University of California
Los Angeles

Dependencies of Voice Source Measures on Age, Sex, Vowel Context, and Prosodic Features

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Electrical Engineering

by

Markus Iseli

2007

The dissertation of Markus Iseli is approved.

Yuanxun Wang

Mani. B. Srivastava

Jody Kreiman

Abeer Alwan, Committee Chair

University of California, Los Angeles

2007

Table of Contents

List of Figures

List of Tables

Acknowledgments

I am deeply grateful to my advisor, Abeer Alwan. Her professional guidance and her great and honest personality helped me find my way at UCLA. Abeer, your lab and its students were a loving and creative family for me. I would also like to express my gratitude to professors Mani Srivastava and Yuanxun Wang for agreeing to be in my Ph. D. committee and for their interest in my research. It was a treat to work with professors Patricia Keating and Jody Kreiman, who introduced me to linguistics and statistics and who were always available to answer my questions and discuss my research. I would like to thank Admela Jukan for inspiring me to go on this journey in research, Sankaran "Panchi" Panchapagesan for helping me polish up my rusty knowledge of signal processing theory and for his friendship, and last but not least, Yen-Liang Shue for his great teamwork and for being a good friend. For all the persons not mentioned here by name and who helped me on my quest to find answers in research as well as in life, please accept my gratitude.

This material is based upon work supported by NSF Grant No. 0326214 and by a Radcliffe Fellowship to Abeer Alwan.

Vita

1968	Born, Zurich, Switzerland.
1992	Diploma in Electrical Engineering Swiss Federal Institute of Technology (ETH), Zürich, Switzerland
1992–1994	Research in the field of room acoustics Akustische und Kino Geräte (AKG), Vienna, Austria
1995–2001	Research and development in the fields of telecommunication, speech recognition, and artificial intelligence. Consulting for speech processing applications. Siemens AG Austria, Vienna, Austria
6-9/2001	Research and development in the field of speaker adaptation Hughes Research Labs, Malibu, California
2003	M.S. in Electrical Engineering University of California, Los Angeles, (UCLA)
2002–2006	Research/Teaching Assistant Electrical Engineering Department, University of California, Los Angeles (UCLA)
2001–2007	Lecturer Electrical Engineering Department, University of California, Los Angeles (UCLA)

Publications and Presentations

M. Iseli, Y.-L. Shue, A. Alwan,"Age, sex, and vowel dependencies of acoustical measures related to the voice source," *The Journal of the Acoustic Society of America*, vol. 121, num. 4, pp. 2283–2295, 2007.

M. Iseli, Y.-L. Shue, M. Epstein, P. Keating, A. Alwan,"Voice source correlates of prosodic features in American English: a pilot study," *Proceedings of ICSLP*, Pittsburgh, PA, pp. 2226–2229, 2006.

M. Iseli, Y.-L. Shue, and A. Alwan,"Age- and Gender-Dependent Analysis of Voice Source Characteristics," *Proceedings of IEEE ICASSP*, Toulouse, France, vol. 1, pp. 389–392, May 2006.

A. Kazemzadeh, H. You, M. Iseli, B. Jones, X. Cui, M. Heritage, P. Price, E. Anderson, S. Narayanan, and A. Alwan, "TBALL Data Collection: the Making of a Young Children's Speech Corpus," *Proceedings of IEEE Eurospeech/Interspeech*, Lisboa, Portugal, pp. 1581–1584, Sep. 2005.

M. Iseli and A. Alwan, "An Improved Correction Formula for The Estimation of Harmonic Magnitudes and Its Application to Open Quotient Estimation," *Proceedings of IEEE ICASSP*, Montreal, Canada, pp. 669–672, May 2004.

X. Cui, M. Iseli, Q. Zhu, and A. Alwan, "Evaluation of noise robust features on the aurora databases," *Proceedings of ICSLP*, Denver, Colorado, vol.1, pp.481-484, Sep. 2002.

Q. Zhu, X. Cui, M. Iseli, and A. Alwan, "Noise Robust Feature Extraction for ASR using the Aurora 2 Database," *Proceedings of IEEE Eurospeech*, Aalborg, Denmark, vol. 1, pp. 185–188, Sept 2001.

M. Iseli and A. Alwan, "Inter- and Intra-speaker Variability of Glottal Flow Derivative using the LF Model," *Proceedings of ICSLP*, Bejing, China, vol. 1, pp. 477–480, Oct. 2000.

M. Iseli and Y.-L. Shue and A. Alwan, "Analysis of vowel and speaker dependencies of source harmonic magnitudes in consonant-vowel utterances," *The Journal of the Acoustic Society of America*, vol. 117, num. 4, p. 2619, 2005.

M. Iseli and A. Alwan, "An improved correction formula for the estimation of voice source harmonic magnitudes," *The Journal of the Acoustic Society of America*, vol. 115, num. 5, p. 2610, 2004.

S. Kadambe and M. Iseli, "Fast on-line speaker/environment adaptation using modified maximum likelihood stochastic matching," *The Journal of the Acoustic Society of America*, vol. 112, num. 5, p. 2321, 2002.

Abstract of the Dissertation

Dependencies of Voice Source Measures on Age, Sex, Vowel Context, and Prosodic Features

by

Markus Iseli

Doctor of Philosophy in Electrical Engineering

University of California, Los Angeles, 2007

Professor Abeer Alwan, Chair

The effects of age, sex, vocal tract configuration, and prosody on the glottal excitation signal in speech are only partially understood, yet understanding these effects is critical to better human speech production models and for improving speech processing applications.

This dissertation evaluates the dependencies of voice source measures on age, sex, vowel context, and prosody. A formula to extract spectral voice source measures by compensating for the influence of formant frequencies is derived and then used to analyze 3145 utterances spoken by 335 native talkers of American English ranging in age between 8 and 39 years old. The measures analyzed are: the fundamental frequency (F_0), the difference between the first two source spectral harmonic magnitudes, $H_1^* - H_2^*$ (related to the open quotient), and the difference between the magnitudes of the first source spectral harmonic and that of the third formant peak, $H_1^* - A_3^*$ (related to source spectral tilt). Asterisks indicate that the measures are corrected for the influence of the vocal tract transfer function. Experimental results show that these three measures are dependent to varying degrees on age, sex, and vowel. The statistical significance of these

results is shown and there seem to exist interdependencies for certain voice source measures.

A pilot study is then conducted to assess the dependencies of five voice source measures F_0, E_e (maximal glottal flow change, related to voice source intensity), R_k (symmetry/skew of the glottal airflow), $H_1^* - H_2^*$, and LIN (spectral linearity, related to source spectral tilt) on three prosodic events: lexical stress, pitch accent, and boundary tone. The study analyzes the speech of one male and two female talkers of American English using a sentence pronounced with different prosodic events. Results show that lexical stress and an increase in tone (F_0) both yield an increase in loudness/intensity and in high-frequency components of the voice source signal, which could be attributed to a tenser voice. The voice source measure $H_1^* - H_2^*$, however, was affected only by stress and not by pitch accent.

A better knowledge of the dependencies of voice source measures on age, sex, vowel context, and prosody will help in the estimation of a talker's age and sex, and the detection of prosodic events and emotion in human speech.

CHAPTER 1

Introduction

1.1 Overview and Motivation

For almost half a century, research has been conducted on the nature of the glottal voice source signal, and glottal source measures have been estimated using various procedures and algorithms. In the past, the voice source signal was mainly studied in the context of text-to-speech synthesis and speech coding systems in the field of engineering. Recent studies [SV96, SVP97, FKL00] have shown that a relationship exists between the acoustic measures of the glottal voice source signal and perceptual voice quality.

The main challenge in estimating voice source measures is the accurate estimation of vocal tract resonances and their bandwidths which is yet an unsolved problem. Because of the complexity of the inverse-filtering task, a few studies have successfully applied voice source measures in practical applications such as speaker identification [PQR99], speech synthesis [GC02], speech analysis and synthesis [CM95], speech coding [SPC97], and speech enhancement [YFL99]. Most applications, however, do not use information about the voice source.

A better knowledge of the relationship between acoustic measures that characterize the voice source and speaker properties such as sex and age, context or

sound type, and prosodic features such as tones and boundaries, would improve our understanding of the human speech production mechanism. This knowledge would also help improve voice analysis for a variety of speech processing applications.

The goal of this research is to develop an approach for extracting voice source measures reliably, and to do subsequent analysis of those measures to unravel dependencies on age, sex, vowel context, and prosodic features.

1.2 Speech production

The human voice production mechanism can be divided into three parts: *lungs*, *vocal folds*, and *vocal tract*. A schematic diagram of the human vocal mechanism is shown in Figure 1.1. Air pressure from the lungs causes air to flow through the *glottis*, which is the airspace between the vocal folds. The vocal folds are two masses of flesh, ligament, and muscle, which stretch between the front and back of the larynx (colloquially known as the "voicebox"). Depending on the adduction or abduction of the vocal folds, they are in different vibratory modes (*voiced* sounds) or are not vibrating at all (*unvoiced* sounds). For voiced sounds, the vocal folds open and close quasi-periodically and thus convert the glottal air flow (air volume velocity) into flow pulses, called the voice source signal. The voice source signal then passes through the vocal tract, which begins at the glottis and ends at the lips. The vocal tract acts as a body with resonances (called *formant frequencies*) and anti resonances (or zeros). It functions as an acoustic filter that shapes the spectrum of the sound. The various speech sounds are produced by adjusting both the shape of the vocal tract as well as the voice source signal. In this dissertation, only voiced sounds are analyzed.

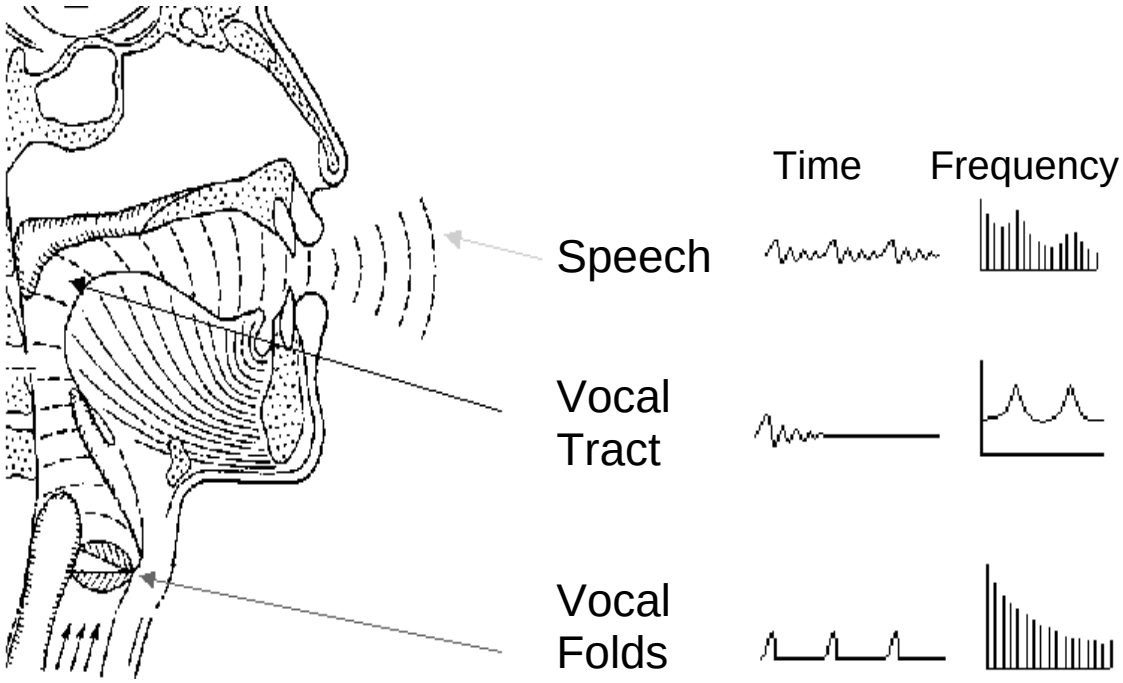

Figure 1.1: Speech production of voiced speech. Air pressure from the lungs produces vibration of the vocal folds, which results in a quasi-periodic pulse-shaped voice source signal. The voice source signal excites the vocal tract, which acts as a resonance body enhancing and attenuating certain frequencies, and voiced speech is produced (from [Ber02]).

1.2.1 The linear source-filter model of speech production

In the linear source-filter model of speech production, described in [Fan60, Fla72], the glottal excitation acts as the source and the vocal tract acts as the linear filter. The speech pressure waveform measured by a microphone in front of the lips can be approximated by the time derivative of the volume velocity signal [RS78]. Because of the assumed linearity of the source-filter model, this lip radiation effect can be included in the source signal of the linear source filter model, i.e. the source function is represented by the derivative of the glottal flow volume

velocity. Figure 1.2 shows the integration of lip radiation into the linear source filter model. From filter theory we know that the output of a linear filter is the convolution of the filter impulse response with the input signal, hence:

$$s(t) = u(t) \star h(t), \tag{1.1}$$

where $\star$ is the symbol for convolution, u(t) is the glottal flow derivative, h(t) is the vocal tract impulse response, and s(t) is the resulting speech signal.

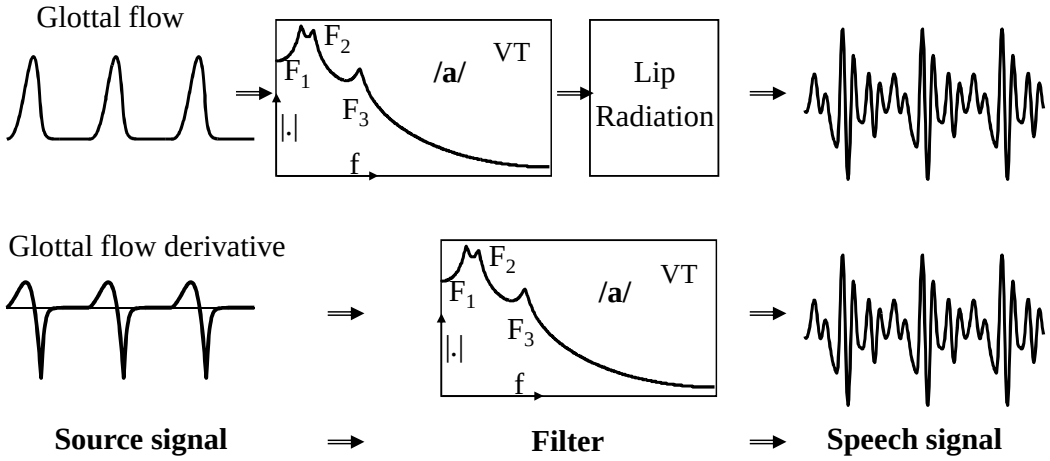

Figure 1.2: The linear source-filter model of speech production. Lip radiation acts as a derivative of glottal airflow and can be integrated into the source: Without (top) and with (bottom) integration of the lip radiation effect into the source. The box representing the vocal tract (VT) filter shows the vocal tract frequency response with the resonance frequencies (formants) F_1, F_2, and F_3.

Consequently, the convolution in the time domain results in a multiplication in the frequency domain, hence:

$$S(f) = U(f) \cdot H(f). \tag{1.2}$$

The linear source-filter model of speech production makes the simplifying

assumptions that the source and vocal tract can be modeled as linear filters and that they are independent and linearly separable. In reality there exists an interaction between the voice source signal and the vocal tract, especially for low-frequency formants. There exist several models of speech production which take this interaction into account. Most of these models are based on the non-linear two-mass mechanical model by [FI78]. More information on source-tract interaction can be found in [FL85] and [Chi94]. Because of its simplicity, the linear source filter model, which is used in this dissertation, is the commonly used model of speech production.

1.2.2 The voice source signal

The voice source signal is usually represented by the glottal airflow volume velocity or its derivative. In voiced speech the vibration of the vocal folds modulates the airflow from the lungs. One vibratory cycle is generated as follows: Air pressure from the lungs forces the vocal folds to open - their tension increases. As the airflow velocity through the glottis increases, the pressure at the glottis decreases (Bernoulli's Principle). The combination of pressure decrease and vocal fold tension increase causes the vocal folds to shut close abruptly, and then the cycle starts again. The duration of one cycle is called the fundamental period (T_0) and its frequency domain equivalent is called the *fundamental frequency* ($F_0 = 1/T_0$). F_0 is sometimes also referred to as "pitch frequency" or "pitch". In singing, the tone height is defined by F_0 and in normal speech, changes in F_0 provide additional information to the spoken text; for example, varying F_0 can indicate the end of a sentence, a speaking turn, or if the sentence is a question or a statement (see Section 1.5). The human voice source production changes for boys between the ages of about 9 and 12 years due to physiological growth

of the larynx ("Adam's apple") and the vocal folds: Adult males have thicker vocal folds with a length of about 15 mm, about 2 mm longer than adult females, which results in lower F_0 values for males.

Early models of the voiced source signal used a simple impulse train. More recent studies model the shape of the glottal airflow or its derivative in the time domain [Hol73, Ana84, Hed84, FLL85, KK90, Ros71]. Several parametric time domain voice source models have been proposed in the literature to represent glottal flow, or its derivative. Among them are the Rosenberg [Ros71], the Liljencrants-Fant [FLL85], the KLGLOTT88 [KK90] and [LS99], and the R++ model [Vel98]. Each model has its own set of four to five parameters. Frequency domain representations of those models, some of them are presented in [Fan95] and [Dd99], can be helpful in the parameter optimization/estimation process. A promising approach to studying the voice source excitation as a mixed causal/anticausal low-pass filter with a spectral glottal peak is presented in [DdH03]. The authors claim that their approach facilitates the estimation of voice quality. A historical survey of glottal models for digital speech processing can be found in [CC95]. In this dissertation, the Liljencrants-Fant (LF) model is used.

The LF model approximates the glottal flow derivative and is shown in Figure 1.3. The model distinguishes between open phase (vocal folds are open), return phase (folds are closing), and closed phase (folds are closed, no airflow through the glottis). The basic equations for the open phase ($E_1(t)$) and the return phase ($E_2(t)$) in continuous time are:

$$E(t) = \begin{cases} E_1(t) &= E_0 e^{\alpha t} \sin(\omega_g t) & (0 \le t \le T_e) \\ E_2(t) &= (\frac{-E_e}{\epsilon T_a})[e^{-\epsilon(t-T_e)} - e^{-\epsilon(T_c-T_e)}] & (T_e < t \le T_c). \end{cases} \tag{1.3}$$

Parameters are the growth factor α, the amplitude scaling factor E_0, the expo-

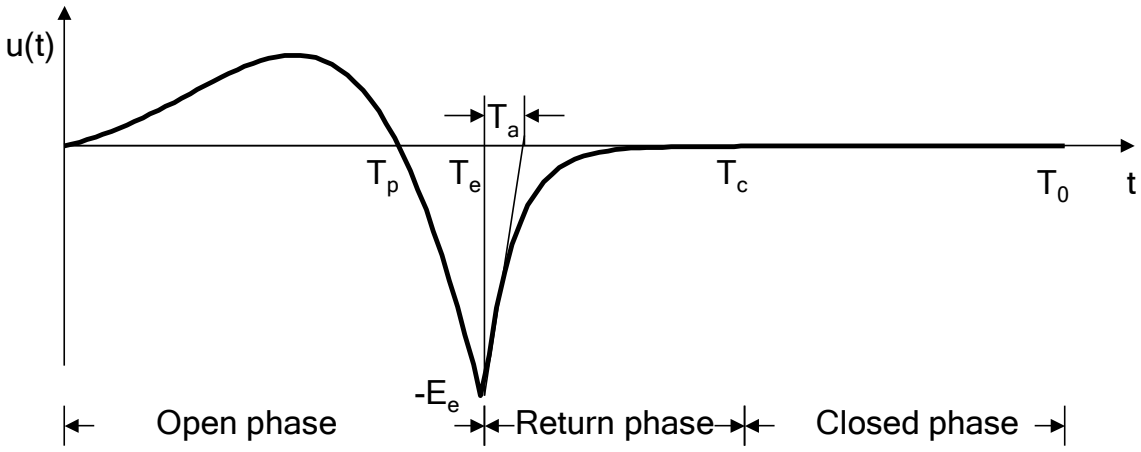

Figure 1.3: The LF model for the glottal flow derivative and its parameters: instant of maximum airflow (T_p), instant of maximum airflow derivative (T_e), effective duration of return phase (T_a), beginning of closed phase (T_c), fundamental period T_o, and amplitude of maximum excitation of glottal flow derivative (E_e).

nential time constant of the return phase ϵ, the duration of the return phase T_a, the instant of glottal closure T_c, the instant of maximal glottal flow derivative T_e, and E_e, which is the magnitude of the signal at time T_e. Defining the instant of maximal glottal airflow as T_p, ω_g is then defined as $\omega_g = \frac{\pi}{T_p}$.

Figure 1.4 depicts an idealized power magnitude spectrum of the voice source signal representing the derivative of the glottal flow volume velocity. Since voiced sounds have a nearly periodic time domain source, their spectrum displays a nearly harmonic frequency domain structure with harmonics at multiples of F_0. The spectral tilt of the glottal flow derivative is typically about -6 dB per octave. The figure also shows the first two source spectral harmonic magnitudes H_1^* and H_2^* (in dB) and the asterisks indicate that the measures are calculated from the source spectrum (without vocal tract information).

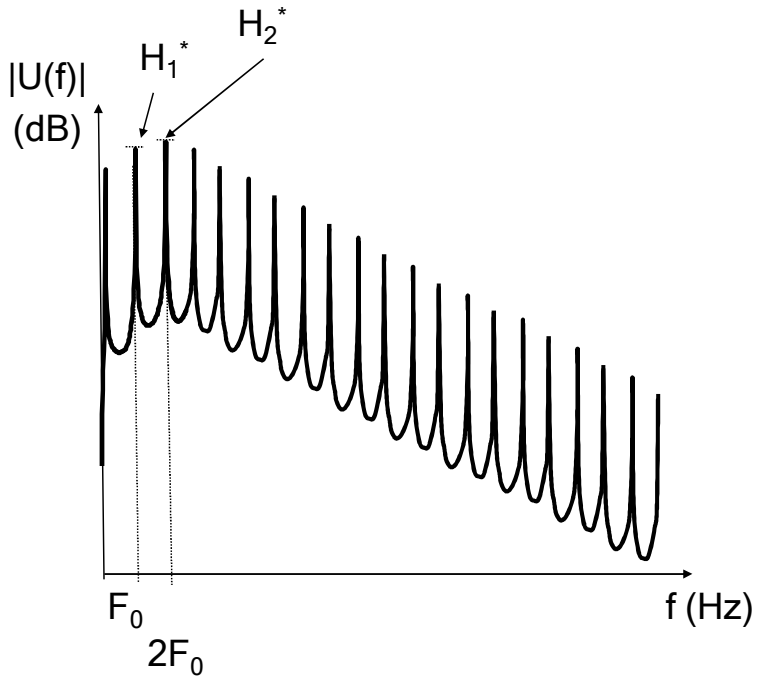

Figure 1.4: Power magnitude spectrum of an idealized voice source signal representing the derivative of the glottal flow volume velocity. H_1^* and H_2^* are the first two source spectral harmonic magnitudes (in dB) at frequencies F_0 and $2F_0$. The asterisks indicate that the magnitudes are measured from the source spectrum and therefore have been corrected for the effect of vocal tract resonances.

1.2.3 The vocal tract

In speech production, the main purpose of the vocal tract is to spectrally "color" the voice source, which is important for making perceptually different speech sounds: For example the vowel /ae/ as in "bat" has different formants than the vowel /uw/ as in "boot". Another purpose of the vocal tract is to produce sound sources for unvoiced sounds either by constricting the air flow and creating turbulence, as in the fricative /f/, or by stopping and suddenly releasing the air flow, as in the plosive /p/.

The shape and cross-sectional profile of the vocal tract is adjusted by articulatory motion, which includes manipulating the tongue, lips, velum, mouth and lower jaw. The vocal tract shape determines the sound energy transfer function from the glottis to the lips and can be described in terms of resonances (*formants*) and anti resonances. Each formant is described by its resonance frequency (*formant frequency*) and its resonance bandwidth (*formant bandwidth*). Certain sounds, especially nasals like the "m" in "meet", produce anti-resonances, called *zeros*, where energy is trapped in the vocal tract. The lowest formant frequency is called the first formant (F_1), the second lowest formant frequency is called the second formant (F_2), and so on. For example in [PB52] it was found that the vocal tract resonance frequencies averaged over all their male talkers of American English for the vowel /a/ as in "Bob" were at 730 Hz, 1090 Hz, and 2440 Hz. Given this information we can set $F_1 = 730$ Hz, $F_2 = 1090$ Hz, and $F_3 = 2440$ Hz, with $F_1 < F_2 < F_3 < ....$ Each formant has a corresponding bandwidth, B_1 (first formant bandwidth), B_2 (second formant bandwidth), etc., in Hz, which represents the resonance damping factor (e.g. vocal tract wall losses, thermal losses, etc.); a larger damping factor results in a wider (larger) bandwidth. Figure 1.5 shows the simplified power magnitude spectrum of an idealized vocal tract with the three formants F_1, F_2, and F_3. Recall that fundamental frequency F_0 is related to the source signal while the formant frequencies F_1, F_2, etc. are related to the vocal tract.

In the discrete frequency domain, the vocal tract is typically modeled as a linear filter with a transfer function $H(z) = B(z)/A(z)$, where poles of $H(z)$ (A(z)=0) model resonances of the vocal tract (formants), and zeros of $H(z)$ ($B(z) = 0$) model anti-resonances of the vocal tract (zeros). This dissertation mainly examines vowels which can be characterized by prominent resonances in their spectra. Hence, to represent the vocal tract transfer function we used the

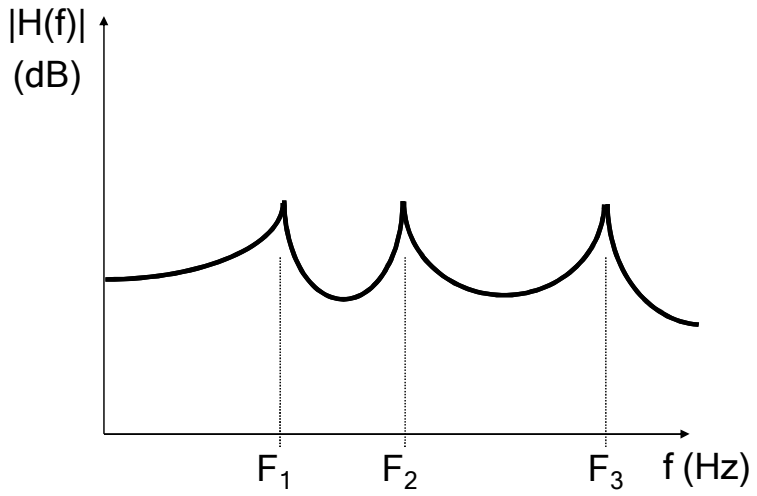

Figure 1.5: Simplified power magnitude spectrum of an idealized vocal tract. $H(f)$ is the vocal tract transfer function. The three formants F_1, F_2, and F_3 are shown.

autoregressive (AR) transfer function which has only poles. In the z-domain, the transfer function is:

$$H(z) = \frac{B}{1 - \sum_{k=1}^{p} a_k z^{-k}}, \tag{1.4}$$

where p is the filter order or number of poles, a_k is the k-th AR coefficient, and B is a constant.

In order to generate a real (non-complex) speech signal, each formant has to be represented by a complex conjugate pole pair: e.g. three formants would require an AR model of order $p = 6$, i.e. three complex-conjugate pole pairs. Typically, below 3500 Hz, there are about three formants [RJ93].

1.2.4 The speech signal

Applying the linear model of speech production with Equation 1.2, the speech spectrum $S(f)$ in Figure 1.6 is the product of its source spectrum $U(f)$ (Fig-

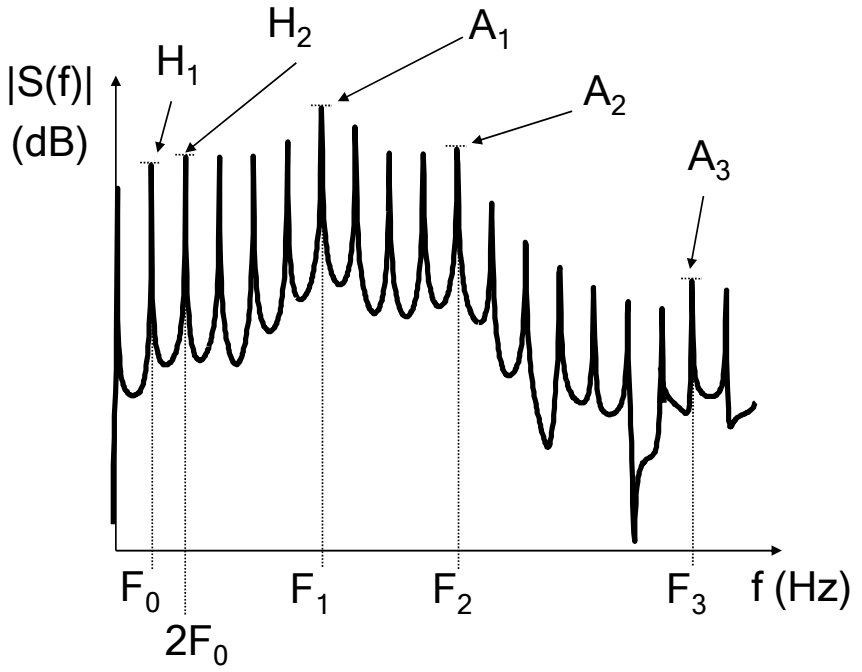

Figure 1.6: Power magnitude spectrum of an idealized speech signal. Spectral harmonic magnitudes H_1 and H_2 (in dB) at multiples of F_0. Formant frequencies F_1 to F_3 (in Hz) with their magnitudes A_1 to A_3 (in dB). Graph is simplified: formant frequencies are not necessarily at multiples of F_0.

ure 1.4) and its vocal tract transfer function $H(f)$ (Figure 1.5) for an idealized voiced speech signal. The periodicity ($T_0 = 1/F_0$) of the voice source signal results in spectral harmonic peaks at multiples of F_0: H_1 and H_2 are the magnitudes in dB of the first and second spectral harmonic peaks, respectively. The resonances of the vocal tract result in spectral formant peaks at frequencies F_1 (first formant frequency), F_2 (second formant frequency), etc., in Hz.

1.3 Recovering the voice source signal from the speech signal

To recover glottal source measures (i.e. the voice source signal) from the acoustic speech signal, vocal tract resonances need to be removed by an "inverse filtering" process. Assuming the linear source-filter model as the underlying model for speech production facilitates inverse filtering, which was first presented by Miller in [Mil59], who applied analog electronic filters to cancel the two lowest formants and the lip radiation effect from the speech pressure waveform captured by a microphone. Rothenberg [Rot73] introduced a different inverse filtering technique that measures the air flow at the mouth and nose with a special mask. This method allows the estimation of absolute flow values, including the DC component, as opposed to the inverse filtering of the pressure signal captured by a microphone, which loses the absolute zero level of flow due to the lip radiation effect. The flow measurement mask is also less sensitive to low-frequency noise and the mask's frequencies are band limited at approximately 1.6 kHz [HG92]. For all recording equipment, be it mask or microphone, it is important that its frequency magnitude response is flat and its phase response is linear from very low frequencies up to high frequencies. Compared to analog filtering, digital sampling, storage, and filtering techniques provide obvious advantages over analog techniques, since they are flexible, repeatable, easy to implement, and cause no phase distortion. Because of these advantages, today, digital inverse filtering methods are almost always used. For sampled signals, the linear source filter model of speech production equations can be written as follows:

$$s(n) = h(n) \star u(n),$$

where "$\star$" stands for convolution, n is the sample index, s(n) is the sampled (digital) speech signal, h(n) is the digital impulse response, and u(n) is the digital source signal. Transferred into the z-domain with the z-transform ($\mathcal{ZT}$) and knowing that convolution in the time domain corresponds to multiplication in the frequency domain, this yields:

$$S(z) = \mathcal{ZT}\left\{s(n)\right\} = H(z) \cdot U(z) = \frac{B}{A(z)} \cdot U(z),$$

where $S(z)$, $H(z)$, and $U(z)$, are the z-transformed $s(n)$, $h(n)$, and $u(n)$, respectively. Assume that an estimate of the vocal tract transfer function ($\tilde{H}(z) = \frac{\tilde{B}}{\tilde{A}(z)}$) is known. Then, inverse filtering is:

$$\tilde{U}(z) = S(z) \cdot \frac{\tilde{A}(z)}{\tilde{B}} = S(z) \cdot \frac{1 - \sum_{k=1}^{\tilde{p}} \tilde{a}_k z^{-k}}{\tilde{B}}.$$

Once $\tilde{U}(z)$ is found, an estimate of the voice source signal in the time domain, $\tilde{u}(n)$, can be calculated via inverse z-transform we have:

$$\tilde{u}(n) = \mathcal{ZT}^{-1}\left\{U(z)\right\}.$$

The inverse-filtering challenge is to find a good estimate of the vocal tract filter parameters $\tilde{B}$, $\tilde{a}_k$, and filter model order $\tilde{p}$, given only the speech signal $s(n)$ and $S(z)$. To find vocal tract filter parameters, typically a linear predictive coding based analysis is applied [HG92]. However, more accurate results can usually be achieved with the method of discrete all-pole modeling (DAP) introduced by [EM91]. DAP uses a cost function which is based on the Itakuro-Saito distance evaluated at the discrete frequencies of the signal power spectrum. A recent publication which uses the DAP method in combination with a code book of source functions, generated with the LF model, and an iterative optimization algorithm is described in [FMS01]. These approaches to obtaining the glottal flow waveform are computationally expensive, and often need manual correction

and tuning. Instead of trying to estimate the time domain parameters of the source models, researchers can study acoustic measures which are correlated with these parameters. This typically involves analyzing the harmonic frequencies in the speech spectrum, such as the magnitudes of the first two spectral harmonics of the source spectrum, located at the fundamental frequency F_0 and at $2F_0$, and the spectral magnitude of various formant peaks. This is less computationally intensive and less prone to error than finding the glottal flow waveform, and is therefore suited for analyzing the extensive amount of data needed for a reliable statistical evaluation. Spectral harmonics, however, are affected by both the source characteristics and by vocal tract resonances (formants). Hence, if one needs only to characterize the voice source signal properties, then the influence of vocal tract resonances, or formant frequencies, need to be compensated for [Mar65, Fan82, Fan95, Han95].

1.4 Voice source measures

Voice source measures can be either voice source model parameters, such as the source measures derived from the time domain $F_0(= 1/T_0)$, E_e, and R_k, or the source measures derived from the frequency domain $H_1^* - H_2^*$, $H_1^* - A_3^*$, and LIN. The frequency domain source measures are explained below.

The time domain source measures $F_0(= 1/T_0)$, E_e, and R_k are LF model parameters or derived from LF model parameters shown in Figure 1.3 [FLL85]. E_e relates to the spectral intensity and is measured as the amplitude of the negative peak of the differentiated glottal pulse. This value is equivalent to the amplitude at the point of maximum discontinuity in the glottal waveform for R_k values up to 0.54. $R_k = \frac{T_e - T_p}{T_p}$ is the ratio of the closing phase to the opening phase of the pulse and is related to the glottal symmetry. Another value which can be

derived directly from the LF model parameters is the open quotient $OQ = T_e/T_0$.

The frequency domain source measures $H_1^* - H_2^*$, $H_1^* - A_3^*$, and LIN are calculated from the voice source spectrum, where $H_1^* - H_2^*$ is the difference between the first two source spectral harmonic magnitudes and $H_1^* - A_3^*$ is the difference between the magnitudes of the first source spectral harmonic and the magnitude of the source spectrum at the frequency location of the third formant (F_3). The asterisk denotes that the measure has been corrected for the influence of vocal tract resonances, thus it is a source measure. LIN is the spectral linearity of the voice source spectrum and is calculated as the correlation coefficient of a linear regression analysis. The larger LIN is, the better the voice source dB-spectrum would fit to a line, the smaller the source spectral tilt, and more high-frequency components can be found [Eps02].

It was shown that $H_1^* - H_2^*$ is correlated with OQ [HHP95], though [HdD01] showed that $H_1^* - H_2^*$ is dependent on both OQ and the ratio $\alpha_m = T_p/T_e$, which they called "asymmetry coefficient". In [Han97] it was found that $H_1^* - A_3^*$ is correlated with the source spectral tilt.

Table 1.1 shows a compilation of voice source measures and their perceptual voice quality correlates (if stated in the literature). Note that recent studies have shown that perceived voice quality depends on a combination of more than one parameter [GC99].

Some of these voice source measures are correlated and variations over longer segments of voiced speech, such as shimmer and jitter are also very important voice quality indicators [KK90].

The term 'voice quality' refers to the perceptual impression related in part to the different modes of phonation, i.e. the different movement patterns of the vocal folds at the glottis. The three main types of phonation are modal, breathy,

	Name and description	**Perceptual correspondence**
F_0	Fundamental frequency: Frequency of glottal vibration cycle	Height of the voice, prosody.
E_e	Excitation strength: The amplitude of maximum negative flow derivative [FLL85].	Related to voice intensity, loudness.
R_k	Glottal symmetry/skew: The ratio of the closing phase to the opening phase of the differentiated glottal flow [FLL85].	Related to breathiness.
$H_1^* - H_2^*$	The difference between the first two source spectral harmonic magnitudes [Han97].	Related to breathiness.
$H_1^* - A_3^*$	The difference between the first source spectral harmonic magnitude and the source spectral magnitude at the frequency location of the third formant [Han97]. This measure is related to source spectral tilt.	Related to the perception of a "weak" or "dull" voice.
LIN	Spectral linearity: Correlation coefficient of a linear regression analysis of the source magnitude spectrum [Eps02]. This measures is related to high-frequency energy and inversely related to source spectral tilt.	Related to the perception of a "weak" or "dull" voice.

Table 1.1: Description of voice source measures used in this dissertation and correspondence to perceived voice quality.

and creaky [CG97]. A modal voice has small high-frequency components. The vibration of the vocal folds is periodic with a full closing of the glottis, therefore no audible friction noise is produced when air flows through the glottis. However, even in perceived modal voice, incomplete closure can be common, particularly in female speech. Breathy voice quality implies that a relatively large amount of air is used during phonation. Usually when there is breathy phonation, the glottis does not fully close during vocal fold vibration. In creaky phonation the vocal folds are strongly adducted and are of weak longitudinal tension. The frequency of vibration (F_0) is usually low. There are many more descriptions of voiced sounds, such as harsh, pressed, tense, and lax, to name a few [CG97].

1.5 Prosody and prosodic features

In connected speech, prosody serves both as a grouping function and a prominence-marking function. The groupings of words into phrases are indicated by prosodic *boundaries*, whereas the prominence of a word within a phrase is marked in English by a change in F_0 patterns (*pitch accent*, phrase focus, focal pitch accent) and the prominence of a syllable within a word is marked by *lexical stress*. English is a stress language that specifies one syllable in a word to have primary word stress. In general it is the primary stressed syllable that is pitch-accented when the word of interest is the focus of a phrase. Prosody conveys important information for understanding connected speech on word, phrase, and content levels. On the word level, the meaning of a word can be changed by placing the primary stress on a different syllable. For example, the word "subject" (stressed syllables are underlined) is a noun and means "topic" or "theme", whereas "subject" is a verb and means "to bring something under control". On the sentence level, pitch accent puts the focus on a word in a sentence. For example the declarative

sentence "John **owns** a car." could mean that John does not rent the car, or "John owns a **car**" could mean that the conversation is about John's car and not his bike (pitch-accented words are in bold). In this dissertation, we analyze the following prosodic features: lexical stress, pitch accent, and boundary tones.

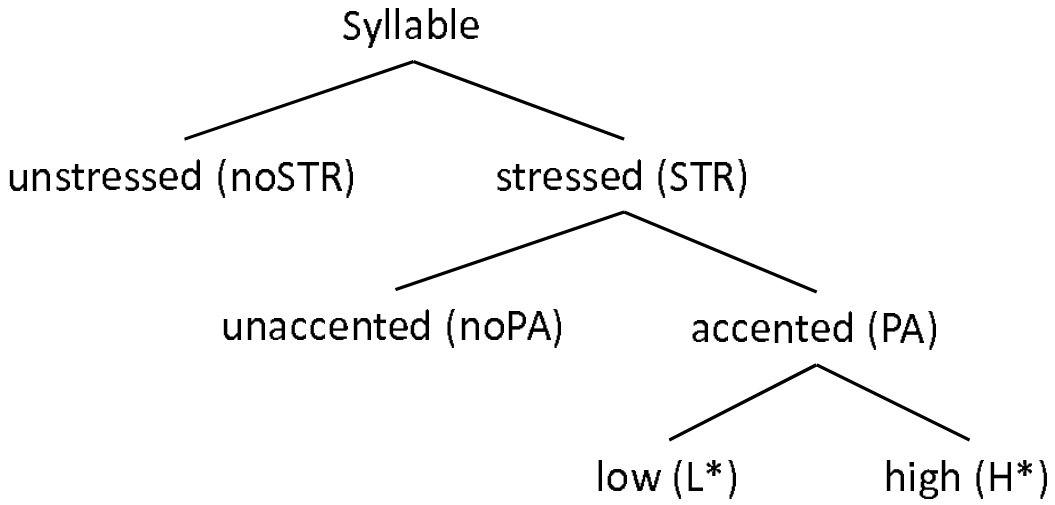

Figure 1.7: Prosodic features: tree diagram for lexical stress and pitch accent (PA, or accent). Note that unstressed syllables are always unaccented and that pitch-accented syllables are always stressed.

Figure 1.7 shows the prosodic features for stress and pitch accent as leaves of a tree diagram. Speech is analyzed on a syllable level. Syllables can either be stressed or unstressed. Stressed syllables can either be (pitch) accented or unaccented. Note that unstressed syllables are always unaccented and that pitch-accented syllables are always stressed. Pitch accent manifests itself in a change in F_0 (pitch) on the accented syllable compared to the adjacent syllables. F_0 can either be low (L*) or high (L*); a low tone or a high tone. The asterisk in L* and H* stands for accentedness and should not be confounded with the asterisk in H_1^*, H_2^*, and A_3^*. Prosodic labeling guidelines are know as ToBI (Tones and Break

Indices) labeling guidelines and can be found in [BE97]. ToBI is a system for transcribing the intonation patterns and other aspects of the prosody of English utterances. There exist more possibilities for labeling pitch accent, such as rising F_0 (R*), but here we focus only on L* and H*.

1.6 Dissertation outline

The dissertation is organized as follows.

Chapter 1 introduces the reader to the human speech production process.

Chapter 2 presents a formula to correct for vocal tract resonances without doing explicit inverse filtering. The correction formula is evaluated on synthetic and naturally-produced speech.

Chapter 3 applies the correction formula presented in Chapter 2 in order to analyze the dependencies of the voice source measures F_0, $H_1^* - H_2^*$, and $H_1^* - A_3^*$ on age, sex, and vowel context.

Chapter 4 analyzes the dependencies of voice source measures on prosodic features, such as lexical stress, pitch accent, and boundary tones. It applies the correction formula from Chapter 2 to calculate the voice source measure $H_1^* - H_2^*$ from a prosodically labeled speech corpus. Additional voice source measures, F_0, E_e, R_k, and LIN, are extracted using explicit inverse filtering.

Finally, Chapter 5 presents a summary of the dissertation and future research directions.

CHAPTER 2

A formula to correct for the influence of vocal tract resonances

The spectral magnitude of the speech signal is the result of interactions from the voice source and the vocal tract. When analyzing the voice source, vocal tract influences need to be removed from the speech signal. In this chapter, we introduce a formula to correct for the influence of vocal tract resonances. The formula, introduced in [IA04], requires no explicit inverse-filtering techniques and in addition to the fundamental frequency (F_0) and F_1, used for the correction in [Han95], the formula can take into account the frequency and bandwidth of any formant. The algorithm operates in the frequency domain, and hence does not have the time domain filtering delay and filter ringing issues. It is not restricted to non-high vowel signals and to signals with low fundamental frequency. After the presentation of the formula, its derivation is shown. The formula is then applied to estimate source spectral harmonics of synthetic and naturally-produced vowels and error and sensitivity analysis are performed. The formula appears to accurately estimate harmonic magnitudes for synthetic and naturally-produced vowel sounds.

2.1 Derivation of the correction formula

The derivation of the correction formula presented in this section is based on the linear source-filter model of speech production [Fan60]. The correction formula removes the effects of the formants on the magnitudes of the source spectrum. This is done by subtracting the amount by which the formants boost the spectral magnitudes. Theoretically, if the formant frequencies and their respective bandwidths are known exactly and the linear source-filter model is applicable, then the corrected spectral magnitudes would represent the actual magnitudes of the source spectrum. Assuming a vocal tract all-pole model, the normalized transfer function $T(s)$ with N formants can be written as

$$T(s) = \prod_{i=1}^{N} \frac{\sigma_i^2 + \Omega_i^2}{(s - (\sigma_i + j\Omega_i))(s - (\sigma_i - j\Omega_i))}. \tag{2.1}$$

The numerator of $T(s)$ is normalized such that $T(s = 0) = 1$. $s_i = \sigma_i + j\Omega_i$, $\sigma_i = -\pi B_i$, $\Omega_i = 2\pi F_i$, where B_i and F_i are the i-th formant bandwidth and frequency, respectively. Note that each vocal tract resonance frequency (formant) is modeled with a complex conjugate pole pair.

Assuming that the axis $s = j\Omega$ lies in the region of conversion (ROC), the Fourier Transform of the magnitude of Eq. 2.1 becomes

$$|T(j\Omega)| = \prod_{i=1}^{N} \left| \frac{\sigma_i^2 + \Omega_i^2}{\sigma_i^2 + \Omega_i^2 - \Omega^2 + j2\sigma_i\Omega} \right|,$$

$$|T(j\Omega)|^2 = \prod_{i=1}^{N} \frac{(\sigma_i^2 + \Omega_i^2)^2}{(\sigma_i^2 + \Omega_i^2 - \Omega^2)^2 + (2\sigma_i\Omega)^2}.$$

Using the definitions of σ_i and Ω_i produces

$$|T(f)|^2 \quad = \prod_{i=1}^{N} \frac{(\pi^2 B_i^2 + 4\pi^2 F_i^2)^2}{(\pi^2 B_i^2 + 4\pi^2 F_i^2 - 4\pi^2 f^2)^2 + 16\pi^4 B_i^2 f^2}.$$

Finally, the total contribution of N formants to the vocal tract power spectrum at frequency f is

$$|T(f)|^2 = \prod_{i=1}^{N} \frac{((B_i/2)^2 + F_i^2)^2}{((B_i/2)^2 + F_i^2 - f^2)^2 + B_i^2 f^2}. \tag{2.2}$$

For $B_i \ll F_i$ the term $(B_i/2)^2$ is often neglected [Fan95]. In this dissertation, however, we will account for the bandwidth.

The aforementioned analysis was done in the continuous frequency domain. For sampled signals (sampling frequency F_s) the contribution of N formants to the vocal tract transfer function can be written in the z domain as

$$T(z) = \prod_{i=1}^{N} \frac{1 - 2\Re(z_i) + |z_i|^2}{(z - z_i)(z - z_i^*)}, \tag{2.3}$$

where T(z) is normalized so that $|T(z=1)| = 1$. $z_i = r_i e^{j\omega_i}$ with $\omega_i = 2\pi F_i/F_s$.

Assuming that the unit circle $z = e^{j\omega}$ lies in the ROC, the Fourier Transform of the squared magnitude of Eq. 2.3 becomes

$$|T(\omega)|^2 = \prod_{i=1}^{N} \frac{\left(1-2r_i\cos(\omega_i)+r_i^2\right)^2}{\left(1-2r_i\cos(\omega-\omega_i)+r_i^2\right)\left(1-2r_i\cos(\omega+\omega_i)+r_i^2\right)}, \tag{2.4}$$

with $r_i = e^{-\pi B_i/F_s}$ and $\omega_i = 2\pi F_i/F_s$.

Eq. 2.4 specifies the amount by which the spectral magnitude at a particular frequency, ω, is boosted by the effects of formants located at frequencies ω_i. Therefore, to obtain the source spectral magnitudes, the effects of the formants need to be subtracted from the magnitudes of the speech spectrum as

$$\begin{array}{ll} H^*(\omega) & = H(\omega) - \\ \sum_{i=1}^{N} & 10\log_{10}\frac{\left(1-2r_i\cos(\omega_i)+r_i^2\right)^2}{\left(1-2r_i\cos(\omega+\omega_i)+r_i^2\right)\left(1-2r_i\cos(\omega-\omega_i)+r_i^2\right)}, \end{array} \tag{2.5}$$

where $H(\omega)$ is the magnitude of the actual signal spectrum (in dB) at frequency ω, N is the number of formants, and $H^*(\omega)$ is the corrected magnitude (i.e. the

magnitude of the source spectrum) at frequency ω. Note that for $B_i = 0$ and $\omega = \omega_i$ this formula is undefined.

For example, the corrected magnitude of the first spectral harmonic located at frequency ω_0, where $\omega_0 = 2\pi F_0/F_s$ and F_0 is the fundamental frequency in Hz, is given by

$$\begin{array}{ll} H^*(\omega_0) & = H(\omega_0) - \\ \sum_{i=1}^{N} & 10\log_{10}\frac{\left(1-2r_i\cos(\omega_i)+r_i^2\right)^2}{\left(1-2r_i\cos(\omega_0+\omega_i)+r_i^2\right)\left(1-2r_i\cos(\omega_0-\omega_i)+r_i^2\right)} \end{array} \tag{2.6}$$

with $r_i = e^{-\pi B_i/F_s}$ and $w_i = 2\pi F_i/F_s$ where F_i and B_i is the frequency and bandwidth of the i-th formant in Hz, respectively, F_s is the sampling frequency and N is the number of formants to correct for. $H(\omega_0)$ is the magnitude of the first harmonic from the speech spectrum and $H^*(\omega_0)$ represents the corrected magnitude and should coincide with the magnitude of the source spectrum at ω_0. Note that all magnitudes are in dB.

2.2 Error analysis of the correction formula

To evaluate the accuracy of the correction formula (with and without bandwidth information) in estimating spectral harmonic magnitudes, H_1 and H_2, careful error analysis is performed. All errors are calculated as "estimated" minus "actual" value, i.e. $H_{1\ est} - H_{1\ act}$, $H_{2\ est} - H_{2\ act}$, or $(H_{1\ est} - H_{2\ est}) - (H_{1\ act} - H_{2\ act})$. In Subsections 2.2.1 and 2.2.2 error analysis is done using synthetic single-, and three-formant vowels, respectively. Specifically, error analysis is evaluated for the $H_1 - H_2$ measure. For the synthetic stimuli, the LF voice source signal is filtered with an all-pole model of the vocal tract. The LF shape is defined by $T_p = 0.48$, $T_e = 0.6$, and $T_a = 0.05$, with $T_c = T_o = 1$. Subsection 2.2.3 presents error analysis results on naturally-produced vowels spoken by three children.

Analysis errors are calculated for the following cases, when applicable:

- **NoC**: Without using correction
- **F1noB1**:With correction for the influence of F_1, without using bandwidth information, that is, by setting $B_i = B_1 = 0$ in Eq. 2.6. Note that when $\omega = \omega_i$, the correction yields an infinite value (see Eq. 2.5).
- **F1B50**: With correction for the influence of F_1, setting $B_1 = 50$ Hz
- **F1B1**$_{est}$: With correction for F_1 using B_1 as calculated from Eq. 2.7
- **F1B1**: With correction for F_1, using exact bandwidth information
- **F12B12**: With correction for F_1 and F_2, using exact bandwidth information for B_1 and B_2
- **F1B1**$_{synth}$: With correction for F_1, using formant frequency and bandwidth information obtained from analysis-by-synthesis
- **F12B12**$_{synth}$: With correction for F_1 and F_2, using formant frequency and bandwidth information obtained from analysis-by-synthesis

2.2.1 Error analysis for single-formant synthetic signals

Formant correction is applied to single-formant synthetic signals with F_0 varying between 100 and 300 Hz, and F_1 between 200 and 800 Hz with constant bandwidth (B_1) of 75 Hz. Since the signals are synthetic, the actual values for H_1 and H_2 are known and the correction error between the actual and estimated harmonics' magnitudes can be calculated.

Figure 2.1 compares the $H_1 - H_2$ error at $F_0 = 250$ Hz for the cases NoC, F1noB1, and F1B1. Maximum errors for NoC and F1noB1 occur at $F_1 = F_0$ and

$F_1 = 2F_0$, where the absolute NoC error is about 24 dB and the F1noB1 error is infinite. The error for F1B1 is zero, which is expected.

Figure 2.2 shows the $H_1 - H_2$ error for the F1noB1 case as a function of F_0 and F_1. It can be seen that the maximum error occurs when F_1 is close to F_0 or $2F_0$.

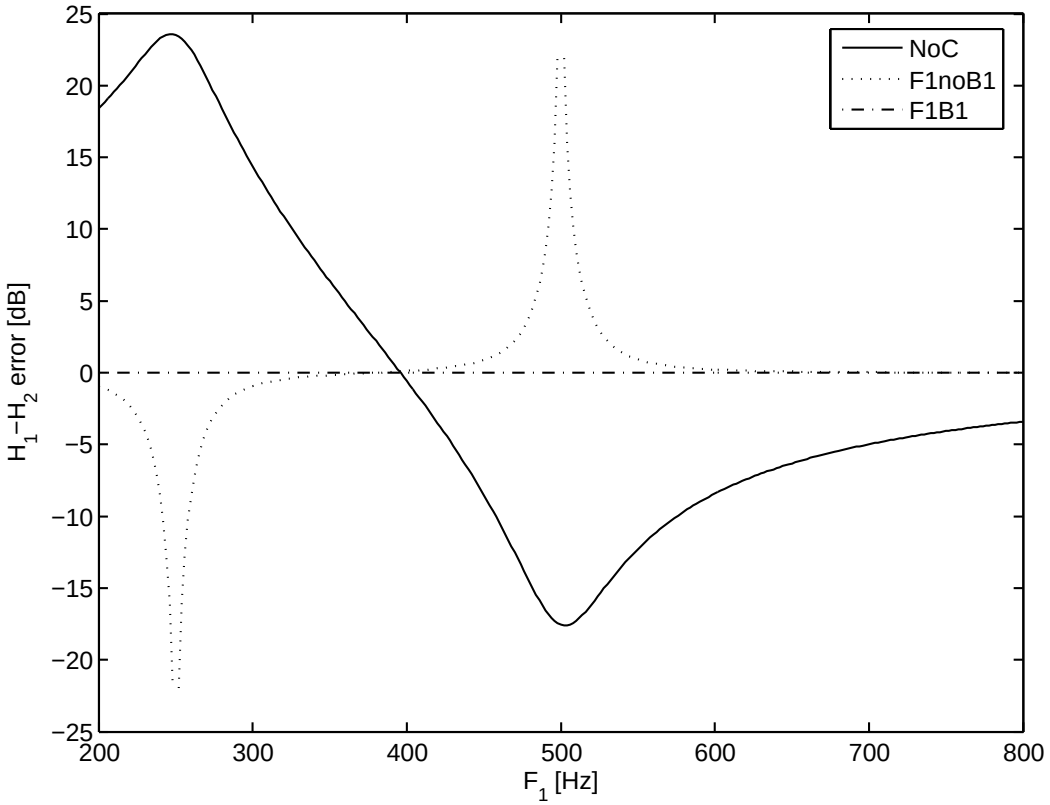

Figure 2.1: $H_1 - H_2$ error in dB with $F_0 = 250$ Hz and $B_1 = 75$ Hz for synthetic one-formant signals. The three curves represent: NoC, no correction (solid line); F1noB1, correction for F_1 not using bandwidth information (dotted line); and F1B1, correction for F_1 using exact bandwidth information (dash-dotted line). The maximum NoC error is about 24 dB. The absolute error for the F1noB1 correction at $F_1 = F_0$ and $F_1 = 2F_0$ is infinite, and the F1B1 error is zero.

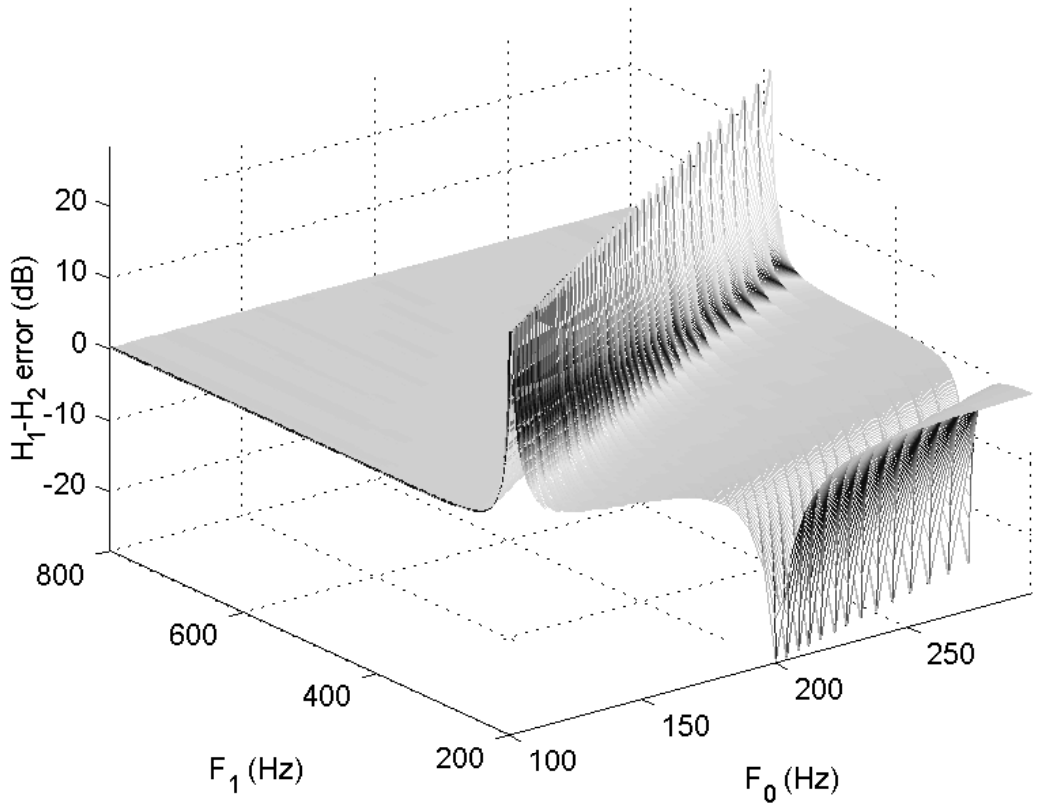

Figure 2.2: $H_1 - H_2$ error in dB using $F1noB1$ for synthetic one-formant signals. F_0 is between 100 and 300 Hz, F_1 is between 200 and 800 Hz, and $B_1 = 75$ Hz. The maximum error ($\pm\infty$) occurs at $F_1 = F_0$ and at $F_1 = 2F_0$ and is capped for display purposes. The solid line in Figure 2.1 is a vertical cut of this figure at $F_0 = 250$.

2.2.2 Error analysis for three-formant synthetic vowels

The vowels /a/, /i/, and /u/, are synthesized using the first three formant frequencies specified in [PB52]. Formant bandwidths are calculated according to the formula in [Man98]:

$$B_i = (80 + 120F_i/5000). \tag{2.7}$$

These values are shown in Table 2.1.

Vowel	F_1	F_2	F_3	B_1	B_2	B_3
Male speech						
/a/	730	1090	2440	98	106	139
/i/	270	2290	3010	86	135	152
/u/	300	870	2240	87	101	134
Female speech						
/a/	850	1220	2810	100	109	147
/i/	310	2790	3310	87	147	159
/u/	370	950	2670	89	103	144
Children speech						
/a/	1030	1370	3170	105	113	156
/i/	370	3200	3730	89	157	170
/u/	430	1170	3260	90	108	158

Table 2.1: Formant frequencies [PB52] and bandwidths [Man98] in Hz used to synthesize the three corner vowels appropriate for male, female, and child talkers.

F_0 is chosen from the ranges provided by [Bak87]: For male talkers, F_0 ranges between 85 and 154 Hz, for female talkers F_0 is between 164 and 256 Hz, and for children F_0 is between 208 and 256 Hz. The sampling frequency (F_s) is at 10 kHz.

For each sex, vowel, and correction method, the minimum, average, and maximum absolute estimation errors for $|H_1 - H_2|$ are calculated over the appropriate range of F_0. The results are shown in Table 2.2. F1noB1 introduces the highest errors especially when F_1 is close to F_0 or $2F_0$. For the vowel /a/, on the other hand, F1noB1 performs similarly to F1B1 because /a/ has a very high F_1, which is greater than $2F_0$, and hence, the influence of F_1 on the first two harmonics

is small. The errors for F1B1 are lower but are not zero since F1B1 does not correct for F_2 and F_3. The highest F1B1 errors are measured for /u/, which has the lowest F_2 of the three vowels.

Vowel	Min/Mean/Max Error in dB		
(F_1 in Hz)	NoC	F1noB1	F1B1
Male speech (F_0: 85-154 Hz)			
/a/ (730)	0.57/1.06/1.99	0.20/0.38/0.69	0.20/0.38/0.69
/i/ (270)	3.04/5.58/8.15	0.41/∞/∞	0.07/0.13/0.23
/u/ (300)	2.66/5.61/9.67	0.00/∞/∞	0.30/0.56/1.04
Female speech (F_0: 164-256 Hz)			
/a/ (850)	1.71/2.84/4.73	0.64/1.02/1.63	0.63/1.02/1.63
/i/ (310)	0.14/5.31/12.20	0.08/1.90/7.82	0.21/0.33/0.52
/u/ (370)	0.05/7.29/11.47	0.03/∞/∞	0.98/1.62/2.67
Child speech (F_0: 208-256 Hz)			
/a/ (1030)	2.05/2.59/3.25	0.86/1.07/1.33	0.83/1.04/1.30
/i/ (370)	0.36/2.91/5.97	0.01/0.73/2.15	0.29/0.36/0.44
/u/ (430)	4.60/9.59/12.48	0.23/∞/∞	1.08/1.36/1.71

Table 2.2: Min/Mean/Max $|H_1-H_2|$ error in dB without correction (NoC), correction for F_1 without bandwidth information (F1noB1), and correction for F_1 using bandwidth information (F1B1). Synthesis included three formants. As a reference, F_1 is given in parentheses for each of the vowels. It can be seen that the errors for NoC and F1noB1 are high when F_1 is close to F_0 or $2F_0$. The error for F1noB1 where $F_1 = F_0$ or $F_1 = 2F_0$ is infinite.

Figures 2.3 and 2.4 show the absolute $|H_1 - H_2|$ error as a function of F_0 for the methods NoC, F1noB1, and F1B1 for synthetic /a/ and /u/ vowels,

respectively. Figure 2.3 shows the error for the synthetic female /a/ ($F_1 = 850$ Hz) where correction without using bandwidth information (F1noB1) works well. As mentioned earlier, this is due to F_1 being much higher than F_0 or $2F_0$, hence, the first formant does not have a significant effect on the magnitudes of the first two harmonics. However, for the female /u/ (Figure 2.4), bandwidth information becomes important in the correction since $F_1 = 2F_0 = 370$ Hz when $F_0 = 185$ Hz. Hence, F1B1 yields significantly better results than F1noB1.

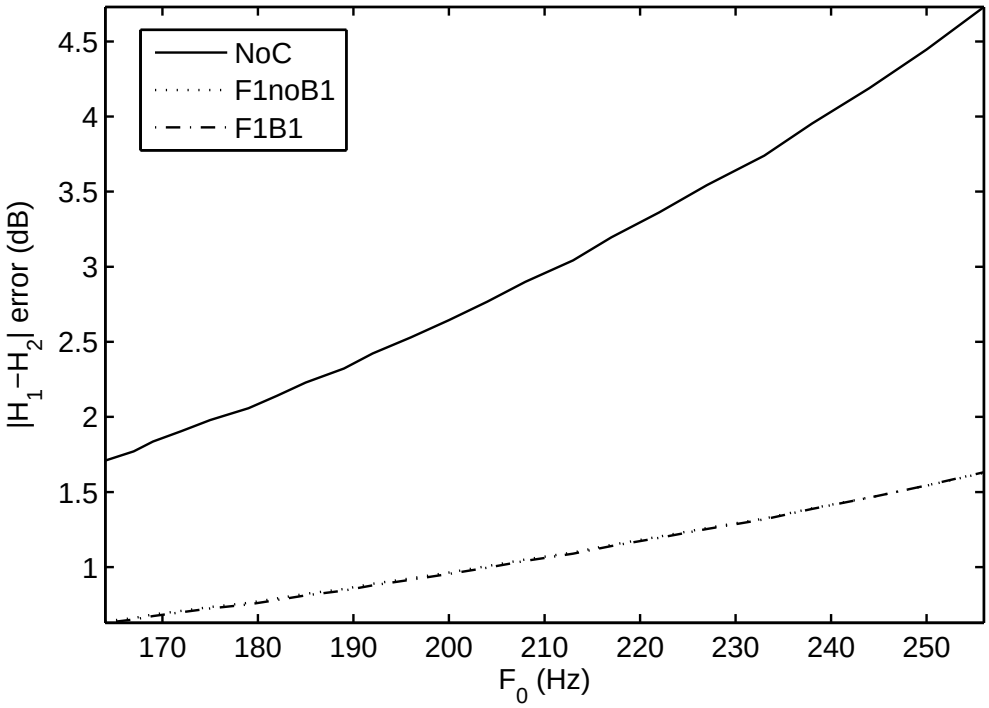

Figure 2.3: $|H_1 - H_2|$ error in dB for a three-formant synthetic female /a/ ($F_1 = 850$ Hz, $F_2 = 1220$ Hz, $F_3 = 2810$ Hz) as a function of F_0. Error using NoC (solid line), with F1noB1 correction (dotted line), and with F1B1 (dash-dotted line). In this case, using bandwidth information is not critical since F_1 is much higher than $2F_0$.

Since it is difficult to estimate bandwidths accurately [HC99], we also compare

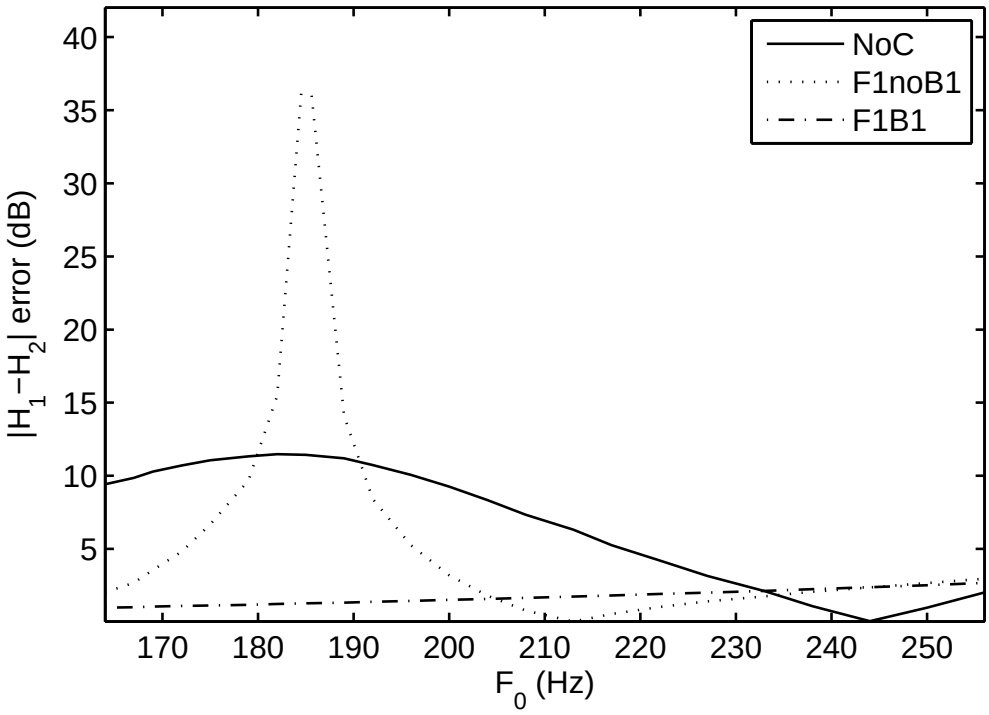

Figure 2.4: $|H_1 - H_2|$ error in dB for a three-formant synthetic female /u/ (F_1 = 370 Hz, F_2 = 950 Hz, F_3 = 2670 Hz) as a function of F_0. Error using NoC (solid line), with F1noB1 correction (dotted line), and with F1B1 (dash-dotted line). F1B1 performed significantly better than F1noB1 since F_1 is quite low. The error for F1noB1 where $F_1 = 2F_0$ is infinite.

these results with the F1B50 case, which applies the correction formula using a constant bandwidth, $B_1 = 50$ Hz. The average absolute errors for the four cases NoC, F1noB1, F1B1, and F1B50, are shown in Figure 2.5. It can be seen that the largest error occurs for the high back vowel /u/, since there is no correction for the low F_2. Using exact bandwidth information (F1B1) or using a fixed B_1 of 50 Hz improves significantly over F1noB1 for /i/ and /u/, which have low F_1. Interestingly, using a bandwidth estimate of 50 Hz (F1B50) yields similar results

to using exact bandwidth information. These results imply that for reducing errors, it is better to use some bandwidth information, even if it is only an educated guess of the true bandwidth.

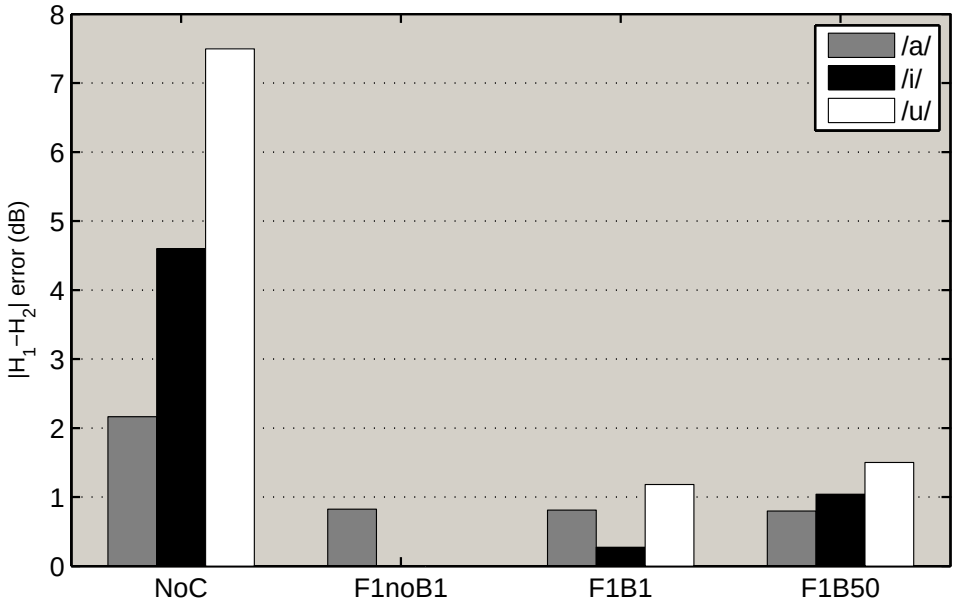

Figure 2.5: A bar diagram comparison of average $|H_1 - H_2|$ error measurements for the three synthetic, three-formant vowels (averaged over both sexes, age groups, and corresponding F_0 values.) Results for NoC, F1noB1, F1B1, and F1B50. No error bars are shown for F1noB1 for /i/ and /u/ since for some values of F_0 they can be infinite.

2.2.3 Error analysis for naturally-produced speech

The performance of the correction formula with naturally-produced speech was analyzed using samples of children's speech collected in the TBALL project [TBA04]. The speech signals were recorded with a commercial dynamic microphone (Shure type SM10A), the sampling frequency was 44.1 kHz with a

16 bit per sample representation, and the recordings were then low-pass filtered at 5 kHz. The children had to read text on flash cards presented on a laptop screen. Recordings of two seven-year old boys were chosen who spoke the single word "food" and the alphabet letter name "B". Steady state parts of the vowel segment /u/ in "food" spoken by "boy 1" and /i/ in "B" spoken by "boy 2" where then extracted and analyzed. Since none of the commonly-used inverse filtering techniques led to a reliable source estimate against which to calibrate the new correction formula, the analysis-by-synthesis approach was adopted and the resulting synthesized signals were then used for calibration. Analysis and synthesis were done "pitch synchronously", where the instants of glottal closure were derived from the linear prediction (LP) residual. In the analysis part, for each recorded signal a few fundamental periods of voiced speech were LP analyzed: The first four formant frequencies were estimated and their formant bandwidths were adjusted by hand. Note that the manual estimation of formant frequencies and bandwidths for children speech is very challenging and prone to error. With these vocal tract parameters the speech signal was then inverse-filtered to obtain a first estimate of the LF source parameters. In the synthesis part, the vocal tract and LF parameters were used to produce a synthetic speech signal. The parameters were fine-tuned so that the spectral magnitude spectrum of the synthesized signal would match the magnitude spectrum of the recorded signal. To have a reference point, the magnitude spectrum of the synthesized speech signal was offset so that H_1 of the synthesized signal was equal to H_1 of the recorded signal. Table 2.3 shows the LF source and vocal tract parameters obtained from analysis-by-synthesis of the children's speech and a comparison between the actual and the synthesized speech spectrum for "boy 1" for frequencies up to 5 kHz is shown in Figure 2.6.

Tables 2.4 and 2.5 show estimated harmonic magnitudes for the children from

LF source parameters in ms				
	$T_0 = T_c$	T_e	T_p	T_a
"boy 1" /u/	4.2	2.6	2.0	0.4
"boy 2" /i/	4.5	2.9	2.2	0.9
Vocal tract parameters (formants) in Hz				
	$F_1(B_1/B_{1\ est})$	$F_2(B_2)$	$F_3(B_3)$	$F_4(B_4)$
"boy 1" /u/	473(100/91)	1260(150)	3260(400)	4043(200)
"boy 2" /i/	399(90/90)	3189(250)	3600(200)	4530(218)

Table 2.3: LF source and vocal tract parameters obtained from analysis-by-synthesis of children speech. $B_{1\ est}$ denotes the first formant bandwidth derived from the formula in Mannell (Eq. 2.7) and used in the correction approach "F1B1$_{est}$."

the different algorithms. The values of the first five rows (NoC, F1noB1, F1B1$_{est}$, F1B1$_{synth}$, and F12B12$_{synth}$) can be compared to the last row, which displays the values obtained from analysis-by-synthesis (SYNTH). Results show that F1B1$_{est}$ performs well for all utterances. Interestingly, the correction using F1B1$_{est}$ outperforms F1B1$_{synth}$ in cases where $B_{1\ est} < B_1$ ("boy 1" /u/). It could be that in these cases, accidentally, the effect of F_2 is partially compensated for.

In Table 2.4, results for the vowel segment /u/ spoken by "boy 1" are shown. It can be seen that the frequency of the second harmonic ($2F_0 = 474$ Hz) is almost equal to F_1. Since $F_2 = 1260$ Hz is relatively low, the additional correction for F_2 should improve H_2 estimation, and indeed, the $H_1 - H_2$ error is 1 dB less for F12B12$_{synth}$ than for F1B1$_{synth}$. F1noB1 correction is adequate for the estimation of H_1, however it is not appropriate for the estimation of H_2, which introduces more than 30 dB difference to its corresponding SYNTH value. F12B12$_{synth}$ works best: its error is only 0.2 dB.

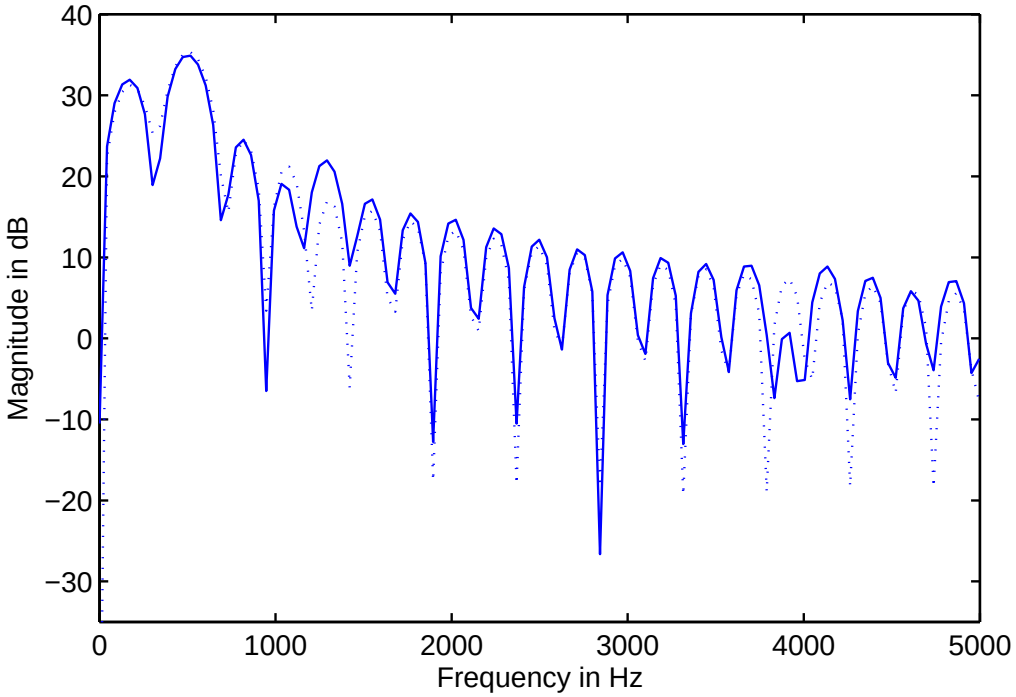

Figure 2.6: Analysis-by-synthesis: A comparison of the actual spectral magnitude of the steady-state part of the vowel segment /u/ in "food" spoken by "boy 1", with the spectral magnitude of the synthesized signal. Actual spectrum (solid line) and spectrum from analysis-by-synthesis (dotted line).

Table 2.5 contains the results for the vowel segment /i/ in "B" spoken by "boy 2". The frequency of the second harmonic ($2F_0 = 446$ Hz) is close to F_1. The table shows that F1noB1 correction is off by about 2 dB for H_2, works well for H_1, and the final value is off by about 2 dB. The corrections F1B1$_{synth}$ and F12B12$_{synth}$ are off by 0.3 dB and 0.2 dB, respectively.

	H_1	H_2	$H_1 - H_2$	$H_1 - H_2$ error
NoC	29.6	34.7	-5.1	-12.4
F1noB1	27.1	-11.2	38.3	31.0
F1B1$_{est}$	27.2	20.4	6.8	-0.5
F1B1$_{synth}$	27.2	21.1	6.1	-1.2
F12B12$_{synth}$	26.9	19.8	7.1	-0.2
SYNTH	26.8	19.5	7.3	0.0

Table 2.4: Harmonic magnitudes and their difference in dB for the vowel segment /u/ in "food" spoken by "boy 1". The corrections NoC, F1noB1, F1B1$_{est}$, F1B1$_{synth}$, and F12B12$_{synth}$ are compared to their corresponding values from analysis-by-synthesis (SYNTH, last row) and their relative errors compared to SYNTH are shown in the last column. $F_0 = 237$ Hz, $F_1 = 473$ Hz, $F_2 = 1260$ Hz, $F_3 = 3260$ Hz, $F_4 = 4043$ Hz.

	H_1	H_2	$H_1 - H_2$	$H_1 - H_2$ error
NoC	35.6	29.3	6.4	-6.6
F1noB1	32.4	17.1	15.3	2.3
F1B1$_{est}$	32.6	19.9	12.7	-0.3
F1B1$_{synth}$	32.6	19.9	12.7	-0.3
F12B12$_{synth}$	32.5	19.7	12.8	-0.2
SYNTH	32.5	19.5	13.0	0.0

Table 2.5: Harmonic magnitudes and their difference in dB for the vowel segment /i/ in "B" spoken by "boy 2". The corrections NoC, F1noB1, F1B1$_{est}$, F1B1$_{synth}$, and F12B12$_{synth}$ are compared against their corresponding values from analysis-by-synthesis (SYNTH, last row). $F_0 = 223$ Hz, $F_1 = 399$ Hz, $F_2 = 3189$ Hz, $F_3 = 3600$ Hz, $F_4 = 4530$ Hz.

2.3 Sensitivity analysis of the correction formula

In this section, the sensitivity of the correction formula to vocal tract parameter estimation errors is evaluated. The resulting spectral magnitude error introduced when applying the correction formula with an inaccurate estimate of either formant frequency or formant bandwidth is henceforth called the "correction error". Empirical results have been obtained by synthesizing one-formant signals with F_1 between 300 and 3500 Hz (in 100 Hz steps) using corresponding bandwidths from Eq. 2.7 and then calculating 1) formant estimation errors letting $F_1 + B_1 \leq F_{1\ est} \leq F_1 + B_1$ (in 1 Hz steps) with $B_{1\ est} = B_1$ and 2) bandwidth estimation errors letting $0.5B_1 \leq B_{1\ est} \leq 1.5B_1$ (in 1 Hz steps) with $F_{1\ est} = F_1$. Overall results show that correction errors are less sensitive to formant bandwidth estimation errors than to formant frequency estimation errors.

2.3.1 Sensitivity of formant correction to formant frequency estimation errors

Assume that for a synthetic one-formant signal with formant frequency $F_1 = 1500$ Hz and formant bandwidth $B_1 = 200$ Hz the estimated formant frequency is 1600 Hz ($F_{1\ est} = F_1 + 100$ Hz). For simplicity, it is assumed that the formant bandwidth was estimated correctly ($B_{1\ est} = B_1$). Figure 2.7 depicts the power spectral magnitudes of the actual signal, the estimated signal, and the correction error (difference between the actual and estimated spectra). The error curve shows maxima around $f \approx F_1$ (maximum) and $f \approx F_{1\ est}$ (minimum). Thus, when $F_{1\ est} \neq F_1$, evaluating the correction formula around the actual formant frequency F_1 or around the estimated formant frequency $F_{1\ est}$ will introduce maximum errors. Towards low frequencies the error reduces to zero and towards high frequencies it remains at a low constant level.

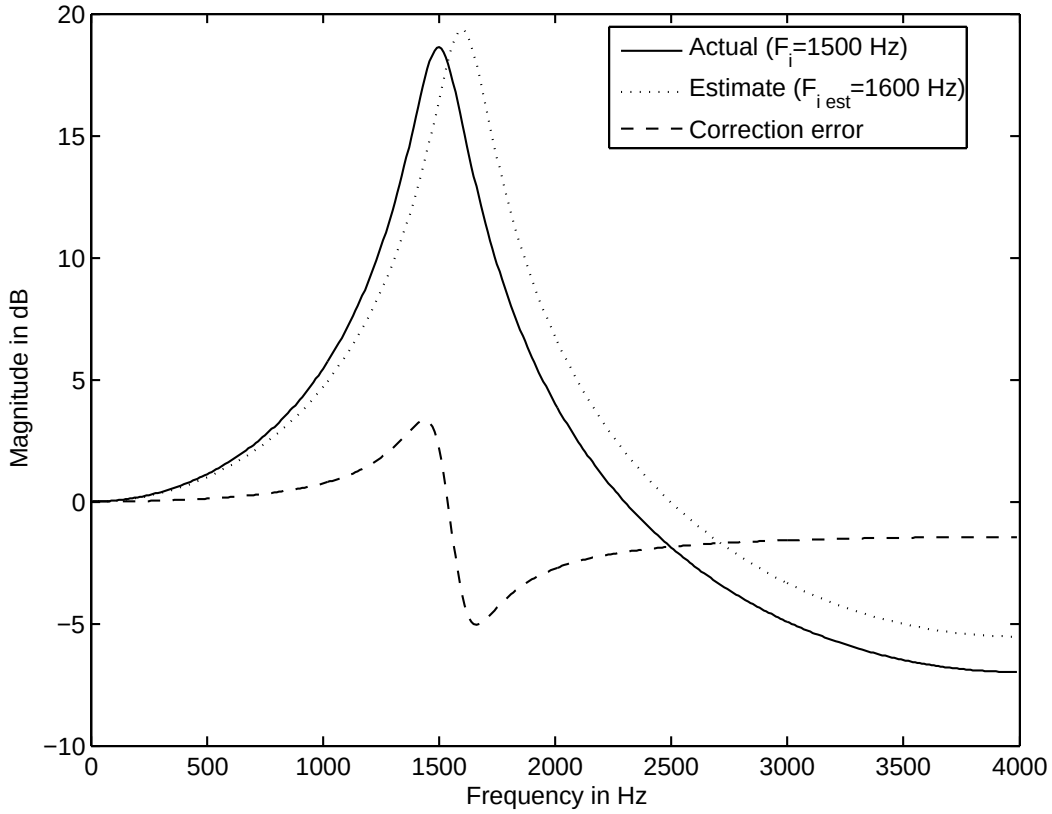

Figure 2.7: Power spectral magnitude for a one-formant signal ($F_1 = 1500$ Hz and $B_1 = 200$ Hz), the estimated signal ($F_{1\ est} = 1600$ Hz and $B_{1\ est} = B_1$), and the resulting correction error.

Results show that the maximum correction error depends primarily on the formant bandwidth (B_i) and on the difference $\Delta = F_{i\ est} - F_i$ and not so much on the formant frequency F_i. As an approximation, the absolute maximum correction error in dB can be linearized for small Δ:

$$|E_{max}| \approx \frac{10\ dB}{B_i}|\Delta|, \qquad \text{for } |\Delta| \leq \frac{B_i}{2} \tag{2.8}$$

Figure 2.8 shows the minimum, average, and maximum absolute correction error as a function of the absolute difference ($|\Delta|$) between estimated and actual formant frequency for $F_i = 1500$ Hz, $B_i = 200$ Hz, and $|\Delta| \leq B_i/2$. Note that the

maximum correction error estimated with the approximation formula in Eq. 2.8 is close to the calculated error and that the minimum error is about 0 dB. From the approximation formula it can be seen that the maximum correction error is proportional to the absolute formant frequency error ($|\Delta|$) and inversely proportional to the formant bandwidth B_i. As an example, assume that $B_i = 100$ Hz. According to Eq. 2.7, a 100 Hz error in formant frequency estimation would introduce a 10 dB correction error.

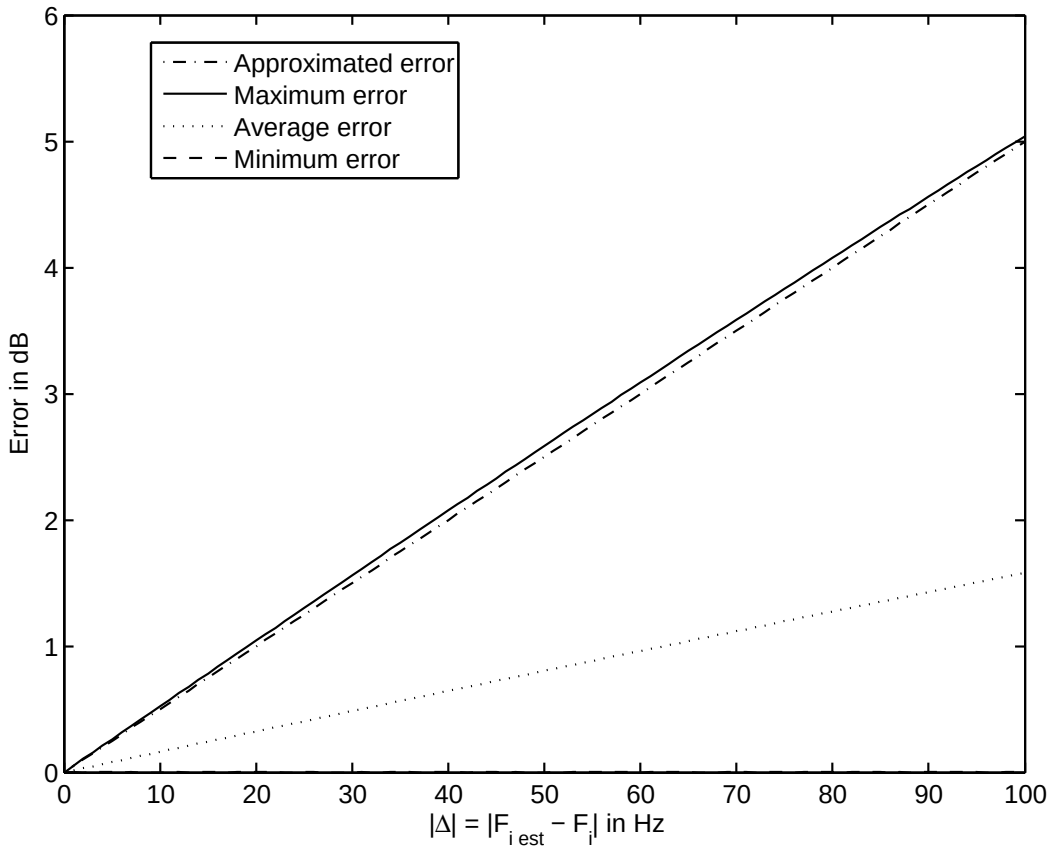

Figure 2.8: Minimum, average, and maximum absolute correction error (in dB) as a function of the absolute difference between estimated and actual formant frequency ($|\Delta|$ in Hz). $F_i = 1500$ Hz, $B_i = 200$ Hz, and $|\Delta| \leq B_i/2$.

2.3.2 Sensitivity of formant correction to formant bandwidth estimation errors

For a synthetic one-formant signal with formant frequency $F_i = 1500$ Hz and formant bandwidth $B_i = 200$ Hz, the correction error introduced by a formant bandwidth estimation error ($B_{i\ est} = B_i - 50$ Hz $= 150$ Hz) is shown in Figure 2.9. For simplicity, it is assumed that the formant frequency was estimated correctly ($F_{i\ est} = F_i$).

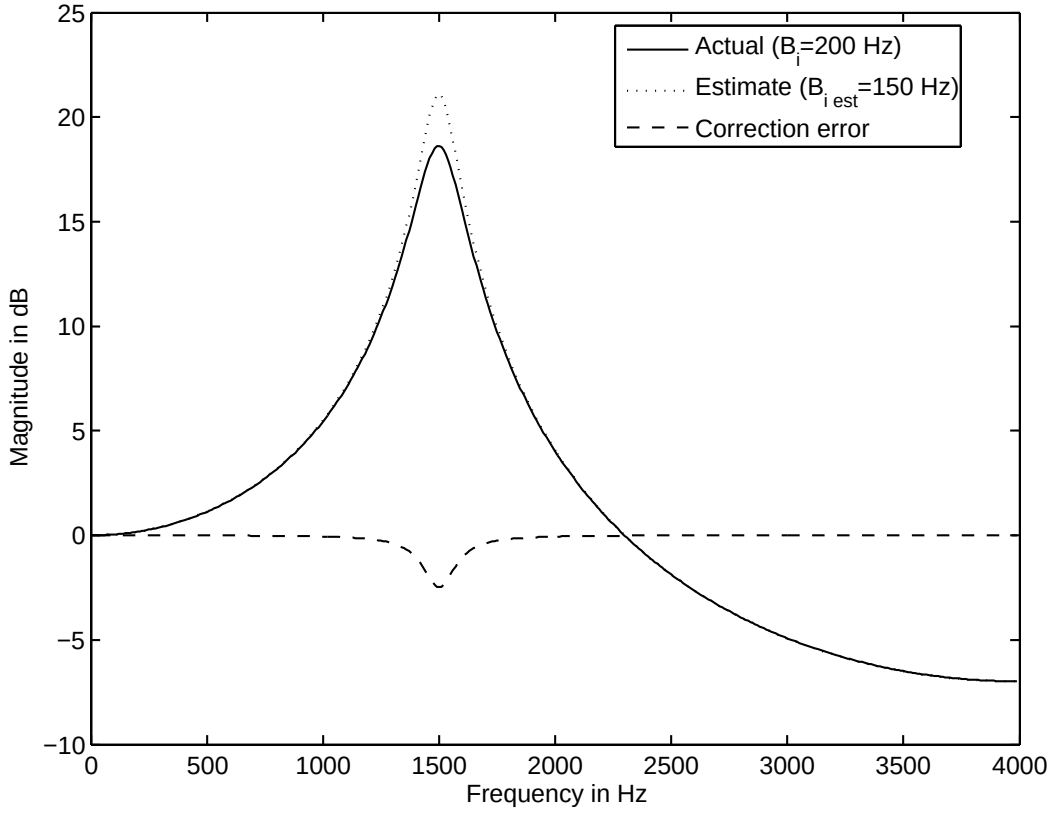

Figure 2.9: Bandwidth estimation error: Power spectral magnitude for a one formant signal ($F_i = 1500$ Hz, $B_i = 200$ Hz), the estimated signal ($F_{i\ est} = F_i$, $B_{i\ est} = 150$ Hz), and the resulting error (difference).

Clearly, the maximum correction error is introduced at exactly the formant frequency $f = F_i$. In order to find an approximation formula for the maximum correction error, we can evaluate Eq. 2.2 at $f = F_i$, assuming that $B_i \ll F_i$:

$$E_{max} \approx 10 \log_{10} \left(\frac{|T(f = F_i, B_i = B_{i\ est}|^2}{|T(f = F_i)|^2} \right).$$

This yields:

$$E_{max} \approx 10 \log_{10} \left(\frac{F_i^2}{B_{i\ est}^2} \cdot \frac{B_i^2}{F_i^2} \right),$$

and we finally get the approximation formula for the maximum correction error in dB:

$$E_{max} \approx 20 \log_{10} \left(\frac{B_i}{B_{i\ est}} \right) \qquad B_{i\ est} \neq 0. \tag{2.9}$$

The maximum correction error introduced by a bandwidth estimation error occurs at $f = F_i$ but is independent of the actual formant frequency value F_i.

Figure 2.10 shows the minimum, average, and maximum absolute correction error as a function of the estimated bandwidth relative to the actual bandwidth in percent $(100 * (B_{i\ est} - B_i)/B_i)$. Note that the maximum correction error estimated with the approximation formula in Eq. 2.9 is close to the calculated error and that the minimum error is about 0 dB. The fact that the maximum correction error curve is not a straight line (the error at -50% bandwidth change is about +6 dB whereas at +50% it is only about -3 dB) illustrates that bandwidth overestimation usually produces less absolute correction error. For $B_{i\ est} = 0$ Hz (not shown in the graph) the error amounts to 51.2 dB.

2.4 Summary

When estimating voice source measures, such as the magnitude of the first two source spectral harmonics, the vocal tract influence on the source spectrum needs

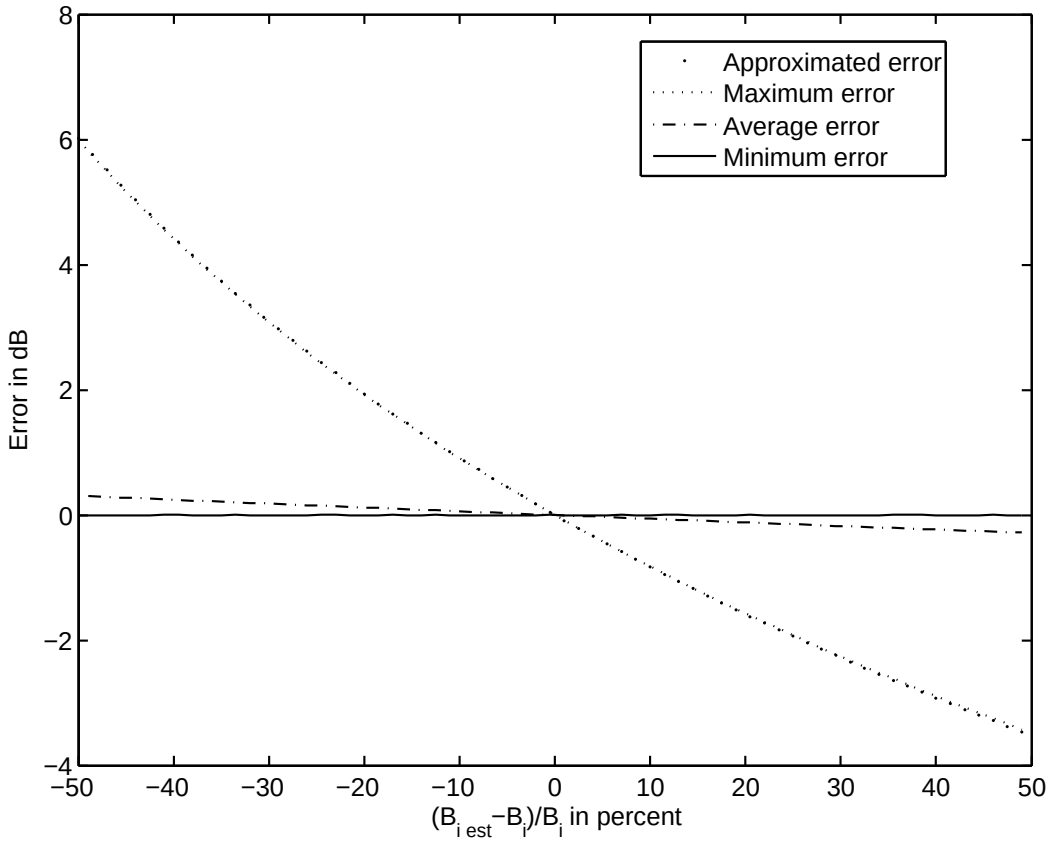

Figure 2.10: Minimum, average, and maximum correction error (in dB) for estimated bandwidth $B_{i\ est}$ in the range from -50% to +50% of actual bandwidth B_i. Note that the approximation formula for the maximum error is very close to the calculated maximum error (curves lie on top of each other).

to be compensated for. A correction formula which corrects for the influence of the vocal tract resonances was presented in this chapter and its importance, especially when applied to high vowels and to high-pitched voices, is shown. To validate the correction formula, it was applied to synthetic and to naturally-produced speech tokens. Synthetic speech was produced with formant frequencies from [PB52] and corresponding bandwidths were calculated using Eq. 2.7. The performance of the correction formula for naturally-produced speech was calibrated with the

analysis-by-synthesis method.

Error analysis of the correction formula, when applied to estimate the first two source spectral harmonics (H_1^* and H_2^*), shows that it is better to use an educated formant bandwidth guess when correcting for the vocal tract influence, rather than not using bandwidth information (i.e. setting $B_i = 0$ in Eq. 2.5). Examples of synthetic vowels show that correction without using bandwidth information can yield larger errors than no correction at all. Sensitivity analysis of the formula to vocal tract estimation errors shows that the maximum correction error can be approximated by Eq. 2.8 for small formant frequency estimation errors and by Eq. 2.9 for formant bandwidth estimation errors.

In conclusion, when estimating voice source measures it is recommended to apply the correction formula, preferably using a bandwidth overestimate, rather than an underestimate or zero bandwidth.

CHAPTER 3

Dependencies of voice source measures on age, sex, and vowel

The effects of age, sex, and vocal tract configuration on the glottal excitation signal in speech are only partially understood, yet understanding these effects is important for both recognition and synthesis of speech as well as for medical purposes. It was shown in [HC99] that the acoustic characteristics of the voice source signal are gender dependent and that open quotient and source spectral tilt are generally higher for adult female than for adult male talkers. Speech acoustics are also affected by age, which was shown in a study by [LPN99]. The study analyzed the fundamental frequency (F_0) and formant frequencies for a large speech database [MLU96] with about 490 subjects in the age range of 5 - 50 years. The study showed that children have higher F_0 and formant frequencies, and greater temporal and spectral variability than adults. These findings are attributed to vocal-tract anatomical differences and possible differences in the ability to control speech articulators. This chapter shows the application of the correction formula presented in Chapter 2 to uncover age, sex, and vowel dependencies of the three voice source measures, F_0, $H_1^* - H_2^*$, and $H_1^* - A_3^*$ in a relatively large speech database. Experimental results show that the three voice source measures are dependent to varying degrees on age and vowel. Age

dependencies are more prominent for male talkers, while vowel dependencies are more prominent for female talkers suggesting a greater vocal tract-source interaction. All talkers show a dependency of F_0 on sex and on F_3, and of $H_1^* - A_3^*$ on vowel type. For low-pitched talkers ($F_0 \leq 175\ Hz$), $H_1^* - H_2^*$ is positively correlated with F_0 while for high-pitched talkers, $H_1^* - H_2^*$ is dependent on F_1 or vowel height. For high-pitched talkers there were no significant sex dependencies of $H_1^* - H_2^*$ and $H_1^* - A_3^*$.

3.1 Speech data

Speech signals recorded from 185 males and 150 females of ages 8, 9, 10, 11, 12, 13, 14, 15, 18, and age group 20–39, from the CID database [MLU96] were analyzed. The vowels /ih/, /eh/, /ae/, /uw/, and /iy/, corresponding to the consonant-vowel-consonant (CVC) words 'bit', 'bet', 'bat', 'boot', and 'bead', were presented in the carrier sentence "I say uh, CVC again". 'uh' was used before the target word to maximize vocal tract neutrality. Most utterances were repeated twice by each speaker. Recordings were made at normal habitual speaking levels with a sampling frequency of 16 kHz. In total, 3145 utterances were analyzed. The age and sex distribution of the analyzed talkers is shown in Table 3.1.

3.2 Methods

The calculation of the three voice source measures requires the estimation of the first three formant frequencies (F_1, F_2, F_3), their respective bandwidths (B_1, B_2, B_3), and F_0. Formant frequencies F_1, F_2, and F_3, as well as F_0 were estimated using the "Snack Sound Toolkit" software [Sj04]. For H_1^* and H_2^*, the correction was for the first and second formant (F_1 and F_2) influence with $N = 2$ in Eq. 2.5.

Age	M	F	Age	M	F
8	25	11	13	16	13
9	24	25	14	11	10
10	25	14	15	11	11
11	24	19	18	10	10
12	22	21	20–39	17	16

Table 3.1: Number of talkers analyzed in each age group separated by sex (males: M; females: F).

For A_3^*, the first three formants were corrected for ($N = 3$) and there was no normalization to a neutral vowel; recall that our correction accounts for formant frequencies and their bandwidths. Figure 3.1 depicts a flowchart of this process.

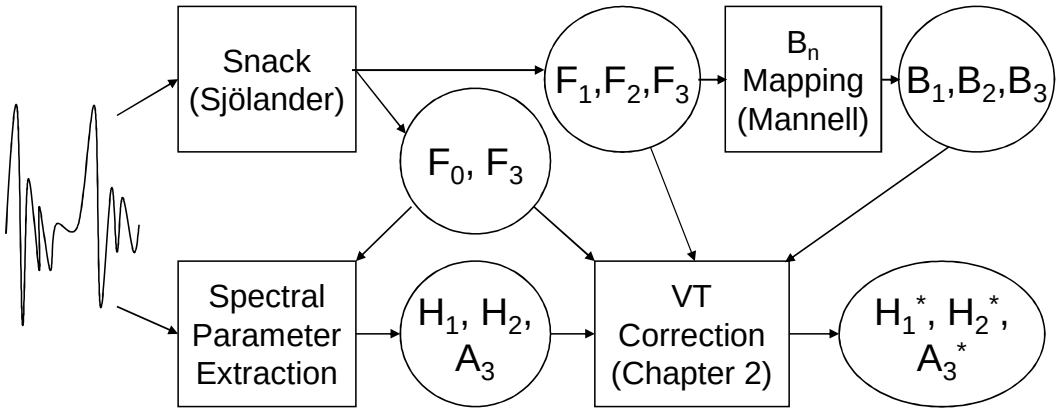

Figure 3.1: Flow chart of feature extraction process. The voice source measures H_1^*, H_2^*, and A_3^* are extracted using the correction formula (VT correction) presented in Chapter 2.

The main parameters that can be changed in Snack are frame length, frame shift, and analysis methods. For formant estimation, the covariance method was chosen because of its accuracy. F_0 was extracted with the ESPS method

(Entropic Signal Processing System), which is a component of Snack. Additional settings were: The pre-emphasis coefficient was 0.9, the length of the analysis window was 25 ms, and the window shift was 10 ms. Using the values extracted with Snack, the amplitudes H_1, H_2, and A_3 were estimated from the speech spectrum. Since the Snack bandwidth estimates varied greatly within the analysis segments and were sometimes unrealistic, all bandwidths were calculated from their corresponding formant frequency using Eq. 2.7. This reduced the bandwidth variance and therefore the variance of bandwidth-dependent results. Analysis segments were chosen at the steady-state part of the vowel, where the context-influence was smaller than in other segments.

The estimates of F_0, F_1, F_2, and F_3 were manually checked for every utterance by viewing the spectrogram, time waveform, and LPC spectral slices. Most formant estimation errors occured with child speech. For example, for high pitched /iy/, Snack typically allocated two formants to the first spectral peak resulting in a much lower second formant frequency. The number of formant estimate corrections in percent, for 8 year old children, was: 86% for /iy/, 44% for /eh/, 32% for /ih/, and 2% for /uw/. Table 3.2 lists the percentage of manual F_0, F_1, F_2, and F_3 corrections by age and sex. The formant values are not listed here as the results are similar to those reported in [LPN99].

3.3 Results

In this section, we refer to males and females from ages 8 to 14, and females 15 years and older as "Group 1", and to male talkers age 15 and older as "Group 2". Group 1 talkers were typically high-pitched (with $F_0 > 175$ Hz) and Group 2 talkers were generally low-pitched (with $F_0 \leq 175$ Hz), although there were F_0 outliers within both groups. The voice source measures F_0, $H_1^* - H_2^*$, and $H_1^* - A_3^*$

Age	Gender	F_0	F_1	F_2	F_3
8	M	0	0	16	24
	F	5	2	19	30
9	M	1	2	14	18
	F	1	3	14	20
10	M	3	1	8	13
	F	4	1	9	14
11	M	2	1	15	18
	F	0	0	9	9
12	M	1	0	15	17
	F	0	0	12	17
13	M	0	1	19	23
	F	5	8	19	20
14	M	0	0	15	16
	F	0	0	9	21
15	M	0	0	20	21
	F	0	0	11	15
18	M	0	0	23	28
	F	7	2	12	18
20-39	M	1	1	23	28
	F	8	5	24	25

Table 3.2: Percentage of manual $F_0/F_1/F_2/F_3$ frequency corrections over all vowels.

were analyzed as a function of age, sex, and vowel type, and their intercorrelations were studied.

3.3.1 Analysis of variance of the three voice source measures

Statistical analysis was performed on the extracted voice source measures by using the three-way analysis of variance (ANOVA) test in the software package

SPSS (v13.0). The factors age (ages 8, 9, 10, 11, 12, 13, 14, 15, 18, and 20-39), sex (M, F) and vowel-type (/iy/, /ih/, /eh/, /ae/, and /uw/) were tested against the variables F_0, $H_1^* - H_2^*$, and $H_1^* - A_3^*$. These factors were tested with: a) all the talkers, b) the talkers separated by sex and c) the talkers separated into Group 1 (children and females, generally high-pitched) and Group 2 (older males, generally low-pitched). Tests where the null hypothesis had a probability of $p < 0.001$ were considered to be statistically significant. This stringent criterion was selected because the statistical tests were highly sensitive due to the large number of degrees of freedom in the analyses. In addition, Pearson correlation coefficients were calculated to test for statistically significant intercorrelations between the three voice source measures.

Table 3.3 shows an overview of the results obtained with a three-way ANOVA for all the talkers showing the F value (ratio of the model mean square to the error mean square) and partial η^2 (calculated as $SS_{effect}/(SS_{effect} + SS_{error})$, where SS_{effect} is the sum of squares of the effect and SS_{error} is the sum of squares of the error). Partial η^2 is a measure of effect size. For all three measures the effect size is greatest with age. For $H_1^* - H_2^*$ and $H_1^* - A_3^*$, the effect size of age is followed by vowel and sex, while for F_0, vowel type shows the smallest effect size. It can be seen that there are a number of factors that are statistically reliable and still have quite low values of F and η^2. These factors contribute complexity to the model without increasing its explanatory power. For example, the effect of sex on $H_1^* - A_3^*$ is very low and it would be expected that adult females have higher spectral tilt than males [HC99]. The reason is that we do not distinguish between adults and children. This problem can be partially solved by splitting up the data to reduce the interactions such as age with sex, vowel with sex, or age with vowel. The data can be split up into male and female talkers, which still does not distinguish between adults and children, or in low-pitched and high-pitched

talkers, which might leave traces of interactions with sex.

Table 3.4 shows ANOVA results when the talkers were separated by sex. It can be seen that across all three voice source measures, the effect size of age is greater for males than for females. This was expected since these measures, for example F_0 [LPN99], vary more substantially with age for male talkers. However, for vowel-type, the effect size is greater for females than for males. This may suggest a greater vocal tract-source interaction for female talkers. For males, an effect of age and vowel interaction on $H_1^* - H_2^*$ can be seen. Given the growth of the larynx and the vocal tract during puberty, it would be expected that all three source measures would be affected.

The results are also interesting when viewed in terms of Group 1 and Group 2 talkers. Table 3.5 shows the F and partial η^2 results for both groups. For Group 1 talkers, it can be seen that nearly all the entries are statistically significant except when sex is tested against $H_1^* - H_2^*$ and $H_1^* - A_3^*$. This result suggests that females of all age groups have a similar OQ and source spectral tilt compared to boys (ages 8 to 14). An interaction effect of age and sex on all three source measures would be expected, since Group 1 contains boys whose articulatory control is limited due to vocal mutation; however, only F_0 is significantly affected. The results for the Group 2 talkers have only one significant entry: vowel type versus $H_1^* - A_3^*$. The lack of any age effect for Group 2 talkers is likely due to the fact that source characteristics for males do not change significantly with age above 15 years old; this has been shown for F_0 in [LPN99]. Sex was not included for Group 2 analysis since all talkers in that group were male.

Table 3.6 shows the Pearson correlation coefficients (PCC's) when the three voice source measures were tested against each other. Although the intercorrelations are statistically significant, there is only one PCC greater than 0.7,

	df	F_0	$H_1^* - H_2^*$	$H_1^* - A_3^*$
Age	9	235.0(0.410)	23.9(0.066)	35.0(0.094)
Sex	1	1012.3(0.250)	57.7(0.019)	*4.1(0.001)*
Vowel	4	28.0(0.036)	52.7(0.065)	68.9(0.083)
Age * Sex	9	95.4(0.220)	14.1(0.004)	5.2(0.015)
Vowel * Sex	4	*1.6(0.002)*	6.3(0.008)	*2.1(0.003)*
Age * Vowel	36	*0.4(0.005)*	2.2(0.026)	*1.9(0.022)*

Table 3.3: Overview table for a three-way ANOVA for all talkers showing F and partial η^2 values (in parentheses). Degree of freedom: df. Degree of freedom for the error is 3045. Values in italics are statistically insignificant ($p \geq 0.001$).

	df	F_0	$H_1^* - H_2^*$	$H_1^* - A_3^*$
Females				
Age	9	26.8 (0.151)	3.4 (0.022)	9.7 (0.060)
Vowel	4	20.4 (0.057)	50.0 (0.128)	36.7 (0.097)
Males				
Age	9	314.3 (0.627)	33.3 (0.151)	34.4 (0.155)
Vowel	4	9.0 (0.021)	11.4 (0.026)	33.7 (0.074)
Age * Vowel	36	–	2.8 (0.057)	–

Table 3.4: ANOVA results for female and male talkers showing F and partial η^2 values (in parentheses). Statistically insignificant values ($p \geq 0.001$) are not shown or are marked with a dash "–". Degree of freedom: df. df for the error is 1359 for females and 1686 for males.

indicating a strong correlation. This occurs for the relationship between $H_1^* - H_2^*$ and F_0 for Group 2 talkers.

	df	F_0	$H_1^* - H_2^*$	$H_1^* - A_3^*$
		Group 1		
Age	2	78.7 (0.208)	3.9 (0.013)	17.2 (0.054)
Sex	1	167.9 (0.059)	–	–
Vowel	4	26.1 (0.037)	75.9 (0.101)	65.1 (0.088)
Age * Sex	6	28.0 (0.059)	–	–
		Group 2		
Age	6	-	-	-
Vowel	4	-	-	6.5 (0.069)

Table 3.5: ANOVA results for Group 1 (children and females) and Group 2 (older males) talkers showing F and partial η^2 values (in parentheses). Statistically insignificant values ($p \geq 0.001$) are not shown or are marked with a dash "–". Sex is not included in the analysis for Group 2 since that group comprises of only male talkers. df for the error is 2697 for Group 1 and 348 for Group 2.

3.3.2 F_0

Table 3.7 shows the range of F_0 values for all talkers. Note that F_0 was not normalized for lexical stress. For males the mean F_0 drops by about 130 Hz between ages 8 and 20 with the largest drop between ages 12 and 15 (105 Hz), while the change is less dramatic for female talkers (overall about 50 Hz). These changes are reflected in Table 3.4 which shows that age has a greater effect size on F_0 for males (F/partial $\eta^2 = 314.3/0.627$) than for females (F/partial $\eta^2 = 26.8/0.151$). As expected, adult females exhibit higher F_0 values than adult male talkers: The difference in the means is about 110 Hz. These trends agree with the results in [LPN99]. We noticed that a few very high F_0 values (above 300 Hz) were due to strong emphasis on the target word.

	F_0	$H_1^* - H_2^*$	$H_1^* - A_3^*$
Group 1			
F_0	1		
$H_1^* - H_2^*$	-0.471	1	
$H_1^* - A_3^*$	-0.356	0.532	1
Group 2			
F_0	1		
$H_1^* - H_2^*$	0.767	1	
$H_1^* - A_3^*$	0.268	0.473	1

Table 3.6: Pearson correlation coefficients (PCC's) for F_0, $H_1^* - H_2^*$ and $H_1^* - A_3^*$ for Group 1 and Group 2 talkers. Correlation coefficients greater than 0.7 indicate strong correlations. All results are statistically significant.

Average F_0 values are highest for /uw/, and higher for /iy/ than for /eh/ and /ae/. The trend of increasing F_0 as the tongue moves from a front to a back position and from open to closed vowels, has been reported for German talkers [Mar96]. This trend can be seen for all ages and genders for the vowels in this study and may partly be explained by vowel-dependent intrinsic pitch [LP61]. ANOVA results in Table 3.4 indicate that although these trends are statistically significant for both males and females, the partial η^2 values, and hence the effect sizes of vowel type, are relatively small for both sexes: F/partial $\eta^2 = 20.4/0.057$ for females and $9.0/0.021$ for males. Interestingly, the vowel effect size on F_0 is about three times higher for females. A further analysis into the vowel dependency was done by performing an ANOVA test on the effects of high and low formant frequencies (thresholds at the formant means) on F_0. It was found that F_0 was positively correlated only with F_3 for all talkers and

Age	F_0 males in Hz	F_0 females in Hz
8	170/255/420	152/283/423
9	160/264/454	187/267/437
10	141/256/407	146/266/367
11	167/256/378	185/254/494
12	125/230/328	178/236/338
13	119/190/285	180/251/394
14	101/177/272	169/228/293
15	95/125/251	179/228/310
18	84/129/239	199/246/310
20–39	88/127/191	156/235/356

Table 3.7: Min/Mean/Max of F_0 (in Hz) per age group for vowels in the target syllables.

this correlation was statistically significant (F/partial $\eta^2 = 133.1/0.041$); again the effect size was relatively small. This positive correlation can be explained by the fact that F_3 is typically correlated with vocal tract length [Wak77]. Hence, a higher F_3, which typically results from a shorter vocal tract, coincides with a higher F_0.

3.3.3 $H_1^* - H_2^*$

The effects of age and sex on $H_1^* - H_2^*$ (related to open quotient) are shown in Figure 3.2. Comparing the values, it is interesting to observe that the $H_1^* - H_2^*$ (mean value) separation between the genders is the clearest at age 15 (5.8 dB). Between ages 8 and 20–39, the mean $H_1^* - H_2^*$ value drops by about 4 dB for male talkers, whereas for female talkers it remains relatively unchanged. Having

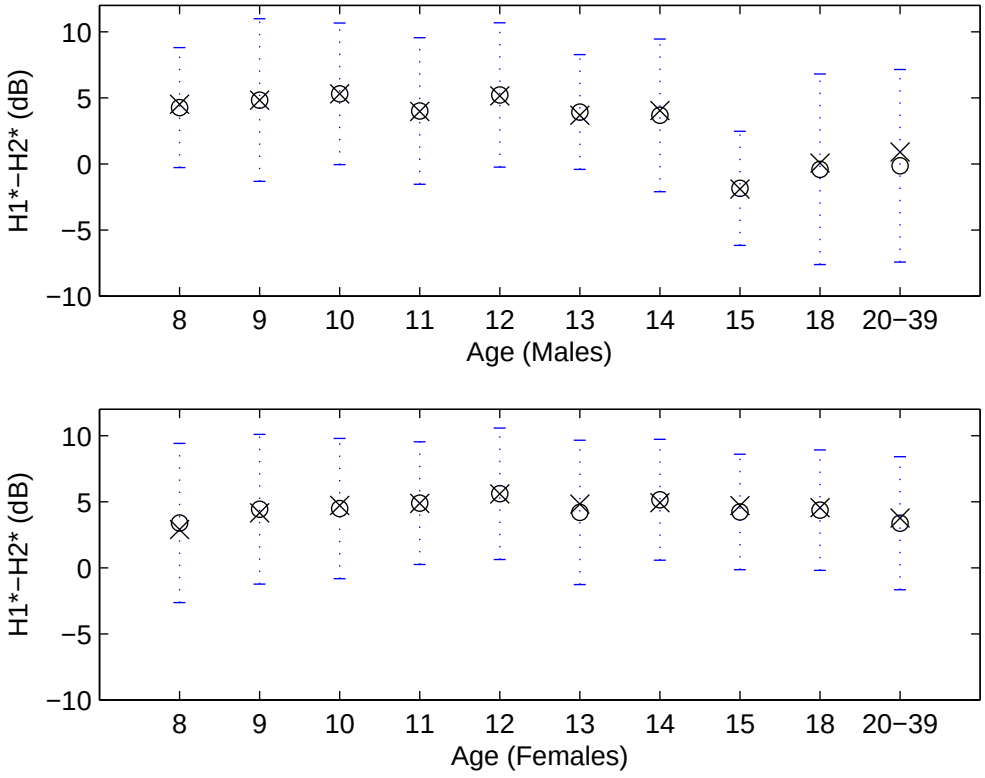

Figure 3.2: $H_1^* - H_2^*$ versus age, separated by sex. Between age 8 and 20–39, $H_1^* - H_2^*$ drops by about 4 dB for males, while for females there is little change. The largest difference between the sexes appears at age 15 where the difference in the means approaches 6 dB. Mean, median, and standard deviation are represented by circles, crosses, and whiskers, respectively.

smaller changes in $H_1^* - H_2^*$ with age is reflected in the statistical analysis of Table 3.4 where the effects of age are less pronounced for females: F/partial $\eta^2 = 3.4/0.022$ vs. $33.3/0.151$ for male talkers. The difference between genders may be related to the fact that F_0 drops significantly between age 12 and 15 for males while it does not change as much for females [LPN99]. Adult females exhibit higher mean $H_1^* - H_2^*$ values (about 3.4 dB) than adult male talkers. A

similar difference (3.1 dB) between adult male and adult female talkers was found in [HC99]. When the talkers are split into Group 1 and Group 2 categories (see Table 3.5), it is interesting to note that the dependence on sex is not significant for Group 1 talkers (children and females).

Vowel effects are larger for female talkers than for males as shown in Table 3.4 (F/partial $\eta^2 = 50.0/0.128$ for females vs. $11.4/0.026$ for males), which suggests a greater vocal tract source interaction for females. When analyzed against Group 1 and Group 2, the results in Table 3.5 indicate that only Group 1 talkers exhibit a dependence on vowel (F/partial $\eta^2 = 75.9/0.101$) whereas Group 2 (older male) talkers do not exhibit a significant dependence on vowel nor on age.

ANOVA tests were also done to study the effects of formant values (thresholds at the formant means). The only statistically significant result is for F_1 with Group 1 talkers (F/partial $\eta^2 = 91.4/0.034$). No significant correlation between $H_1^* - H_2^*$ and F_1 (vowel height) can be observed for Group 2, nor can a correlation with F_2 and F_3 be shown for any group. This effect can be seen in Figure 3.3 which depicts $H_1^* - H_2^*$ as a function of vowel for the Group 1 talkers. Vowels are sorted from left to right as a function of their average F_1 value. $H_1^* - H_2^*$ values for /iy/ and /uw/ are the lowest, suggesting that high vowels have lower OQ. As F_1 increases for /iy/, /uw/, /ih/, /eh/, and /ae/, $H_1^* - H_2^*$ becomes larger. Figure 3.4 shows $H_1^* - H_2^*$ as a function of F_1 and agrees with Figure 3.3 trends. Hanson showed in [Han97] that, for adult female voices, the mean value of $H_1^* - H_2^*$ was slightly lower for /eh/ than /ae/ which agrees with our results.

The lack of significant trends of $H_1^* - H_2^*$ values with F_1 for Group 2 talkers may be due to the physiology associated with voice production in different genders. This difference could be due to increased vocal tract-source interaction when F_0 or its harmonics are close to F_1 [Tit04], which is often the case for low

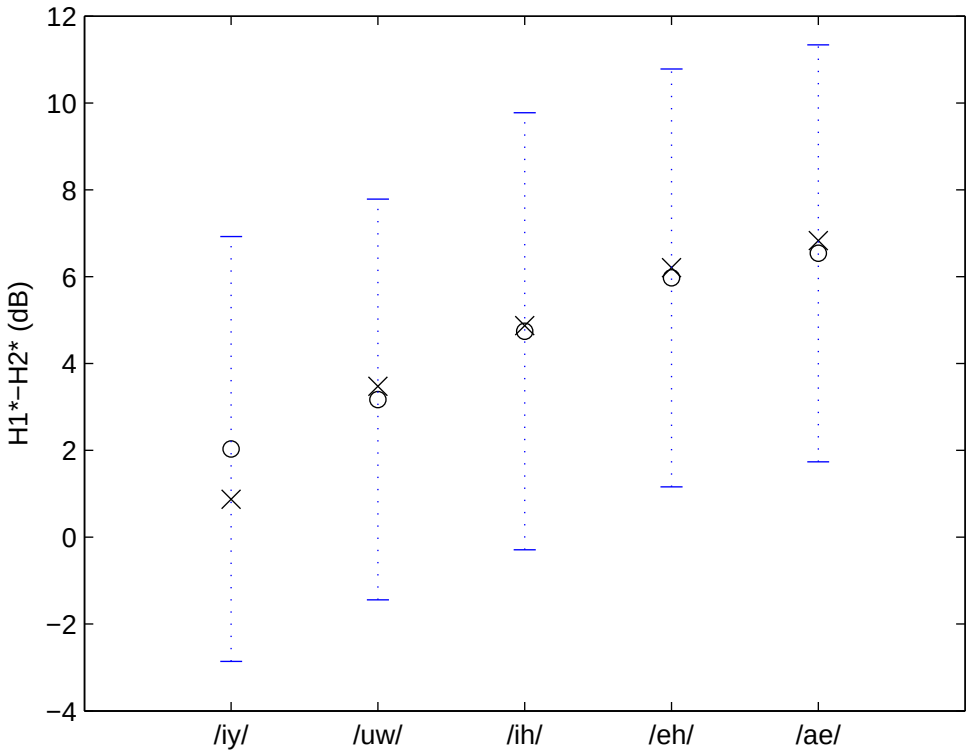

Figure 3.3: $H_1^* - H_2^*$ as a function of vowel for Group 1 talkers (females and children). Vowels are sorted according to their F_1 value from low to high. Note that the lowest values occur for the high and tense vowels /iy/ and /uw/.

F_1 and high F_0.

For both sexes $H_1^* - H_2^*$ for /iy/ is about 3 dB lower than for /ih/. This could be due to the tense/lax difference. As reported in [ML85], for four minority languages in China, the amplitude difference between the first two harmonics was smaller for tense vowels than lax ones, which would agree with our findings.

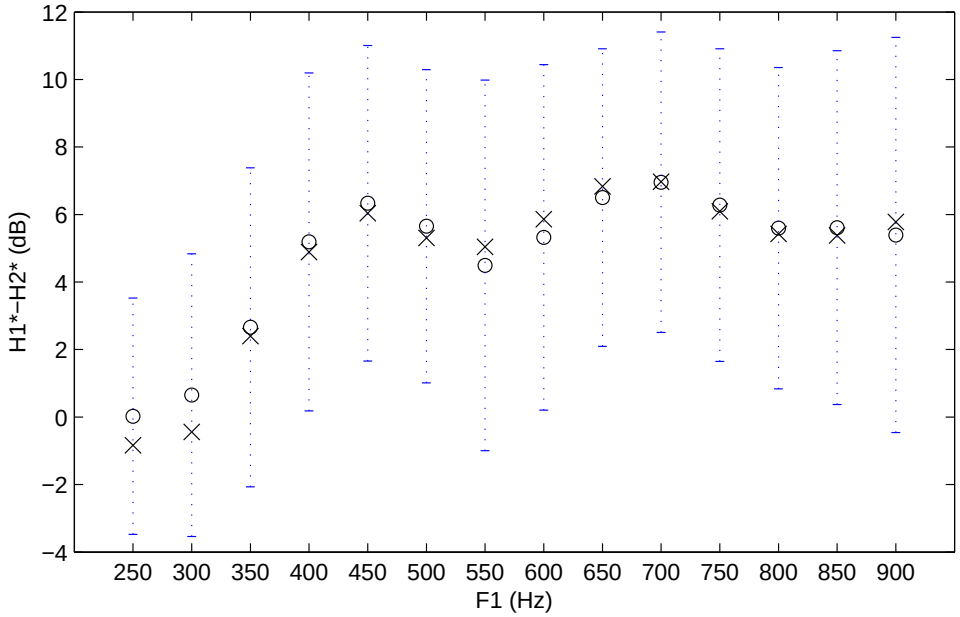

Figure 3.4: $H_1^* - H_2^*$ versus F_1 for Group 1 talkers. $H_1^* - H_2^*$ monotonically increases, on average, by about 6 dB when F_1 increases between 250–450 Hz.

3.3.4 Relationship of $H_1^* - H_2^*$ with F_0 and $H_1^* - A_3^*$

Figure 3.5 shows the relationship between $H_1^* - H_2^*$ and F_0 for both groups. As can be seen in Table 3.6, the Pearson correlation coefficient (PCC) between $H_1^* - H_2^*$ and F_0 yields a value of 0.767 for Group 2 and a weak negative correlation (PCC=-0.471) for Group 1. An approximate mapping for $H_1^* - H_2^*$ and F_0 for Group 2 is:

$$H_1^* - H_2^* \approx 0.22F_0 - 28 \quad \text{for } F_0 \text{ between 80–175 Hz} \tag{3.1}$$

A possible interpretation for this result is that the Group 1 talkers (females and children, generally high-pitched) and the Group 2 talkers (older males, generally low-pitched) use OQ differently during the phonation of vowels. In a study

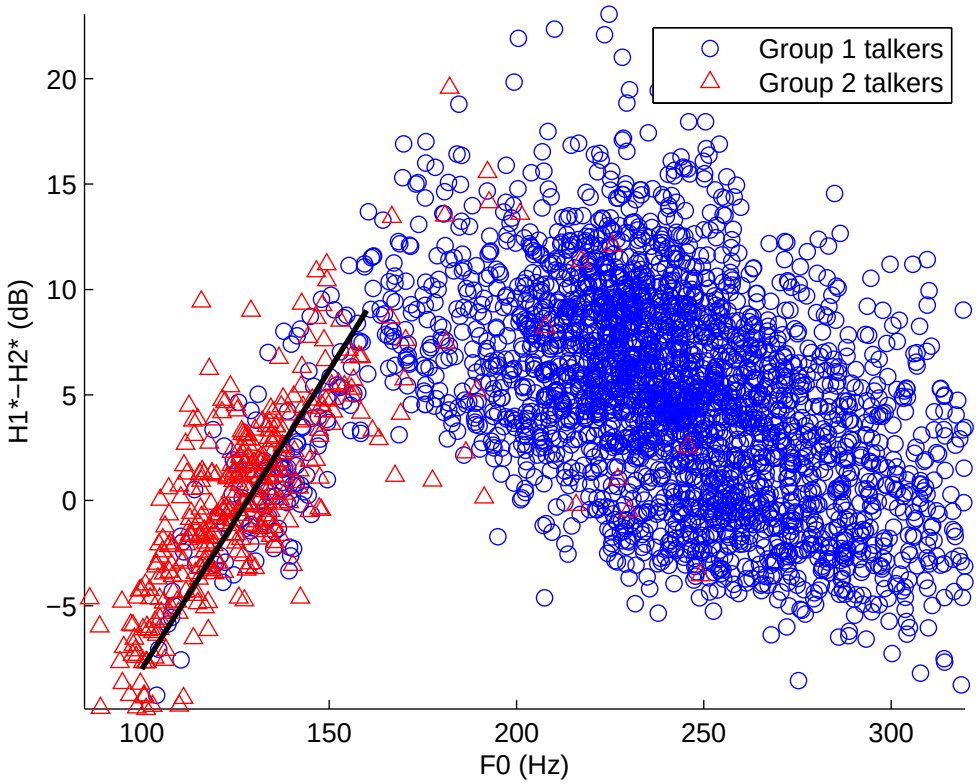

Figure 3.5: $H_1^* - H_2^*$ versus F_0 for Group 1 and Group 2 talkers. A linear relationship for F_0 between 80 and 175 Hz is observed.

by [Esp05] utilizing electroglottography (EGG) of Zapotec talkers, females were shown to produce phonation differences by altering OQ while males did not. It has also been observed in [Kor96] that increased tension of the cricothyroid muscle in the larynx induces a simultaneous increase of F_0 and OQ, and therefore also of $H_1^* - H_2^*$. However, we observed a strong positive correlation only for low F_0 values. Similar results were found in [SV01].

As seen in Table 3.6 the intercorrelation between $H_1^* - H_2^*$ and $H_1^* - A_3^*$ for both groups is weak: 0.532 (Group 1), 0.473 (Group 2). A weak correlation was

also reported in [Han97] for adult female talkers.

3.3.5 $H_1^* - A_3^*$

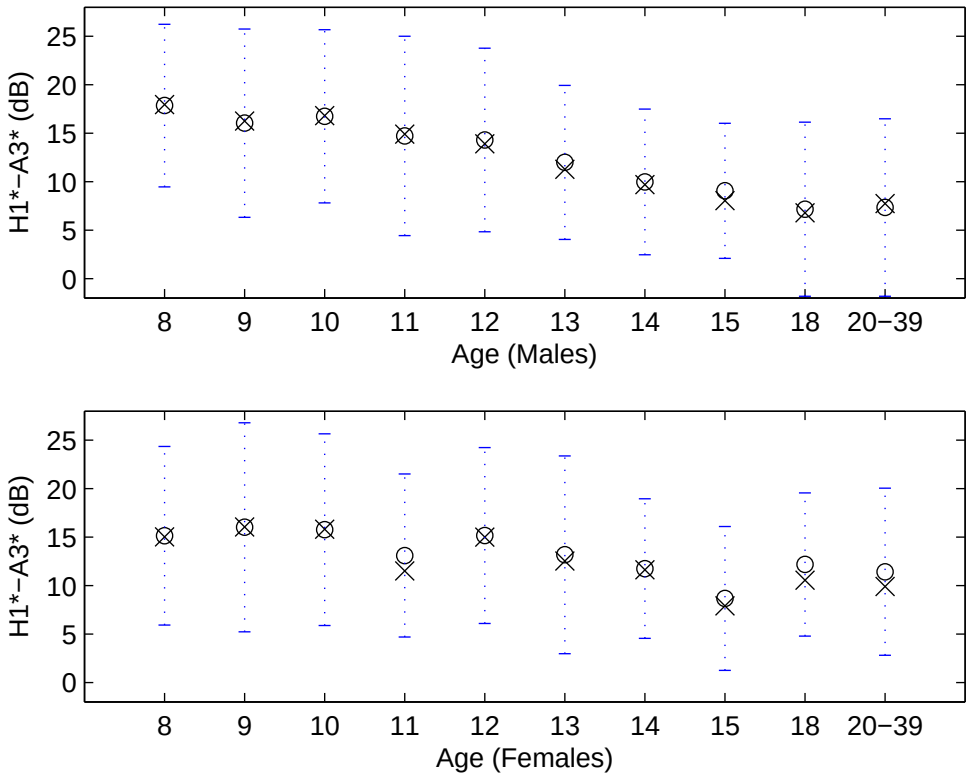

Figure 3.6: $H_1^* - A_3^*$ versus age; the top panel represents data for male talkers and the lower panel represents data for female talkers. For both sexes there is a drop of $H_1^* - A_3^*$ between age 8 and age group 20–39: The drop is about 4 dB for females, and 10 dB for males.

The age and sex effects on $H_1^* - A_3^*$ (related to source spectral tilt) are shown in Figure 3.6. Between ages 8 and 20–39, the mean $H_1^* - A_3^*$ value drops for male talkers by about 10 dB, whereas for female talkers it drops by about 4 dB

resulting in higher values (by about 4 dB) for adult females than for adult males. The higher effect size for males (F/partial $\eta^2 = 34.4/0.155$) compared to females (F/partial $\eta^2 = 9.7/0.060$) in Table 3.4 confirms this result. When the talkers are split into groups (see Table 3.5), Group 1 shows a dependence on age (F/partial $\eta^2 = 17.2/0.054$), whereas Group 2 does not. It is also interesting to note that the dependence on sex is not significant for Group 1. These trends are similar to those shown for $H_1^* - H_2^*$ (see Section 3.3.3), thus they can be interpreted similarly. That is, females (children and adults) and young males (8–14 years old) exhibit statistically similar OQ and source spectral tilt characteristics.

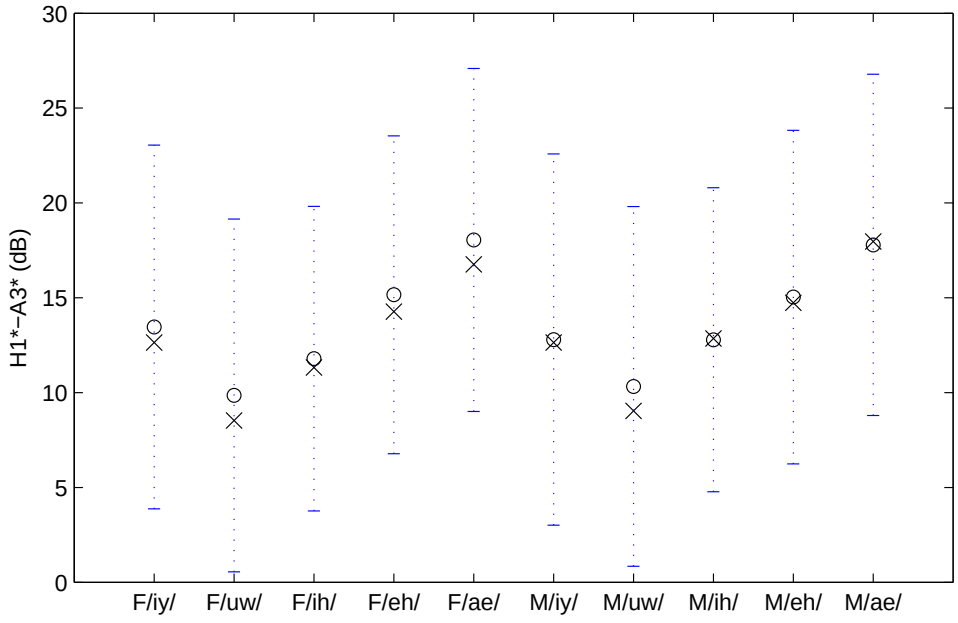

Figure 3.7: $H_1^* - A_3^*$ as a function of vowel for all talkers; M and F indicate data from male and female talkers, respectively. /ae/ and /eh/ have the highest values, while /uw/ has the lowest value. This result might be related to the dependence of the parameter on formant values.

In Figure 3.7, $H_1^* - A_3^*$ is plotted as a function of vowel and sex. The largest difference is observed between the vowels /ae/ and /uw/ where /ae/ is a low front vowel (high F_1, high F_2) and /uw/ is a high back vowel (low F_1, low F_2). Values for $H_1^* - A_3^*$ for /ae/ and /eh/ are the highest, and for /uw/ they are the lowest. These trends are similar for both sexes and indeed it can be seen from ANOVA analysis that the effect sizes of vowel are similar when male talkers are compared with females (Table 3.4).

To find the effects of formants on $H_1^* - A_3^*$, an ANOVA analysis based on high and low values of F_1, F_2 and F_3 (thresholds at the formant means) yielded F/partial η^2 values of 210/0.063, 42.7/0.013 and 100.0/0.031, respectively. Thus, the first three formants have an effect on $H_1^* - A_3^*$ for all talkers. To visualize these effects, Figures 3.8, 3.9, and 3.10 show $H_1^* - A_3^*$ gradually rising for increasing F_1, F_2, and F_3, respectively. Since /uw/ on average has lower F_2 and F_3 compared to the other vowels used in this study, this can explain why $H_1^* - A_3^*$ values for /uw/ are lowest.

The dependency of $H_1^* - A_3^*$ on F_1 is somewhat similar to the dependency of $H_1^* - A_3^*$ on $H_1^* - A_1$ (related to F_1) which was observed in [HC99]. The dependency of the measure on F_2 and F_3 was expected since a high F_2 is normally associated with a high F_3, which in term will affect the source spectral tilt. Since A_3^* represents the magnitude of the source spectrum at F_3, it is affected by the position of F_3 due to the source spectral tilt. A_3^* can also be influenced by the presence of higher formants, such as F_4, for which the parameter was not corrected for, and which would boost the value of A_3^* when evaluated close to F_4.

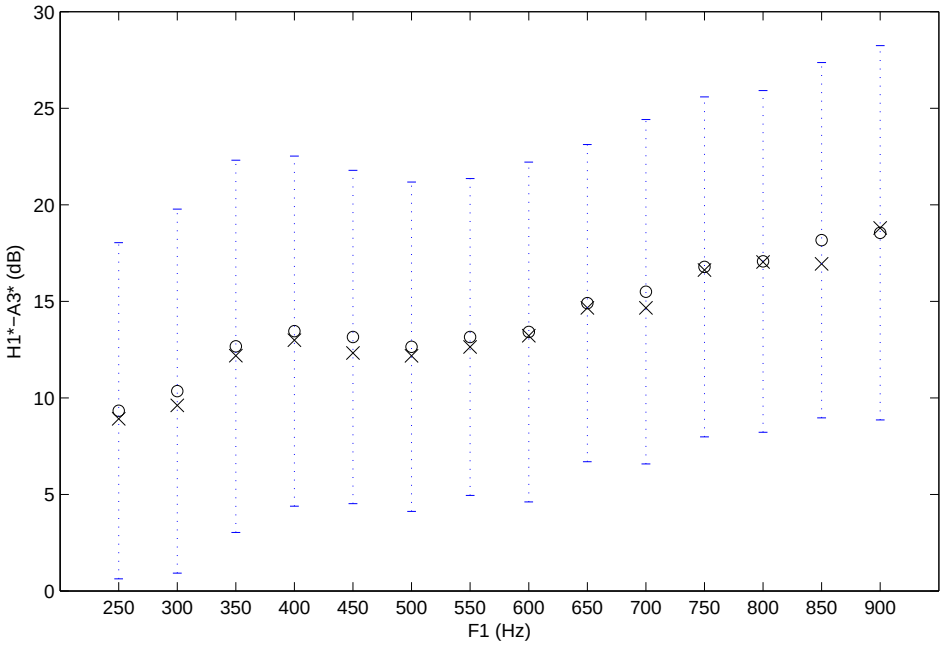

Figure 3.8: $H_1^* - A_3^*$ versus F_1 for all talkers. $H_1^* - A_3^*$ increases for increasing F_1.

3.4 Summary

In this chapter, the effects of age, sex, and vocal tract configuration on the three voice source measures F_0, $H_1^* - H_2^*$, and $H_1^* - A_3^*$ are studied, applying the correction formula presented in Chapter 2.

Statistical analysis of variance (ANOVA) is performed for all three voice source measures and the three factors, age, sex and vowel. These factors are tested with: a) all talkers, b) talkers separated by sex and c) talkers separated into Group 1 (children ages 8 to 14 and females ages 15 and older: generally high-pitched) and Group 2 (males ages 15 and older: generally low-pitched). In addition, where applicable, Pearson correlation coefficients are calculated for the different measurements. For Group 1, all effects are statistically significant except when

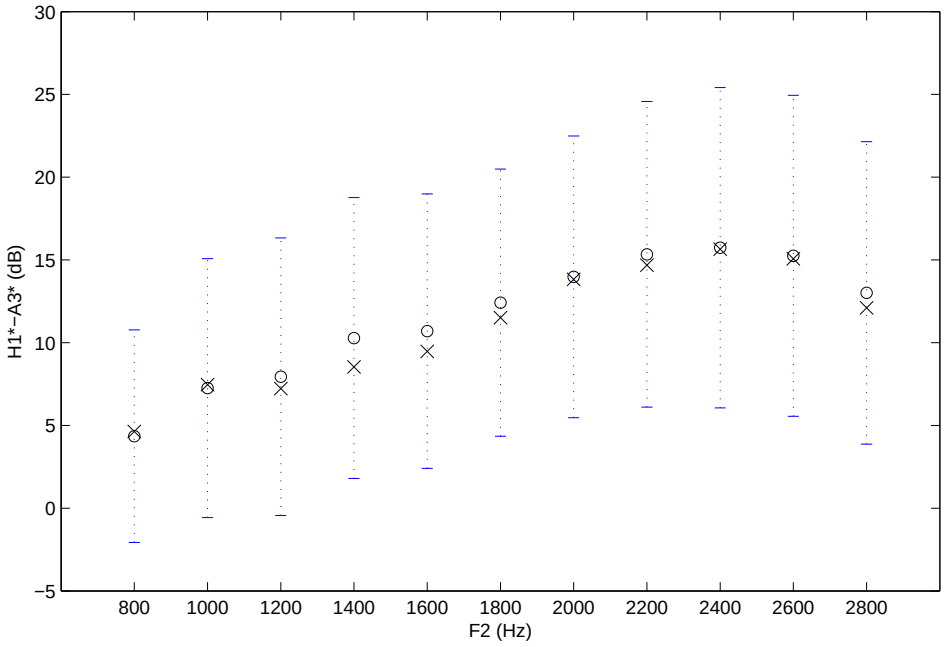

Figure 3.9: $H_1^* - A_3^*$ versus F_2 for all talkers. $H_1^* - A_3^*$ monotonically increases for F_2 increasing between 800 and 2400 Hz.

sex is tested against $H_1^* - H_2^*$ and $H_1^* - A_3^*$. This result suggests that females of all age groups and boys (ages 8 to 14) have similar OQ and source spectral tilt values. For Group 2 the only significant result occurs when $H_1^* - A_3^*$ is tested against vowel type.

F_0 for male talkers drops between ages 8 and 20-39 (by about 130 Hz), whereas the overall drop for females is only about 50 Hz. F_0 is shown to be vowel dependent, with the highest values for /uw/, and higher for /iy/ than for /eh/ and /ae/. This trend may be attributed to intrinsic pitch. Furthermore, F_3 is shown to have a statistically significant relationship with F_0 which can be explained by the dependency of F_3 on vocal tract length: A higher F_3 indicates a shorter vocal tract length which coincides usually with smaller and shorter vocal cords or a

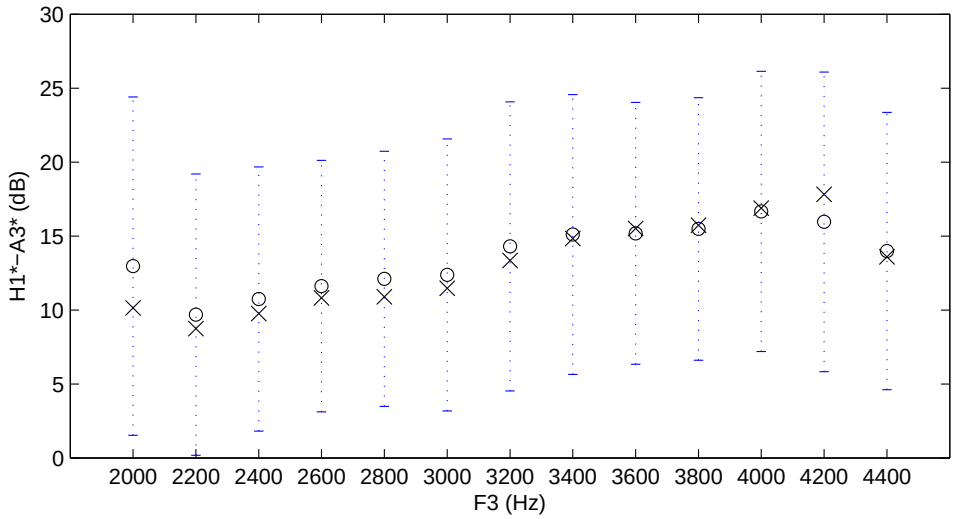

Figure 3.10: $H_1^* - A_3^*$ versus F_3 for all talkers. $H_1^* - A_3^*$ monotonically increases for F_3 increasing between 2200 and 4000 Hz.

higher F_0.

$H_1^* - H_2^*$ (hence, the open quotient) is age dependent and for male talkers a drop of about 4 dB between the ages of 9 and 20-39 is found. For females, there is less dependency on age. On average, $H_1^* - H_2^*$ values are higher by about 3 dB for adult female compared to male talkers. There is no significant dependency on age and vowel for Group 2 talkers. $H_1^* - H_2^*$ is proportional to F_0 for F_0 below 175 Hz. Above that frequency a weak negative correlation with F_0 could be found. For Group 1 talkers and for F_1 below 450 Hz, $H_1^* - H_2^*$ is proportional to F_1, resulting in low $H_1^* - H_2^*$ values for high vowels. For Group 2 talkers, on the other hand, no significant correlations between the $H_1^* - H_2^*$ values and vowel height could be observed. The different OQ dependencies between females and children (ages 8–14), and older males (ages 15 and older) could be due to phonological differences, where females alter OQ to signal acoustic differences

while males do not [Esp05], and/or to vocal tract-source interaction when F_0 or its harmonics are close to F_1 [Tit04], which is often the case for low F_1 and high F_0 values. For both sexes $H_1^* - H_2^*$ for /iy/ is about 3 dB lower than for /ih/ which could be due to a tense/lax difference.

$H_1^* - A_3^*$ (hence source spectral tilt) values drop by about 10 dB between ages 8 and 20-39 for males, whereas for females the values drop by only about 4 dB within the same age period. This results in generally lower values for adult males (by about 4 dB) compared to adult females. Until age 10, the values are similar for both sexes. Statistical analysis shows a high dependence of the measure on age and vowel for all talkers. Also, $H_1^* - A_3^*$ shows a strong dependence on all formant frequencies for all talkers and age groups: Increasing F_1, F_2, or F_3 yields an increase in $H_1^* - A_3^*$. These findings imply that source spectral tilt is vowel dependent and, in fact, it can be seen that tilt values are highest for /ae/ and /eh/ and lowest for /uw/.

	Age (from 8 to 39 years old)		Vowel dependencies and intercorrelations
	Females	Males	
F_0	$\downarrow 50\ Hz$	$\downarrow 130\ Hz$	linearly related to $H_1^* - H_2^*$ for low-pitched talkers, and to F_3 for all talkers
$H_1^* - H_2^*$	–	$\downarrow 4\ dB$	linearly related to F_0 for low-pitched talkers, and to F_1 for high-pitched talkers.
$H_1^* - A_3^*$	$\downarrow 4\ dB$	$\downarrow 10\ dB$	dependent on F_1, F_2, and F_3 for all talkers

Table 3.8: Summary of key results.

CHAPTER 4

Dependencies of voice source measures on prosodic features: A pilot study

In this chapter, we examine dependencies of the voice source measures F_0 (fundamental frequency), E_e (maximal glottal flow change), R_k (glottal symmetry/skew), LIN (spectral linearity, related to source spectral tilt), and $H_1^* - H_2^*$ (difference of formant-corrected magnitudes of the first two voice source spectral harmonics) on prosodic features such as lexical stress, pitch accent, and boundary tone. In addition to the five source measures, syllable duration (DUR) was added to compare with previous work, which found that duration was affected by prosody. A small, carefully designed corpus consisting of a sentence in different prosodic configurations was used in this study. Statistical analysis was performed using two-way ANOVAs to test for the voice source parameter dependencies.

4.1 Previous work

Accurate detection of prosodic events in speech processing applications would benefit from knowledge of voice source parameter dependencies on prosody. Pre-

vious studies of prosody have focused on F_0, syllable and word duration, intensity, high-frequency energy, and spectral balance (comparison of high- to low-frequency components) as acoustic correlates. A framework for studying voice source measures in connected speech was provided in [Fan97], in the context of the Liljencrants-Fant (LF) model parameters [FLL85]. In [FK96] it was shown that the LF parameters vary systematically as a function of both stress and pitch accent in Swedish: Increasing stress produced an increase in duration and F_0, whereas pitch accent was seen to produce increased F_0, duration, intensity, and high-frequency energy values. In [SV96] it was shown that for American English and Dutch speakers, spectral balance, duration, overall intensity, and vowel quality all varied with lexical stress (with and without pitch accent). Stressed syllables were shown to be longer and had higher spectral balance (i.e. more high-frequency energy). Spectral balance here refers to the relative spectral energy above 500 Hz compared to the total energy, and is related to the speed of glottal closure.

Recent publications [Eps02, JHC05] have used the ToBI framework, which provides labels for the following prosodic events: pitch accent, boundary tone, and break indices. In [Eps02], normalized LF model parameters were shown to vary with the presence of accents and boundary tones in a small set of short read sentences. Epstein suggested that, at least in English, prosodic strengthening is seen in voice measures in much the same way as elsewhere in speech (e.g. [Kea]). She found tenser voice (lower OQ), utterance-initially and with pitch accent, suggesting greater laryngeal tension in prosodically strong positions.

In [JHC05], the influence of pitch accents and boundary tones was evaluated for the measures: duration, F_0, harmonic structure, spectral tilt, and energy for the Boston University Radio Corpus, a relatively large database of American

English. The study did not study lexical stress. It was reported that duration and energy were useful for detecting pitch accents, while $H_1 - H_2$ was helpful for boundary detection. Note that their harmonic structure measures, such as $H_1 - H_2$, were not corrected for the influence of the formants. F_0 was higher for H-H% (high tone) boundary tones compared to L-L%. Interestingly, the time course of these measurements (slope and convexity) served as good indicators for prosodic events.

A thorough analysis, separating the effects of stress and pitch accent for the voice measures F_0, duration, $H_1^* - A_3^*$ (related to source spectral tilt), high-frequency noise (measured around F_3), and H_1^* (amplitude of voicing), was presented in [Oko06]. The study found that the effects of high tone pitch accent (H*) could be shown up to two syllables after the high tone accented syllable. Only H* was analyzed. Stress was manifested by an increase in syllable duration and high-frequency noise, and a decrease of $H_1^* - A_3^*$. H* produced an increase in F_0, syllable duration, and H_1^*, and a decrease of $H_1^* - A_3^*$. Overall, H* was shown to add to the effects of stress, e.g. high tone accent plus stress showed stronger effects on above measures than stress alone. No correlations with stress nor pitch accent could be found for $H_1^* - H_2^*$ and $H_1^* - A_1$ (related to B_1).

4.2 Data

The corpus [Eps02] consisted of the following eight-syllable sentences, where the bold word was accented:

- **Dagada** gave Bobby doodads.
- Dagada gave Bobby **doodads**.

- **Dagada** gave Bobby doodads?
- Dagada gave Bobby **doodads**?

These sentences were designed to contain no nasals and to have all vowels surrounded by voiced consonants. Lexical stress was on the following underlined syllables: "Dagada gave Bobby doodads." Boundary tones were on the syllable "dads" and their tone height depended on the declarative or interrogative nature of the sentence. Speech signals were recorded from three native talkers of Western American English between 25-35 years old: two females (F-1, F-2) and one male (M-1). Signals were collected in a sound-attenuated booth with a 1.0" Bruel & Kjaer condenser microphone placed 5 cm from the subjects' lips. The signals were sampled at 20 kHz and then downsampled to 10 kHz. Each sentence was recorded 10 times for each talker and the first and last recordings were then discarded in the final analysis.

Syllables were extracted by hand and then labeled and classified depending on the talker and prosodic features such as lexical stress (stressed vs unstressed) and pitch accent (accented vs unaccented, L* vs H*). The labeling system was based on the ToBI [SBP92] transcription standard, where each pitch accent is denoted by L and H indicating low and high pitch (F_0), respectively. A more detailed description of the data collection procedure and corpus labeling can be found in [Eps02].

Figure 4.1 shows a syllable distribution tree ordered by the prosodic events lexical stress (STR) and pitch accent (PA), where the number of analyzed syllables is shown in parentheses. Syllables are divided into stressed (STR) and unstressed (noSTR); stressed syllables are split into accented (PA) and unaccented (noPA); accented syllables are distinguished by low (L*) and high (H*) tones. A detailed syllable distribution by talker including the low (L–L%) and

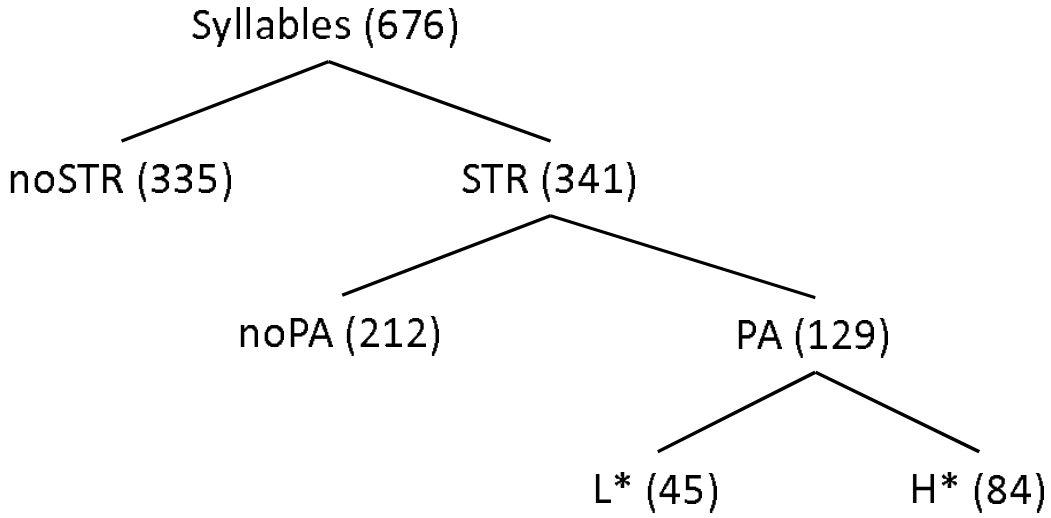

Figure 4.1: Syllable distribution tree ordered by the prosodic events lexical stress (STR) and pitch accent (PA). The number of analyzed syllables is shown in parentheses. Syllables are divided into stressed (STR) and unstressed (noSTR); stressed syllables are split into accented (PA) and unaccented (noPA); accented syllables are distinguished by low (L*) and high (H*) tones. Note that unstressed syllables are always unaccented and that pitch-accented syllables are always stressed.

high (H–H%) boundary tones for the phrase final syllable "dads" is shown in Table 4.1.

By using the same string of words with different pitch accent locations, and different pitch accent and boundary tones, the design of this database allows to directly compare the effects of these prosodic variables on voice source measures by standard factorial analysis of variance.

	Talkers			
Label	F-1	F-2	M-1	Total
L*	16	7	22	45
H*	30	33	21	84
PA	46	40	43	129
noPA	58	78	76	212
STR	104	118	119	341
noSTR	100	113	122	335
Total	204	231	241	676
L–L%	8	8	12	28
H–H%	14	16	15	45
Total	22	24	27	73

Table 4.1: Syllable distribution numbers sorted by prosodic features, with respect to each talker in the test corpus. The table includes syllable distribution numbers for the boundary syllable "dads" from the word "doodads" (L-L%, H-H%).

4.3 Methods

Our algorithms estimate the five voice source measures F_0, E_e, R_k, LIN, and $H_1^* - H_2^*$. For comparison with previous results from literature, syllable duration was calculated as well.

The voice source measures F_0, E_e, R_k, and LIN were estimated from explicit inverse filtering and LF-fitting by using the signal analysis tool developed at UCLA's Bureau of Glottal Affairs [Glo]. To prepare the data for this tool, a fundamental cycle was taken from the steady-state portion in the middle of the vowel in each syllable for each word of the corpus. The cycles were then

concatenated with themselves 10 times in order to produce a long enough signal for inverse filtering. After discarding 68 syllables, deemed to be non-LF-fittable by the inverse filtering program, there remained a total number of 676 syllables.

The amplitudes of the harmonics were estimated from the signal spectrum by using F_0 information provided by the STRAIGHT algorithm [KCP98]. The effects of the first two formant frequencies were then removed using the correction formula presented in Chapter 2 with formant frequencies and bandwidths (F_1, B_1 and F_2, B_2) obtained with the "Snack Sound Toolkit" software [Sj04]. Snack settings were: pre-emphasis factor of 0.9, analysis window length of 25 ms, and window shift of 1 ms. The small window shift time was matched to the time resolution of STRAIGHT's F_0 values; i.e. one F_0 value every millisecond.

For each syllable, the five voice source measures were estimated and since the goal was to compare parameter values on a sentence level, the voice source measures were standardized relative to each sentence's mean and standard deviation. The resulting standard scores – also called z-scores or normal scores – are dimensionless and were calculated by subtracting the sentence mean from the individual parameter values within that sentence and then by dividing the difference by the sentence standard deviation. The standardizing equation is

$$z = \frac{x - \mu}{\sigma}, \tag{4.1}$$

where z is the z-score, x is the "raw" score, μ the sentence mean, and σ the sentence standard deviation.

Manual sentence segmentation yielded syllable durations in ms. The syllable durations were not normalized, since the speaking rate and sentence length were similar for all talkers and sentences.

Syllables from different nodes in the distribution tree from Figure 4.1 were then compared using a statistical two-way ANOVA test in the software package

SPSS (v13.0). Factors for each of the two-way analyses consisted of talker plus one other factor chosen from the prosodic features.

4.4 Results

This section analyzes statistically significant dependencies of voice source measures on the prosodic features lexical stress, pitch accent, and boundary tone. The significance level was chosen as $p < 0.01$, where p is the probability of the null hypothesis. A detailed correlation evaluation of the voice source measures F_0 with E_e, $H_1^* - H_2^*$, R_k, and LIN concludes this section.

4.4.1 Lexical stress

To analyze the effect of stress on voice source measures, syllables at boundaries (i.e. "Da" and "dads") as well as syllables with pitch accent were excluded from this analysis and 181 unstressed (noSTR) and 212 stressed unaccented (noPA) syllables remained. Statistically significant dependencies of the voice source measures on stress are shown in Table 4.2. The table shows that for stressed syllables (noPA) duration increases by an average of 40 ms for our talkers. This result agrees with the findings in [Oko06]. The change in F_0 can be due to the effect of adjacent pitch-accented syllables, which was shown in [Oko06]. This effect should be avoided and a better suited set of syllables, which are not surrounded by syllables with pitch accent had to be found.

In order to reduce the effects of adjacent pitch-accented syllables in our analysis on lexical stress, we further restricted our syllable set and analyzed just the two-syllable word "Bobby". In total, we analyzed 86 unstressed unaccented (noSTR) syllables "bby" vs 80 stressed unaccented (noPA) syllables "Bo". It is

noSTR→noPA	F_0	DUR
p	0.008	0.000
	z-score means	ms
F-1	*-.091* ↗ *.085*	83 ↗ 117
F-2	.049 ↘ -.217	90 ↗ 129
M-1	.210 ↘ -.442	82 ↗ 127
Total	.060 ↘ -.170	85 ↗ 125

Table 4.2: Statistically significant dependencies of voice source measures on stress: comparing unstressed unaccented (noSTR) vs stressed unaccented (noPA) syllables. All unaccented syllables, except syllables at boundaries, were analyzed. Up arrows indicate higher values for noPA than for noSTR, down arrows mean the opposite. DUR stands for syllable duration. p is the probability of the null hypothesis.

assumed that the pitch-accented syllable "doo" in "doodads", which comes after the "bby" will not introduce any unwanted effects on the syllable "bby". The results for this restricted analysis of "Bobby" are shown in Table 4.3. All voice source measures except F_0 show a statistically significant dependency on stress. Compared with the unstressed syllable "bby", the stressed unaccented syllable "Bo" exhibits higher values of E_e and LIN and lower values of $H_1^* - H_2^*$ for all talkers. This result for $H_1^* - H_2^*$ is unlike in [Oko06], who found no correlation of $H_1^* - H_2^*$ with stress. One explanation for this could be the different vowels compared. Also, here the syllable "Bo" does not necessarily contain a full vowel "o". For female talkers, the value of R_k is lower and for the male talker it is slightly higher. Overall, this result is in line with the increase of spectral balance for Dutch stressed syllables described in [SV96]. The observed increase of E_e for stressed syllables agrees with studies that found an increase in intensity [SV96]

and H_1^* [Oko06]. The larger syllable duration - on average about 20 ms - for unaccented stressed syllables agrees with [FK96], [SV96], and [Oko06].

noSTR → noPA	E_e	$H_1^* - H_2^*$	
p	0.000	0.000	
	z-score means	z-score means	
F-1	-.587 ↗ .476	.107 ↘ -.200	
F-2	-.688 ↗ .154	.051 ↘ -.304	
M-1	-.764 ↗ -.055	.915 ↘ -.009	
Total	-.684 ↗ .160	.380 ↘ -.170	
noSTR → noPA	R_k	LIN	DUR
p	0.000	0.000	0.000
	z-score means	z-score means	ms
F-1	.687 ↘ -.971	-.789 ↗ .516	107 ↗ 126
F-2	.613 ↘ -.629	-.895 ↗ .324	117 ↗ 134
M-1	*.270 ↗ .336*	-.629 ↗ -.186	95 ↗ 120
Total	.513 ↘ -.369	-.766 ↗ .189	106 ↗ 127

Table 4.3: Statistically significant dependencies of voice source measures on stress: comparing unstressed (noSTR) vs stressed unaccented (noPA) syllables. Only the unaccented two-syllable word "Bobby" was analyzed. DUR stands for syllable duration. p is the probability of the null hypothesis.

4.4.2 Pitch accent

To find statistically significant dependencies of voice source measures on pitch accent, we were interested in the comparison of unaccented stressed (noPA) with accented stressed syllables (PA) (see tree nodes in Figure 4.1). After filtering

out all unstressed syllables and all boundary syllables, 212 unaccented vs 129 accented (stressed) syllables remained for analysis. Analysis of variance found that, compared to unaccented stressed syllables, LIN is larger for accented syllables. This would correspond to an increase in high-frequency energy. However, as will be shown later, F_0 and LIN are strongly correlated, and since pitch accent can be viewed as a combination of pitch accent tone (F_0) and stress, it is crucial to control for the influence of F_0. Therefore we compared the 45 L* vs the 84 H* syllables instead of noPA vs. PA syllables. Table 4.4 shows the dependencies of the voice source measures on pitch accent tone. Comparing the accented tones L* with H* yielded a very interesting result which, independent of talker, shows that compared to L*, H* exhibits larger values for F_0, E_e, and LIN, and smaller values for R_k. No significant dependency of $H_1^* - H_2^*$ ($p = 0.015$) and duration ($p = 0.106$) on pitch accent tone can be found.

In [Eps02] it is stated that prominent and phrase initial syllables display a tenser voice quality than non-prominent and phrase-final syllables. The 106 prominent syllables studied in [Eps02] were a subset of the 153 pitch-accented syllables studied here. Citation from [Eps02]: "Both prominent words and phrase-initial words displayed a tenser voice quality than their non-prominent and phrase-final counterparts. A tense voice quality is associated in theory with greater compression of the vocal folds and greater force of closure of the arytenoids. Acoustically, tense voice quality is correlated with low values of open quotient and glottal skew, and high values of spectral intensity and spectral linearity". Our findings are similar, however we see this behavior only for high tone accented syllables and tend to attribute this behavior more to the influence of stress which is further accentuated by the presence of a high tone.

To determine the effect of low and high tones on voice source measures, 84

L* → H*	F_0	E_e
p	0.000	0.000
	z-score means	
F-1	-1.298 ↗ .176	.155 ↗ .373
F-2	-.339 ↗ .703	-.769 ↗ -.090
M-1	-.875 ↗ .846	-.662 ↗ .712
Talkers Mean	-.942 ↗ .551	-.388 ↗ .276
L* → H*	R_k	LIN
p	0.000	0.000
	z-score means	
F-1	-.357 ↘ -.449	.061 ↗ .376
F-2	1.053 ↘ -.424	-.728 ↗ .195
M-1	.752 ↘ -.727	-.491 ↗ .523
Talkers Mean	.405 ↘ -.509	-.332 ↗ .342

Table 4.4: Statistically significant dependencies of voice source measures on pitch accent: comparing low (L*) vs high (H*) tone pitch accent. The probability (p) of the null hypothesis and the standardized means are shown.

accented high (H*) and 45 low tone (L*) syllables were separately compared to 212 unaccented stressed (noPA) syllables, which are assumed to be of average tone height. The results are presented in Table 4.5. The F value is defined as the ratio of the model mean square to the error mean square. Partial η^2 is a measure of effect size and is calculated as $SS_{effect}/(SS_{effect} + SS_{error})$, where SS_{effect} is the sum of squares of the effect and SS_{error} is the sum of squares of the error. ΔH is the parameter change for H* relative to noPA (average tone height), ΔL is the parameter change for L* relative to noPA, and ΔH-ΔL is their difference

noPA → H*	F_0	E_e	R_k	LIN	DUR
p	0.000	0.000	0.000	0.000	0.000
F	72.0	14.5	18.4	14.1	35.2
η^2	0.199	0.048	0.060	0.046	0.108
ΔH	+.721	+.363	–.501	+.410	+14 ms
noPA → L*	F_0	(E_e)	R_k	(LIN)	(DUR)
p	0.000	*0.028*	0.003	*0.056*	*0.048*
F	39.5	*4.9*	8.8	*3.7*	*3.9*
η^2	0.136	*0.019*	0.034	*0.015*	*0.015*
ΔL	–.772	*–.301*	+.413	*–.264*	*+7 ms*
L* → H*	F_0	E_e	R_k	LIN	(DUR)
p	0.000	0.000	0.000	0.000	*0.106*
F	109.4	19.4	25.9	16.5	*2.6*
η^2	0.471	0.136	0.174	0.118	*0.021*
ΔH–ΔL	+1.493	+.664	–.914	+.674	*+7 ms*

Table 4.5: Influence of low and high pitch accent on voice source measures: comparing unaccented stressed syllables (noPA) to low (L*) and high (H*) tone accented syllables separately. All values are means over all talkers. ΔH is the parameter change for H* relative to noPA (average tone height), ΔL is the parameter change for L* relative to noPA, and ΔH–ΔL is their difference representing the parameter change for H* relative to L*. The probability (p) of the null hypothesis and the standardized means are shown. For a significance level of $p < 0.01$, the following measures are statistically insignificant (in italics): E_e, LIN, and duration (DUR) for noPA→L* and DUR for L*→H*.

representing the parameter change for H* relative to L*. Unlike the results shown in Table 4.4, when comparing noPA vs L* and H* separately, some of the results

were talker dependent and therefore mean values over all talkers are presented. It can be seen that: 1) For all voice source measures in Table 4.5 the effect of high tone pitch accent is larger compared to the effect of low tone accents (η^2 is larger for noPA→H* compared to noPA→L*). Furthermore, the effect of noPA→L* on E_e, LIN, and DUR is statistically insignificant. 2) Comparing noPA→L* to noPA→H*, source measures change in opposite direction, whereas duration (DUR) changes in the same direction: e.g. for our talkers DUR increases by about +14 ms for H* and by about +7 ms for L*. 3) Independent of tone type, no significant change of $H_1^* - H_2^*$ was found with pitch accent.

4.4.3 Boundary-related tone

Significant dependencies of the voice source measures on boundary tone are evaluated here. Since for our eight-syllable sentence the boundary tone was always on the phrase-final syllable, only the unstressed unaccented final syllable "dads" in "doodads" was analyzed. As a precaution, only unaccented "doodads" were chosen which resulted in 10 low boundary tone syllables (L-L%, declarative sentences) and 22 high boundary tone syllables (H-H%, interrogative sentences).

Table 4.6 shows an expected increase of F_0 at phrase-final syllables for high boundary tones (interrogative sentences). When comparing Table 4.6 to Table 4.4 it can be seen that they are very similar: compared to low-tone syllables, measures for high-tone syllables always have higher F_0, E_e, and LIN values, and lower R_k values. Additionally the following can be found: 1) F_0 values are generally smallest for L-L% and largest for H-H% when compared to L* and H*. This can be explained with a larger F_0 excursion at boundaries 2) E_e values are smaller for boundary tones than for their corresponding accented tones, e.g. $E_{eL-L\%} < E_{eL^*}$ and $E_{eH-H\%} < E_{eH^*}$. E_e is smallest for L-L% and largest for H* which can

L-L% → H-H%	F_0	E_e
p	0.000	0.000
	z-score means	
F-1	.723 ↗ 1.155	-.716 ↗ -.635
F-2	-.746 ↗ 1.151	-1.065 ↗ -.190
M-1	-1.002 ↗ 1.014	-1.303 ↗ .444
Total	-.408 ↗ 1.102	-1.055 ↗ -.081
L-L% → H-H%	R_k	LIN
p	0.008	0.000
	z-score means	
F-1	.335 ↘ .231	-.692 ↗ -.284
F-2	.665 ↘ -.258	-.412 ↗ .388
M-1	.746 ↘ -.276	-1.178 ↗ .579
Total	.598 ↘ -.131	-.803 ↗ .274

Table 4.6: Statistically significant dependencies of voice source measures on boundary tone: comparing low (L-L%) vs high (H-H%) boundary tones. The probability of the null hypothesis (p) and the standardized means are shown. Only the phrase-final boundary syllable "dads" in the unaccented word "doodads" was analyzed.

be attributed to phrase-final effects [Sli06]. 3) Contrary to the trends for E_e, R_k values are larger for boundary tones than for their corresponding accented tones, e.g. $E_{eL-L\%} > R_{kL^*}$ and $R_{kH-H\%} > R_{kH^*}$. R_k is smallest for H*. 4) No significant dependencies of $H_1^* - H_2^*$ ($p = 0.656$) and duration ($p = 0.984$) on the type of boundary tone can be shown.

4.5 Summary

When analyzing the dependencies of voice source measures on only lexical stress, special care was taken to analyze only unaccented syllables which did not follow an accented syllable. The two-syllable word "Bo-bby" with lexical stress on the syllable "Bo", was chosen for analysis. Results showed that:

- Compared to unstressed syllables, stressed syllables yielded longer syllable durations, higher values for E_e and LIN, and lower values for $H_1^* - H_2^*$ and R_k, which would indicate a louder and tenser voice quality.
- No significant changes of F_0 were found.

The analysis of the dependencies of voice source measures on pitch accent showed the importance of controlling for F_0. Understanding pitch accent as a combination of its tone (F_0) and stress, it was important to see what effect low or high pitch accent tones (L*, H*) had on voice measures. Results show that:

- When compared to L*, independent of talker, H* provoked higher values of F_0, E_e, and LIN, and lower values of R_k. This indicates higher intensity and tenser voice quality for H* when compared to L*.
- Compared to noPA (stressed, average tone height syllable), the effect of adding H* was larger than the effect of adding L*. Furthermore, the effect of adding L* to noPA on E_e, LIN, and duration was statistically insignificant.
- Voice measures changed in opposite direction when adding H* to noPA compared to adding L*, whereas duration increased for both L* and H*.
- No significant change of $H_1^* - H_2^*$ was found with pitch accent tone type.

The results for the boundary tones L-L% and H-H%, were very similar compared to the results for pitch accent tones L* and H*, respectively:

- When compared to low tone syllables (L*, L-L%), high tone syllables (H*, H-H%) had higher values of F_0, E_e, and LIN, and lower values for R_k. Furthermore it was found that:

- F_0 values are generally smallest for L-L% and largest for H-H% when compared to L* and H*, i.e. $F_{0L-L\%} < F_{0L*orH*} < F_{0H-H\%}$. This can be explained with a larger F_0 excursion at boundaries.

- When comparing L-L% with L* and H-H% with H*, E_e values are smaller for boundary tones than for their corresponding accented tones, e.g. $E_{eL-L\%} < E_{eL^*}$ and $E_{eH-H\%} < E_{eH^*}$. E_e is smallest for L-L% and largest for H*.

- Contrary to the trends for E_e, R_k values are larger for boundary tones than for their corresponding accented tones, e.g. $R_{kL-L\%} > R_{kL^*}$ and $R_{kH-H\%} > R_{kH^*}$. R_k is smallest for H*.

- No significant change of $H_1^* - H_2^*$ and duration was found with the type of boundary tone.

These results indicate that stress and an increase in tone (F_0) both yield an increase in loudness/intensity and in high-frequency components, which could be attributed to a tenser voice. On the other hand, unstressed or low tone syllables tend to have lower intensity and less high-frequency components. Regardless of pitch accent, stressed syllables have lower $H_1^* - H_2^*$ than unstressed syllables. Furthermore, the effect of pitch accent is stronger for H* than for L* and the duration of pitch accented syllables is longer compared to unaccented ones. If pitch accent is understood as a combination of stress and tone, then the effect on

voice source measures would be a combination of the effects of stress and tone. The main results of this chapter are summarized in Table 4.7. The effects of stress were studied on syllables which were not adjacent to pitch accented syllables and which were not at boundaries, the effects of pitch accent were determined for syllables which were not at boundaries, and the effects of boundary tones were studied for syllables which were not adjacent to pitch accented syllables.

	Stress[a]	Pitch accent[b]			Boundary tone[c]
	noSTR→noPA	noPA→L*	noPA→H*	L*→H*	L-L%→H-H%
F_0	–	↘	↗	↗	↗
E_e	↗	(↘)	↗	↗	↗
$H_1^* - H_2^*$	↘	–	–	–	–
R_k	↘[d]	↗	↘	↘	↘
LIN	↗	(↘)	↗	↗	↗
DUR	↗	(↗)	↗	–	–

Table 4.7: Summary table: dependencies of voice source measures and syllable duration on stress, pitch accent tone, and boundary tone. DUR stands for syllable duration. Dashes (–) indicate that there is no significant dependency. If not stated otherwise, results are consistent for all talkers. [a]Determined for the two-syllable word "Bobby". [b]Determined for non-boundary syllables. Results for noPA→L* and noPA→H* are averaged over all talkers. [c]Determined for the syllable "dads" in unaccented "doodads". [d]Except talker M-1.

CHAPTER 5

Summary and future work

5.1 Summary

In this dissertation, the dependencies of the voice source signal on age, sex, vowel context, and prosody are presented. The focus is on the voice source measures F_0, E_e, R_k, $H_1^* - H_2^*$, $H_1^* - A_3^*$, and LIN.

Chapter 1 explains the various scientific terms used in speech processing and shows how human speech is produced using the lungs as a source of air pressure, the vocal folds as a source of excitation for voiced sounds, and the vocal tract as a resonance body to form different sounds.

In Chapter 2, a formula to remove the influence of the vocal tract resonances on the voiced speech signal is derived. The formula is based on the linear source filter model of speech production, which assumes that the human speech production system can be modeled by a source, the voice source signal produced at the glottis, and a linear filter, the vocal tract. The vocal tract is modeled with an all-pole filter, with the poles representing the formant frequencies. The formula provides magnitudes of the source power spectrum at select frequencies and is evaluated at the frequencies F_0 (to find H_1^*), $2F_0$ (to find H_2^*), and F_3 (to find A_3^*). Compared to explicit inverse filtering, which calculates the actual voice source

signal in the time domain, this method can be partly automated and needs fewer manual corrections. As a result, more data can be evaluated for use in statistical data analysis.

Chapter 3 evaluates the dependencies of the three voice source measures F_0, $H_1^* - H_2^*$ and $H_1^* - A_3^*$ on age, sex, and vowel context. The correction formula presented in Chapter 2 is applied to calculate $H_1^* - H_2^*$ and $H_1^* - A_3^*$ from a large speech database containing voice samples from American English talkers of different ages and gender, spoken in different vowel contexts.

In Chapter 4, a pilot study assesses the dependencies of five voice source measures F_0, E_e, R_k, $H_1^* - H_2^*$, and LIN on three prosodic events: lexical stress, pitch accent, and boundary tone. The measures E_e and R_k, found by explicit inverse filtering, were added. The pilot study analyzes voice source measures using one sentence spoken by one male and two female talkers of American English and pronounced with different prosodic events.

The following two sections summarize the results of the dependencies of voice source measures on age, sex, vowel context, and prosodic features.

5.1.1 Dependencies on age, sex, and vowel context

Our study shows that for male talkers, F_0 drops by about 130 Hz between ages 8 and 20-39, whereas the overall drop for females is only about 50 Hz. The source measures depended on the value of F_0, and therefore talkers are split into low pitched (males ages 15 and older) and high pitched (children ages 8 to 14 and females ages 15 and older) groups. F_0 is shown to be vowel dependent which may be attributed to intrinsic pitch. Furthermore, F_3 is shown to have a statistically significant relationship with F_0 which can be explained by the dependency of F_3 on vocal tract length: A higher F_3 indicates a shorter vocal tract length which

coincides usually with smaller/shorter vocal cords or a higher F_0.

$H_1^* - H_2^*$, related to OQ, drops by about 4 dB between the ages of 9 and 20-39 for male talkers, whereas for females no strong age dependency is shown. As a result, average $H_1^* - H_2^*$ values are about 3 dB lower for adult males when compared to adult females. $H_1^* - H_2^*$ is proportional to F_0 for $F_0 < 175$ Hz. For high-pitched talkers and for F_1 below 450 Hz, $H_1^* - H_2^*$ is proportional to F_1, resulting in low $H_1^* - H_2^*$ values for high vowels. For low-pitched talkers, no significant dependency of $H_1^* - H_2^*$ on age and vowel is shown. This could be due to phonological differences, where females alter OQ to signal acoustic differences while males do not, and/or to vocal tract-source interaction when F_0 or its harmonics are close to F_1. For both sexes $H_1^* - H_2^*$ for /iy/ is about 3 dB lower than for /ih/ which could be due to a tense/lax difference.

$H_1^* - A_3^*$, related to source spectral tilt, drops by about 10 dB between ages 8 and 20-39 for males, whereas for females the values drop by only about 4 dB within the same age period. This results in generally lower values for adult males (by about 4 dB) compared to adult females. Until age 10, the values are similar for both sexes. Statistical analysis shows a high dependence of the measure on age and vowel for all talkers. Also, $H_1^* - A_3^*$ shows a strong dependence on all formant frequencies for all talkers and age groups: Increasing F_1, F_2, or F_3 yields an increase in $H_1^* - A_3^*$. These findings imply that source spectral tilt is vowel dependent and, in fact, it can be seen that tilt values are highest for /ae/ and /eh/ and lowest for /uw/.

5.1.2 Dependencies on prosodic features

For the analysis of voice source dependencies on lexical stress, voice source measures estimated from unaccented stressed syllables were compared to unaccented

unstressed syllables. In addition, the two-syllable word "Bo-bby" with lexical stress on the syllable "Bo" was also analyzed. The interesting thing about "Bobby" is that the preceding pitch-accented syllables (e.g. "ga" from "Dagada") occurred at least three syllables before "Bobby", thus minimizing any possible influence pitch accent might have [Oko06]. Results show that, compared to unstressed syllables, unaccented stressed syllables have higher values for E_e, LIN, and duration, and lower values for $H_1^* - H_2^*$ and R_k, which would indicate a louder and tenser voice quality. No significant changes of F_0 are found.

The analysis of the dependencies of voice source measures on pitch accent shows the importance of distinguishing low and high pitch-accented tones, L* and H*. Understanding pitch accent as a combination of lexical stress and tone height, special attention is given to the analysis of changes in tone height, since the effects of lexical stress were already examined: The dependencies of voice source measures on L*, H*, and stressed unaccented syllables (noPA) of average tone height are analyzed. Compared to noPA, H* provokes higher values of F_0, E_e, and LIN, and lower values of R_k, independent of talker. This indicates higher intensity and tenser voice quality for H*. Compared to noPA, L* causes the measures to change in the opposite direction, however only the changes for F_0 and R_k are statistically significant. Duration is found to be longer for H* compared to noPA. No significant change of $H_1^* - H_2^*$ is found.

The results for the boundary tones L-L% and H-H%, are very similar compared to the results for pitch-accented tones L* and H*: When compared to low tone syllables (L*, L-L%), high tone syllables (H*, H-H%) have higher values of F_0, E_e, and LIN, and lower values for R_k. F_0 values are generally smallest for L-L% and largest for H-H% when compared to L* and H*, i.e. $F_{0L-L\%} < F_{0L*orH*} < F_{0H-H\%}$. This can be explained with a larger F_0 excursion

at boundaries. When comparing L-L% with L* and H-H% with H*, E_e values are smaller for boundary tones than for their corresponding accented tones, e.g. $E_{eL-L\%} < E_{eL^*}$ and $E_{eH-H\%} < E_{eH^*}$. E_e is smallest for L-L% and largest for H*. Contrary to the trends for E_e, R_k values are larger for boundary tones than for their corresponding accented tones, e.g. $R_{kL-L\%} > R_{kL^*}$ and $R_{kH-H\%} > R_{kH^*}$. R_k is smallest for H*.

These results indicate that lexical stress and an increase in tone (F_0) both yield an increase in loudness/intensity and in high-frequency components, which could be attributed to a tenser voice. Furthermore, pitch-accented syllables are longer than non-accented ones independent of the pitch tone being H* or L* and that stressed syllables have lower $H_1^* - H_2^*$ than unstressed syllables regardless of pitch accent. On the other hand, unstressed and low pitch accent tone and boundary tone syllables tend to have lower intensity and less high-frequency components. If pitch accent is understood as a combination of lexical stress and tone, then the effect on voice source measures could be a combination of the effects of lexical stress and tone.

5.2 Challenges and Outlook

This dissertation developed an approach for the extraction of voice source measures and unraveled the dependencies on the factors age, sex, vowel context, and prosodic features for six voice source measures. It was seen that certain voice source measures as well as the factors age, sex, and vowel are intercorrelated and that splitting the data into low- and high-pitched talkers helped reduce these intercorrelations. In a pilot study, the effects of prosodic events, such as stress, pitch accent, and boundary tone, on voice source measures were presented. It was found that a clear separation of syllable context was necessary to evaluate

the dependencies of the voice source measures on prosody: For example, when analyzing stressed syllables, it was made sure that pitch accented syllables were not adjacent. Since the acoustic correlates of pitch accent do not necessarily occur in the middle of syllables, it is recommended that contours are used or that measurements are taken around F_0 minima and maxima when evaluating the influence of pitch accent.

Future approaches to voice source analysis should try to quantify the intercorrelations of factors and of source measures. In order to obtain more independent results subset of the speech data could be analyzed. It would be interesting to see how new and different voice source measures depend on the factors presented in this dissertation or other factors such as voice pathologies. A better knowledge of voice source correlations with voice quality (breathy, creaky, tense, etc.) and prosodic features (lexical stress, boundaries, etc.) would benefit speaking style detection, emotion classification, and other higher-level information in speech. Additionally, on a lower level, the speech source measures themselves, would provide a better understanding of the mechanism of speech production and the voice source signal. All in all, a better understanding of the speech production mechanism and the (inter)correlations of voice source measures can help improve practical applications such as speaker identification and recognition, speech recognition, speech analysis and synthesis, speech coding, and speech enhancement.

References

[Ana84] T. V. Ananthapadmanabha. "Acoustic analysis of voice source dynamics." *STL-QPSR*, **25**(2–3):1–24, 1984.

[Bak87] R. J. Baken. *Clinical Measurement of Speech and Voice*. Taylor and Francis Ltd, London, 1987.

[BE97] M. E. Beckman and G. A. Elam. "Guidelines for ToBI labeling." Ohio State University, 1997. http://www.ling.ohio-state.edu/research/phonetics/E_ToBI/ (last viewed Jan. 2007).

[Ber02] A. Bernard. *Source and channel coding for speech transmission and remote speech recognition*. Ph. d. thesis, University of California, Los Angeles, 2002.

[CC95] K. E. Cummings and M. A. Clements. "Glottal models for digital speech processing: a historical survey and new results." *Digital Signal Processing*, **5**:21–42, 1995.

[CG97] A. N. Chasaide and C. Gobl. *The handbook of phonetic sciences*, chapter Voice Source Variation, pp. 428–461. Blackwell Publishers Inc., 1997.

[Chi94] D. G. Childers. "Measuring and modeling vocal source-tract interaction." *IEEE Transactions on Biomedical Engineering*, **41**(7):663–671, July 1994.

[CM95] G. Cohen and D. Malah. "Speech analysis and synthesis using a glottal excited AR model with DTW-based glottal determination." In *18th Convention of Electrical and Electronics Engineers*, Tel Aviv, Israel, March 1995.

[Dd99] B. Doval and C. d'Alessandro. "The spectrum of glottal flow models." Technical report, LIMSI-CNRS, Orsay, France, May 1999.

[DdH03] B. Doval, C. d'Alessandro, and N. Henrich. "The voice source as a causal/anticausal linear filter." In *Proceedings of VOQUAL'03*, pp. 15–20, Geneva, Switzerland, August 2003. ISCA.

[EM91] A. El-Jaroudi and J. Makhoul. "Discrete all-pole modeling." *IEEE Trans. Sig. Proc.*, **39**(2):411–423, 1991.

[Eps02] M. Epstein. *Voice Quality and Prosody in English.* Dissertation, University of California, Los Angeles, 2002.

[Esp05] C. Esposito. "An Acoustic and Electroglottographic Study of Phonation in Santa Ana del Valle Zapotec." Poster at the 79th meeting of the Linguistic Society of America, 2005, 2005.

[Fan60] G. Fant. *Acoustic theory of speech production.* Mouton, The Hague, Paris, 1960.

[Fan82] G. Fant. "The voice source - Acoustic modeling." *STL-QPSR*, **23**(4):28–48, 1982.

[Fan95] G. Fant. "The LF model revisited. Transformations and frequency domain analysis." *STL-QPSR*, **36**(2–3):119–156, 1995.

[Fan97] G. Fant. "The voice source in connected speech." *Speech Communication*, pp. 125–139, 1997.

[FI78] J. L. Flanagan and K. Ishizaka. "Computer model to characterize the air volume displaced by the vibrating vocal cords." *J. Acoust. Soc. Am.*, **63**(5):1559–1565, May 1978.

[FK96] G. Fant and A. Kruckenberg. "Voice source properties of speech code." *TMH-QPSR*, **37**(4):45–56, 1996.

[FKL00] G. Fant, A. Kruckenberg, J. Liljencrants, and S. Hertegård. "Acoustic-phonetic studies of prominence in Swedish." *TMH-QPSR*, **41**(2–3):1–52, 2000.

[FL85] G. Fant and Q. Lin. "Glottal source - vocal tract acoustic interaction." *STL-QPSR*, **28**(1):13–27, 1985.

[Fla72] J. Flanagan. *Analysis, synthesis, and perception of speech.* Springer-Verlag, Berlin, 2nd edition, 1972.

[FLL85] G. Fant, J. Liljencrants, and Q. Lin. "A four-parameter model of glottal flow." *STL-QPSR*, **26**(4):1–13, 1985.

[FMS01] M. Fröhlich, D. Michaelis, and H. W. Strube. "SIM - simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals." *J. Acoust. Soc. Am.*, **110**(1):479–488, July 2001.

[GC99] C. Gobl and A. N. Chasaide. "Perceptual correlates of source parameters in breathy voice." In *Proceedings of the XIVth International Congress of Phonetic Sciences*, pp. 2437–2440, San Francisco, 1999.

[GC02] C. Gobl and A. N. Chasaide. *Improvements in Speech Synthesis*, chapter Dynamics of the glottal source signal: implications for naturalness in speech synthesis, pp. 273–283. Wiley and Sons, New York, 2002.

[Glo] The Bureau of Glottal Affairs. "Inverse Filter Software." UCLA. Available as open source shareware at http://www.surgery.medsch.ucla.edu/glottalaffairs/software.htm.

[Han95] H. M. Hanson. *Glottal characteristics of female speakers.* Ph. D. Dissertation, Harvard University, Cambridge, MA, 1995.

[Han97] H. M. Hanson. "Glottal characteristics of female speakers: Acoustic correlates." *J. Acoust. Soc. Am.*, **101**:466–481, 1997.

[HC99] H. M. Hanson and E. S. Chuang. "Glottal characteristics of male speakers: Acoustic correlates and comparison with female data." *J. Acoust. Soc. Am.*, **106**:1064–1077, 1999.

[HdD01] N. Henrich, C. d'Alessandro, and B. Doval. "Spectral correlates of voice open quotient and glottal flow asymmetry: theory, limits and experimental data." In *Proceedings of EUROSPEECH*, pp. 47–50, Scandinavia, 2001.

[Hed84] P. Hedelin. "A glottal LPC-vocoder." In *Proc. IEEE*, pp. 1.6.1–1.6.4, 1984.

[HG92] S. Hertegård and J. Gauffin. "Acoustic properties of the Rothenberg mask." *STL-QPSR*, **33**(2–3):9–18, 1992.

[HHP95] E. B. Holmberg, R. E. Hillman, J. S. Perkell, P. Guiod, and S. L. Goldman. "Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice." *J. Speech Hear. Res.*, **38**:1212–1223, 1995.

[Hol73] J. N. Holmes. "Influence of the glottal waveform on the naturalness of speech from a parallell formant synthesizer." *IEEE Trans. on Audio and Electroacoustics*, pp. 298–305, 1973.

[IA04] M. Iseli and A. Alwan. "An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation." In *Proceedings of ICASSP*, volume 1, pp. 669–672, Montreal, Canada, May 2004.

[JHC05] J-Y.Choi, M. Hasegawa-Johnson, and J. Cole. "Finding intonational boundaries using acoustic cues related to the voice source." *The Journal of the Acoustical Society of America*, **118**(4):2579–2587, 2005.

[KCP98] H. Kawahara, A. de Cheveign, and R. D. Patterson. "An instantaneous-frequency-based pitch extraction method for high quality speech transformation: revised TEMPO in the STRAIGHT-suite." In *Proceedings ICSLP'98*, Sydney, Australia, December 1998.

[Kea] P. Keating. "Phonetic Encoding of Prosodic Structure." to appear in J. Harrington and M. Tabain (eds) *Speech Production: Models, Phonetic*

Processes and Techniques, Psychology Press: New York.

[KK90] D. H. Klatt and L. C. Klatt. "Analysis, synthesis, and perception of voice quality variations among female and male talkers." *J. Acoust. Soc. Am.*, **87**(2):820–857, February 1990.

[Kor96] J. Koreman. *Decoding Linguistic Information in the Glottal Airflow.* Ph.d thesis, University of Nijmegen, 1996.

[LP61] I. Lehiste and G. E. Peterson. "Some basic considerations in the analysis of intonation." *J. Acoust. Soc. Am.*, **33**(4):419–425, April 1961.

[LPN99] S. Lee, A. Potamianos, and S. Narayanan. "Acoustics of childrens speech: Developmental changes of temporal and spectral parameters." *J. Acoust. Soc. Am.*, **105**(3):1455–1468, March 1999.

[LS99] H.-L. Lu and J. O. Smith, III. "Joint estimation of vocal tract filter and glottal source waveform via convex optimization." In *Proceedings of the IEEE workshop on applications of signal processing to audio and acoustics*, pp. 79–82, Piscataway, NJ, USA, 1999.

[Man98] R. H. Mannell. "Formant diphone parameter extraction utilising a labelled single speaker database." In *Proceedings of the ICSLP*, volume 5, pp. 2003–2006, Sydney, Australia, 1998. ASSTA.

[Mar65] J. Mártony. "Studies of the voice source." *STL-QPSR*, **6**(1):4–9, 1965.

[Mar96] K. Marasek. "Glottal correlates of the word stress and the tense-lax opposition in German." In *Proceedings ICSLP*, pp. 1573–1576, Philadelphia, PA, Oct. 1996.

[Mil59] R. L. Miller. "Nature of the vocal cord wave." *J. Acoust. Soc. Am.*, **31**:667–677, 1959.

[ML85] I. Maddieson and P. Ladefoged. "Tense and lax in four minority languages of China." *Journal of Phonetics*, **13**:433–454, 1985.

[MLU96] J.D. Miller, S. Lee, R.M. Uchanski, A.F. Heidbreder, B.B. Richman, and J. Tadlock. "Creation of two children's speech databases." In *Proceedings of ICASSP*, volume 2, pp. 849–852, May 1996.

[Oko06] A. O. Okobi. *Acoustic correlates of word stress in American English.* Dissertation, Massachusetts Institute of Technology, 2006.

[PB52] G. E. Peterson and H. L. Barney. "Control methods used in a study of the vowels." *J. Acoust. Soc. Am.*, **24**(2):175–184, March 1952.

[PQR99] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds. "Modeling of the glottal flow derivative waveform with application to speaker identification." *IEEE Trans. Speech Audio Processing*, **7**(5):569–585, September 1999.

[RJ93] L. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition.* Prentice Hall, Englewood, New Jersey, 1993.

[Ros71] A. E. Rosenberg. "Effect of glottal pulse shape on the quality of natural vowels." *J. Acoust. Soc. Am.*, **49**(2):583–590, February 1971.

[Rot73] M. Rothenberg. "A new inverse-filtering technique for deriving the glottal airflow during voicing." *J. Acoust. Soc. Am.*, **53**(6):1632–1645, 1973.

[RS78] L. R. Rabiner and R. W. Schafer. *Digital processing of speech signals.* Prentice Hall, Englewood Cliffs, NJ, 1978.

[SBP92] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, and J. Pierrehumbert. "ToBI: a standard for labeling English prosody." In *Proc. ICSLP*, volume 2, pp. 867–870, Banff, Alberta, Canada, Oct. 1992.

[Sj04] Kåre Sjölander. "Snack Sound Toolkit." KTH Stockholm, Sweden, 2004. http://www.speech.kth.se/snack/ (last viewed Jan. 2007).

[Sli06] J. Slifka. "Some physiological correlates to regular and irregular phonation at the end of an utterance." *J. Voice*, **20**(2):171–186, June 2006.

[SPC97] X. Q. Sun, F. Plante, B. M. G. Cheetham, and W. T. K. Wong. "Phase modelling of speech excitation for low bit-rate sinusoidal transform coding." In *Proceedings ICASSP97*, volume 3, pp. 1691–1694, Munich, Germany, April 1997.

[SV96] A. Sluijter and V. Van Heuven. "Spectral balance as an acoustic correlate of linguistic stress." *J. Acoust. Soc. Am.*, **100**(4):2471–2485, 1996.

[SV01] M. Swerts and R. Veldhuis. "The effect of speech melody on voice quality." *Speech Communication*, **33**:297–303, 2001.

[SVP97] A. Sluijter, V. Van Heuven, and J. Pacilly. "Spectral balance as a cue in the perception of linguistic stress." *J. Acoust. Soc. Am.*, **101**(1):503–513, 1997.

[TBA04] TBALL Project. "Technology Based Assessment of Language and Literacy.", 2004. http://diana.icsl.ucla.edu/Tball/.

[Tit04] I. R. Titze. "A Theoretical Study of F0-F1 Interaction with Application to Resonant Speaking and Singing Voice." *Journal of Voice*, **18**(3):292–298, 2004.

[Vel98] R. Veldhuis. "A computationally efficient alternative for the Liljencrants-Fant model and its perceptual evaluation." *J. Acoust. Soc. Am.*, **103**(1):566–571, January 1998.

[Wak77] H. Wakita. "Normalization of vowels by vocal-tract length and its application to vowel identification." *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **25**(2):183–192, April 1977.

[YFL99] A. Yasmin, P. Fieguth, and D. Li. "Speech enhancement using voice source models." In *Proceedings ICASSP'99*, volume 2, pp. 797–800, Phoenix, AZ, March 1999.