "Electromagnetic articulography: use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," *Brain and Language*, 31: 26–35.

Schroeter, J. and Sondhi, M.M. (1992). "Speech coding based on physiological models of speech production," in S. Furui and M. Sondhi (Eds.), *Advances in Speech Signal Processing* (pp. 231-267), New York: Dekker.

Schroeter, J. and Sondhi, M.M. (1994). "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE Trans. Speech and Audio Proc.*, 2(1): 133-150.

Sen, A. and Srivastava, M. (1990). *Regression analysis: theory, methods, and applications*. Springer-Verlag.

Shepard, R.N. (1962). "The analysis of proximities: multidimensional scaling with an unknown distance function. Part 1," *Psychometrika*, 27: 125-140.

Shepard, R.N. (1987). "Toward a universal law of generalization for psychological science," *Science*, 237: 1317-1323.

Shirai, K. and Kobayashi, T. (1991). "Estimation of articulatory motion using neural networks," *J. Phonetics*, 19: 379-385.

Slaney, M. and Covell, M. (2001). "FaceSync: a linear operator for measuring synchronization of video facial images and audio tracks," *Proc. NIPS*.

Smith, S. (1989). "Computer lip reading to augment automatic speech recognition," *Speech Tech.*, 175-181.

Soli, S.D. and Arabie, P. (1979). “Auditory versus phonetic accounts of observed confusions between consonant phonemes,” *J. Acoust. Soc. Am.*, 66: 46-59.

Sondhi, M.M. and Schroeter, J. (1987). “A hybrid time-frequency domain articulatory speech synthesizer,” *IEEE Trans. Acoust. Speech Sig. Proc.*, 35(7): 955-967.

SPSS Base 9.0 User’s Guide.

Stevens, K.N (1972). “The quantal nature of speech: evidence from articulatory-acoustic data,” in E.E. David and P.B. Denes (Eds.), *Human communication: a unified view* (pp. 51-66), New York: McGraw-Hill.

Stevens, K.N. (1989). “On the quantal nature of speech,” *J. Phonetics*, 17: 3–45.

Stevens, K.N. and Blumstein, S. (1981). “The search for invariant acoustic correlates of phonetic features,” in P.D. Eimas and J.L. Miller (Eds.), *Perspectives on the Study of Speech* (pp. 1-38), Hillsdale, NJ: Erlbaum.

Stevens, K.N. and House, A.S. (1955). “Development of a quantitative description of vowel articulation,” *J. Acoust. Soc. Am.*, 27: 484-493.

Stevens, K.N. and Keyser, S.J. (1989). “Primary features and their enhancement in consonants,” *Language*, 65: 81–106.

Sugamura, N. and Itakura, F. (1986). “Speech analysis and synthesis methods developed at ECL in NTT - from LPC to LSP,” *Speech Commun.*, 5: 199-215.

Sumby, W.H. and Pollack, I. (1954). “Visual contribution to speech intelligibility in noise,” *J. Acoust. Soc. Am.*, 26: 212-215.

Terzopoulos, D. and Waters, K. (1990). “Physically-based facial modeling, analysis, and animation,” *J. Visualization and Computer Animation*, 1(4): 73-80.

Torgerson, W.S. (1965). “Multidimensional scaling of similarity,” *Psychometrika*, 30: 379-393.

Tversky, A. (1977). “Features of similarity,” *Psychological Review*, 84: 327-352.

Vatikiotis-Bateson, E. and Ostry, D.J. (1995). “An analysis of the dimensionality of jaw motion in speech,” *J. Phonetics*, 23: 101-117.

Vignoli, F. (2000). “From speech to talking faces: lip movements estimation based on linear approximators, ” *Proc. ICASSP*, Istanbul, Turkey, 2381-2384.

Walden, B.E. and Montgomery, A.A. (1975). “Dimensions of consonant perception in normal and hearing-impaired listeners,” *J. Speech Hear. Res.*, 18: 444-455.

Walden, B.E., Montgomery, A.A., and Prosek, R.A. (1987). “Perception of synthetic visual consonant-vowel articulations,” *J. Speech Hear. Res.*, 30: 418-424.

Walden, B.E., Prosek, R.A., Montgomery, A.A., Scherr, C.K., and Jones, C.J. (1977). “Effects of training on the visual recognition of consonants,” *J. Speech Hear. Res.*, 20: 130-145.

Waters, K. (1987). “A muscle model for animating three-dimensional facial expression,” *Computer Graphics*, 21: 17-24.

Waters, K. (1992). “A physical model of facial tissue and muscle articulation derived from computer tomography data,” *Proc. SPIE Visualization in Biomedical Computing*, N. Carolina, 574-583.

Waters, K. and Levergood, T.M. (1993). “DECface: an automatic lip synchronization algorithm for synthetic faces,” *Technical Report CRL 93/4*, DEC Cambridge Research Laboratory.

Waters, K. and Terzopoulos, D. (1991). “Modeling and animating faces using scanned data,” *J. Visualization and Animation*, 2(4): 123-128.

Williams, J.J., Katsaggelos, A.K., and Randolph, M.A. (2000). “A hidden markov model based visual speech synthesizer,” *Proc. ICASSP*, Istanbul, Turkey, 2393-2396.

Yamamoto, E., Nakamura, S., and Shikano, K. (1998). “Lip movement synthesis from speech based on Hidden Markov Models,” *Speech Commun.*, 26(1-2): 105-115.

Yehia, H.C., Kuratate, T., and Vatikiotis-Bateson, E. (1999). “Using speech acoustics to drive facial motion,” *Proc. ICPhS*, San Francisco, CA, 631-634.

Yehia, H.C., Rubin, P., and Vatikiotis-Bateson, E. (1998). “Quantitative association of vocal-tract and facial behavior,” *Speech Commun.*, 26(1): 23-43.

Young, F.W. and Hamer, R.M. (1987). *Multidimensional scaling: history, theory and applications*. Hillsdale, NJ: Erlbaum.

Zhu, Q. and Alwan, A. (2000). “On the use of variable frame rate analysis in speech recognition,” *Proc. ICASSP*, Istanbul, Turkey, 1783-1786.