

HUMAN AND MACHINE RECOGNITION OF SPEECH SOUNDS IN NOISE

Abeer Alwan, Qifeng Zhu, and Jeff Lo

Department of Electrical Engineering, UCLA
Los Angeles, CA 90095
* NEC Electronics, Santa Clara, CA

ABSTRACT

In this paper, we report on studies of human perception and machine recognition of speech signals in noise. As a case study, the perception of the linguistic feature ‘place of articulation’ for consonants in adverse conditions is examined through a series of perceptual experiments. The experiments examined the effects of additive noise on the perception of Consonant-Vowel (CV) syllables. Results are analyzed in terms of vowel context and SNR. Results show a striking vowel-context effect. In general, the /Ca/ context is more noise robust than other contexts. Another important result is that certain attributes which cue a feature acoustically and perceptually in quiet conditions (such as nasal murmur for /n, m/) do not correlate well with the perceptual robustness of the feature in noise. A Hidden Markov Model (HMM)-based automatic speech recognition (ASR) system was then constructed to identify the features at various signal-to-noise ratios. Modifications to a standard ASR system were made that were inspired by the results of the perceptual experiments. The modifications improved recognition performance by 20-40 percent in noise.

Keywords: noise-robust speech recognition; human perception in noise; vowel-context effects.

1. INTRODUCTION

Although noise is frequently the limiting factor in communication, most previous studies that examined the perceptual importance of acoustic cues in signaling phonetic contrasts have been based on experiments conducted in quiet. This study focuses on the perception of the place of articulation for syllable-initial nasal consonants /m, n/ in adverse conditions. These sounds are typically characterized by an initial segment (a murmur), which has most of its energy in the low-frequency region, and by distinct formant transitions into the neighboring vowel. The place of maximum constriction is at the lips for /m/, whereas it is at the alveolar ridge for /n/. Hence, the spectral characteristics of these two sounds, in the murmur region and formant transitions, are different. In quiet, nasal place of articulation is thought to be signaled by both the murmur

and formant transitions into the adjacent vowel [3, 4, 7]. The only study that examined place perception in noise is [6] in which the perception of place, manner, and voicing of syllable-initial consonants (including the nasals) were examined. Unfortunately, the study was limited to the vowel /a/.

We examine the perceptual role of both acoustic features (murmur and formant transitions) in identifying nasal place through an extensive series of perceptual experiments. The experiments examine the effects of additive white Gaussian noise, and additive speech-shaped noise, on place perception. The experiments also examine human perception of altered /CV/ syllables, whereby the murmur or the formant transitions are removed, in the presence of AWGN. An automatic speech recognition (ASR) system was then constructed to take into account the results of the perceptual experiments. System performance was compared to a baseline ASR system.

2. PERCEPTUAL EXPERIMENTS

2.1 Stimuli and Protocol

Stimuli consisted of CV syllables where the consonant was either /m/ or /n/, and the vowel was /a/, /i/, or /u/. Eight tokens of each syllable were recorded by two male and two female talkers of American English, resulting in a total of 192 syllables. The sampling rate was 16 kHz and the speech was coded with 16 bits. Perceptual experiments were a combination of identification and adaptive forced choice tasks, and were conducted in a sound-isolated chamber. Four healthy-hearing subjects participated in the experiments.

2.2 Additive Noise Experiments

In these experiments, white Gaussian noise (WGN) or speech shaped noise (SSN), modeled after the specifications of [1], was added, digitally, to the speech stimuli. The level of the noise varied in 5 dB steps. The noisy signals were 150 msec longer than the speech tokens. The speech tokens were placed 150 msec after the onset of the noise so that artifacts caused by the sudden onset of noise are avoided. The Signal-to-Noise Ratio

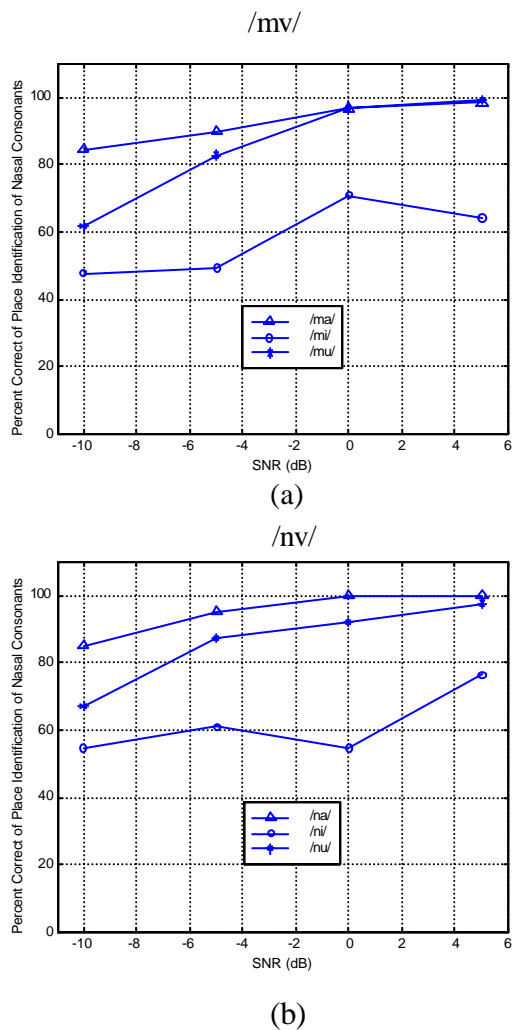


Figure 1: Average percent correct identification for nasal place in the presence of additive white Gaussian noise.

(SNR) was calculated based on the average energy of the speech signal and the calculation precluded silence in the speech segments, if any.

2.2.1 Results. Figures 1 and 2 summarize the results of the AWGN and SSN experiments, respectively. Notice the strong vowel-context effect, with /Ca/ being the most robust in the presence of noise and /Ci/ being the least robust. In the AWGN case, and at -10 dB SNR, percent correct recognition is above 80 for /Ca/ syllables. For /Ci/ syllables, on the other hand, place perception is difficult even at a high SNR of 5 dB. We speculate that /Ca/ is the most robust in noise because formant transitions are longer than they are in the other syllables. In our database, /na/ transitions were about 50-60 msec long, while they were about 15-

20 msec long for the other syllables. Shorter signals are more difficult to hear especially in the presence of noise [2]. Spectrograms of the syllables /na/ and /ni/ as spoken by a male talker are shown in Fig. 3. Notice the longer formant (especially F2) transition in /na/. In addition, F2, which carries important place information has the highest amplitude (relative to F1) in /Ca/ syllables, and the least in /Ci/ syllables.

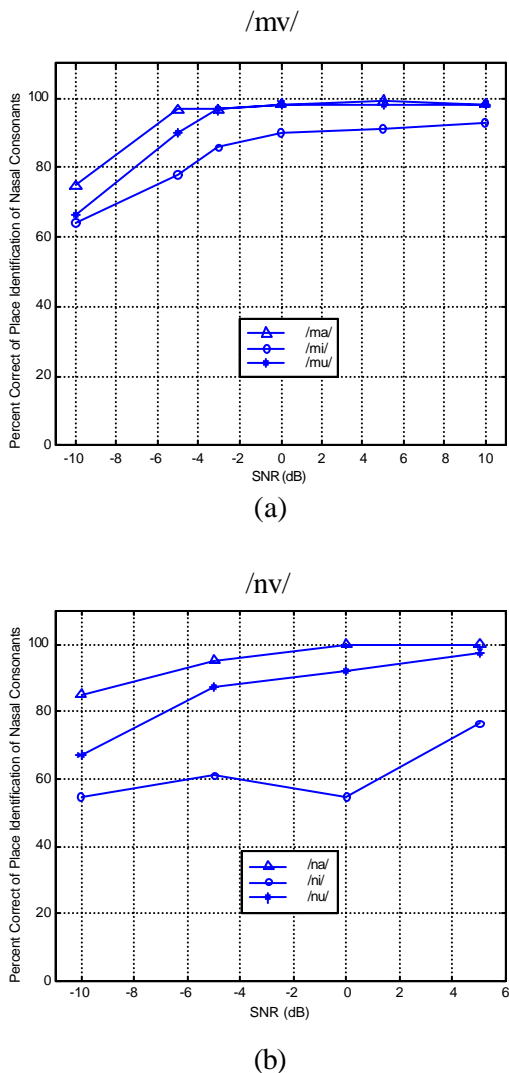


Figure 2: Average percent correct identification for nasal place in the presence of additive speech shaped noise.

The type of noise also affects perception. For example, a comparison of Figures 1 and 2 reveals that, at the same SNR, nasals are more difficult to perceive correctly in the presence of WGN than in the presence of SSN. This could be explained by the fact that speech-shaped noise is low-pass and as such, high-frequency

spectral cues can contribute to place perception if these cues are not masked by noise. This is especially true for /Ci/ syllables since F2 in this case is high (above 2000 Hz). The results clearly imply that the study of Miller and Nicely [6] does not generalize to all vowel contexts and to different noise shapes.

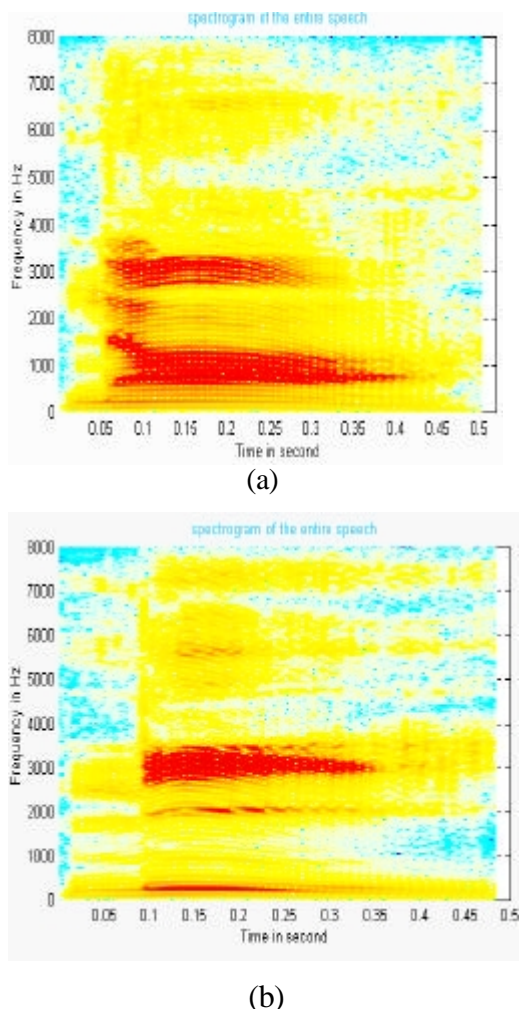


Figure 3: (a) Spectrogram of /na/, (b) Spectrogram of /ni/.

2.3 Examining the Role of the Murmur and Formant Transitions in Noise

To better quantify the role of the murmur and formant transitions on nasal perception, the following experiment was undertaken. Subjects were asked to identify nasal consonants in three different types of speech tokens: (1) CV syllables; (2) CV syllables without the murmur, and (3) CV syllables with 150 msec of the formant transition in the following vowels removed. The speech tokens were then added to WGN and presented

to listeners. An adaptive procedure based on the transformed up-down method by [5] was implemented. A correct response results in a reduction in threshold and an incorrect response results in a threshold increase. The convergence of the threshold occurs when there are 79% correct responses.

2.3.1 Results. Table 1 illustrates experimental results. A threshold increase implies that the sound can be identified reliably only if the the additive noise is lower than it was for the baseline case. For /Ca/ and /Cu/, removing the nasal murmur raises the threshold by about 2-3 dB, while removing the transition results in raising the threshold by about 24 dB for /Ca/ and 12 dB for /u/. Thresholds could not be found (procedure did not converge) for /Ci/ syllables when either the murmur or the formant transition was removed. These results clearly indicate that, in the presence of AWGN, formant transitions seem to play a critical role in identifying place for /Ca/ and /Cu/ syllables. In /Ci/ syllables, since the formant transitions are short and the amplitudes of F2 are relatively weak, the existence of both the murmur and the formant transitions is important for identifying place.

vowel	CV	w/o murmur	w/o transition
/a/	-12.5	-10.6	12
/u/	-7.5	-4.8	5.4
/i/	8.9	N/A	N/A

Table 1: The 79% correct threshold in dB for identifying nasal place in AWGN.

3. VARIABLE FRAME RATE (VFR) METHOD

3.1 The Algorithm As implied in the perceptual experiments, changes in spectral characteristics are important cues for discriminating and identifying speech sounds. These changes can occur over very short time intervals. Computing frames every 10 ms, as commonly done in recognition systems, is not sufficient to capture such dynamic changes. To illustrate this point, Figure 4 shows plots of MFCC vectors along a 100 ms segment surrounding the formant transition region in a /ma/ syllable. The frame length is 20 ms, but the frame step size is 10 ms in (a) and 2.5 ms in (b). Note that the murmur and steady-state region of the vowel are represented by (perhaps an unnecessarily large) number of MFCC vectors, while the critical formant transition region (13 ms) is only represented by one vector with a 10 ms frame step size and 2 (distinct) vectors when the step size is reduced to 2.5 ms.

We propose a Variable Frame Rate (VFR) algorithm [9]. The algorithm results in an increased number

of frames for rapidly-changing segments with relatively high energy and less frames for steady-state segments. Using MFCC feature vectors, the variable frame rate algorithm is implemented as shown in Figure 5. First, speech is analyzed with frame lengths of 25 ms (Hamming window) and a step size of 2.5 ms. We refer to these frames as the dense frames. Second, the difference ($d(i)$, where i is the time index,) between every two adjacent dense frames” is calculated. The average of these differences is then calculated over the whole utterance. Third, based on the weighted differences, some frames are kept and others are discarded. In particular, dense frames around a formant transition will be kept, while at the steady part of the signal, frames will be picked sparsely. It is important to note that the distance $d(i)$ is calculated as the energy weighted Euclidean MFCC distance: first the Euclidean distance of the MFCC vectors of two adjacent frames are calculated, then it is weighted by $(E - b)$, where E is the log energy of that frame, and b is a constant offset. This is different from the method proposed in where the Euclidean MFCC distance was used. Energy weighting is important so that segments which exhibit changes but are low in energy are discarded, since they may not be noise robust. Our previous experiments have shown a clear relationship between the energy of formant transitions and perceptual noise robustness. In addition, our pilot ASR experiments using Euclidean MFCC distance did not yield high recognition accuracy in noise. The two parameters a , the threshold, and b , log energy offset, are chosen experimentally. The choice of a will determine the average data rate. For example, if a is 4 (ratio of the 10 ms step size and the dense step size of 2.5 ms), then the resulting total number of frames will be nearly the same as that in a front-end with a frame step size of 10 ms. If a is larger than 4, then the average data rate will be less than 100 frames per second and vice versa. In our implementation, a was chosen to be 6.8. The log energy offset b was set to be the average E (over the entire utterance) divided by 1.5.

3.2 An Example of VFR Analysis Figure 6 illustrates how frames are picked for the utterance /ma/ as spoken by a male speaker. Part (a) shows a time waveform of the utterance. The upper part of (b) plots $d(i)$, the weighted feature distance between two adjacent frames - with a step size of 2.5 ms - and the lower part shows the result of the frame-picking algorithm where each bar indicates that a frame has been chosen for recognition. Note that near the transition region from the consonant to the vowel $d(i)$ is large. For this example, 50 out of 200 dense frames are picked. Around the transition region, all the dense frames (spaced by 2.5 ms) are kept while in the steady-state part of the vowel, only 3-4 frames, out of 20 frames, are selected corresponding to a step size which is larger than 10ms.

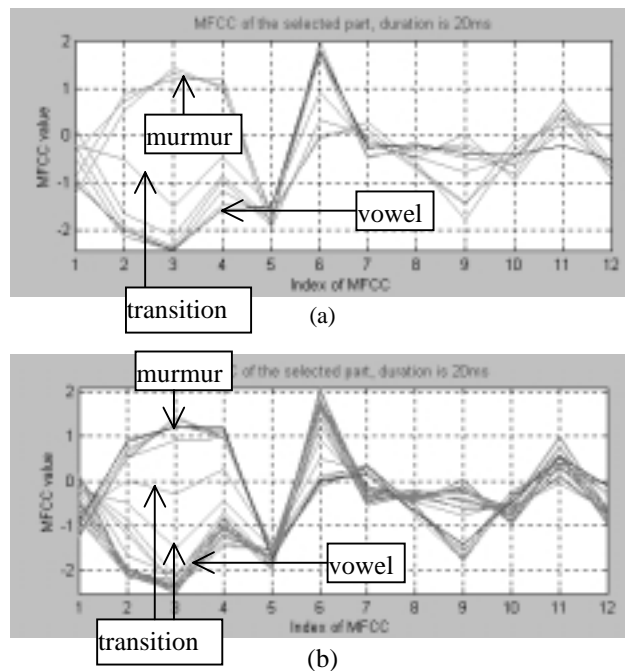


Figure 4: MFCC vectors around the transition of /ma/. (a) Window step = 10ms. (b) Window step = 2.5ms.

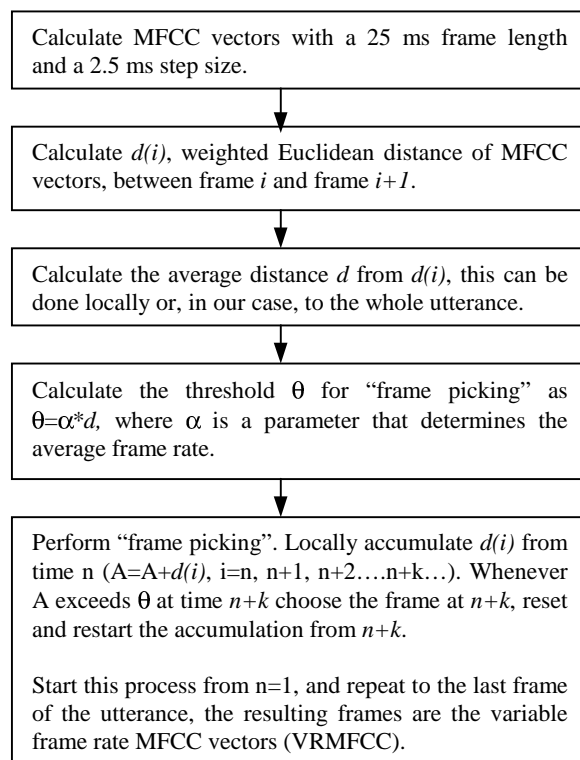


Figure 5: Flow chart of computing variable frame rate MFCC vectors.

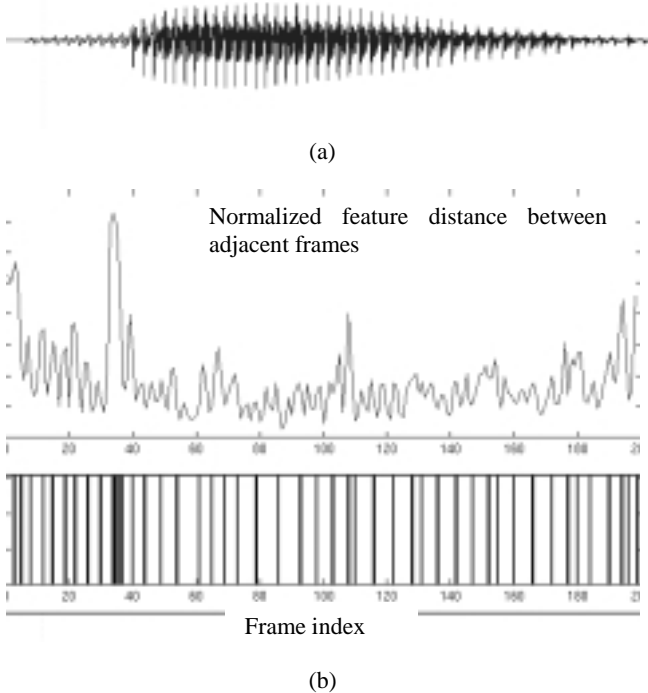


Figure 6: (a). The wave form of /ma/. (b). The "frame picking" results. The upper panel is the normalized $d(i)$ the lower panel is the frames that are picked out from the dense frame.

3.3 Recognition with the VFR Front End The variable frame rate method was used in ASR experiments using the nasal database described in Section 2, and the TIDIGITS database. In the experiments, the performance of the recognition system with two feature vectors were compared: MFCC, and MFCC vectors with peak enhancement [8] (hereafter referred to as MFCCP). First and second derivatives of these features were used. Training was done using clean data while testing was done with either clean or noisy data. Results for the nasal recognition experiment are shown in Table 2. Clearly, the variable frame rate approach together with a method for enhancing spectral peaks, gives the best performance at low SNRs. The VFR method was also used with the database Studio Quality Speaker Independent Connected Digit Corpus (TIDIGITS). Each left to right digit HMM model had 4 states, 2 mixtures, and a diagonal covariance matrix. 80 utterances from 80 speakers, (40 male and 40 female) were used to train each model. Test data were from the other 32 speakers (half male and half female).

We compared MFCC and MFCCP with their variable frame rate versions. The results are shown are summarized in Table 3. The results clearly illustrate that applying the VFR method to each feature vector improves recognition performance especially at low

SNRs. Increasing time resolution for rapidly changing segments, while keeping the time resolution low for steady parts, results in improved robustness.

	Clean	15 dB	5 dB	0 dB
MFCC	90	89	68	34
MFCCP	96	91	77	68
VFRMFCCP	100	96	81	71

Table 2: Recognition accuracy for the nasal database for different front ends.

SNR	20 dB	15 dB	5 dB	0 dB
MFCC	97	87	71	56
MFCCP	98	96	93	78
VFRMFCCP	97	97	96	89

Table 3: Recognition accuracy for MFCC, MFCCP, and VFRMFCCP front ends using the TIDIGITS database.

SUMMARY AND CONCLUSION

In this paper, we investigated the perception of place of articulation for the nasal consonants /m, n/ in adverse conditions which included AWGN, and additive speech-shaped noise. In addition, identification thresholds for the two consonants in AWGN were measured using an adaptive procedure. These thresholds were measured for the entire syllable, and for the syllable with either the murmur or formant transitions removed. Results show a strong vowel-context effect. For example, for /Ca/ and /Cu/ syllables, the formant transitions seem to play a bigger role in place identification (in the presence of additive noise) than the murmur. For /Ci/, both the murmur and the formant transitions appear to play an important role in identifying place. To investigate whether or not placing a larger emphasis on formant transitions would improve machine recognition performance, a recognition system was constructed which was sensitive to dynamic changes of the signal over short periods of time. The system had better performance than a baseline ASR system especially at low SNRs.

ACKNOWLEDGMENTS

Work is supported in part by NSF and NIH.

REFERENCES

- [1] Byrne, D. and H. Dillon, "The National Acoustic Laboratories' (NAL) new procedure for selecting the gain and frequency response of a hearing aid," *Ear and Hearing*, vol. 7, pp. 257-265, Aug. 1986.

- [2] Hant, J., Strobe, B., and A. Alwan "A psychoacoustic model for the noise masking of plosive bursts", J. Acoust. Soc. Am. (JASA), Vol. 101, No. 5, Pt. 1, 2789-2802, May 1997.
- [3] Kurowski, K., and Blumstein, S.E. 1984. "Perceptual integration of the murmur and formant transitions for place of articulation in nasal consonants," JASA, 76, 383-390.
- [4] Kurowski, K., and Blumstein, S.E. 1987. "Acoustic properties for place of articulation in nasal consonants," JASA, 81, 1917-1927.
- [5] Levitt, H. 1971. "Transformed Up-Down Methods in Psychoacoustics," JASA, 49, 467-477.
- [6] Miller, G.A., and Nicely, P.E. 1955. "An analysis of perceptual confusions among some English consonants," JASA, 27, 338-352.
- [7] Repp, B. 1986. "Perception of the [m] - [n] distinction in CV syllables," JASA, 79, 1987-1999.
- [8] Strobe, B., and Alwan, A. 1997. "A model of dynamic auditory perception and its application to robust word recognition," IEEE Trans. on Speech and Audio Processing, Vol. 5, No. 5, 451-464.
- [9] Q. Zhu and A. Alwan, "On the use of variable frame rate analysis in speech recognition", Proc. IEEE ICASSP, Istanbul, Vol. III, p. 1783-1786, 2000.