HUMAN AND MACHINE RECOGNITION OF NASAL CONSONANTS IN NOISE

Abeer Alwan, Jeff Lo*, and Qifeng Zhu
Department of Electrical Engineering, UCLA
Los Angeles, CA 90095
**NEC Electronics, Santa Clara, CA*

ABSTRACT

The nasal consonants /m, n/ are often confused in the presence of background noise. In addition, these consonants are difficult to recognize reliably by machine. In this study, the perception of the place of articulation for nasal consonants in adverse conditions is examined through a series of perceptual experiments. The experiments examined the effects of additive white Gaussian noise (AWGN), and additive speech-shaped noise on nasal place perception in CV syllables. Results show a strong vowel-context effect. For example, it appears that the role of the formant transitions is more critical than that of the murmur in signaling place for /Ca/ and /Cu/ syllables while both the murmur and formant transitions appear to be important in signaling place for /Ci/ syllables. A Hidden Markov Model (HMM)-based automatic speech recognition (ASR) system was then constructed to identify the nasals 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 allowed a greater focus on formant transitions significantly improving recognition performance in noise.

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 (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, /Ca/ 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 Figure 3; note that the formant

/mv/

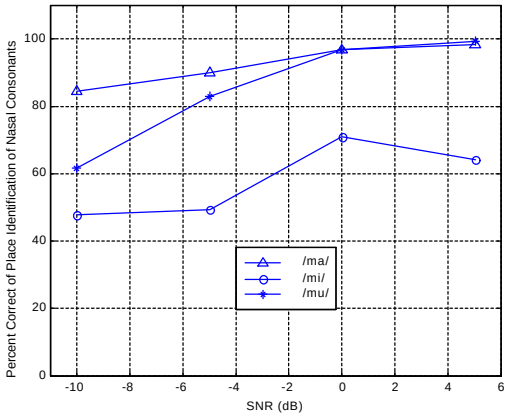

Fig 1a.

/nv/

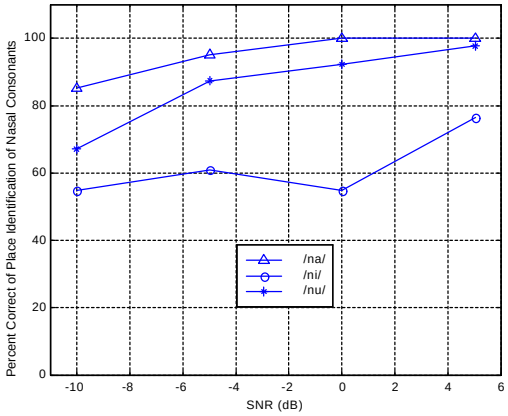

Fig 1b.

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

/mv/

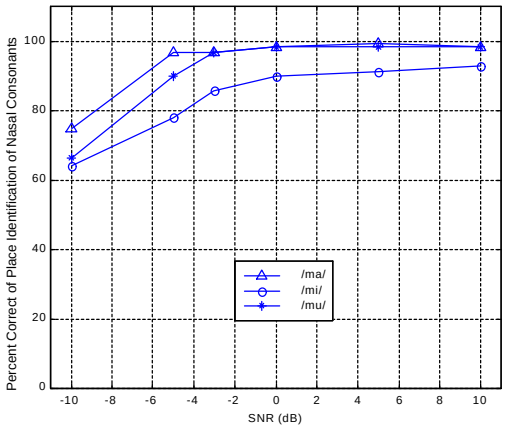

Fig 2a.

/nv/

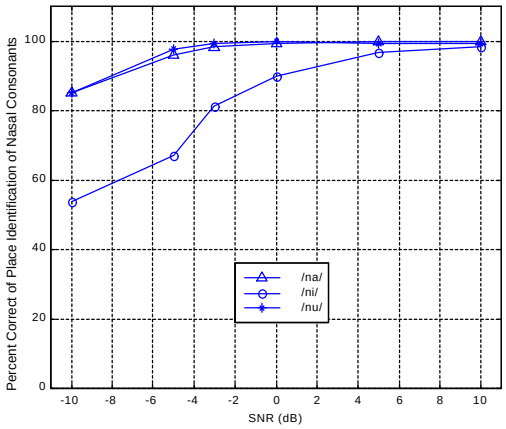

Fig 2b.

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

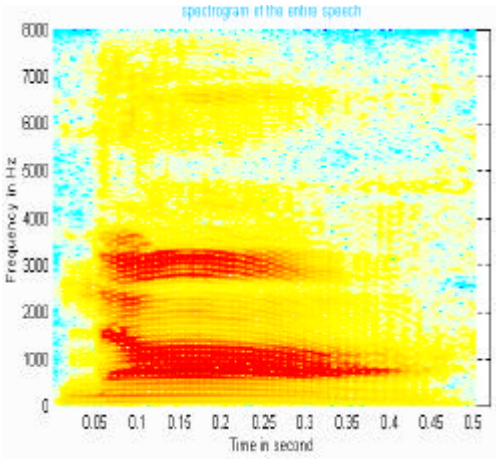

Fig 3a. Spectrogram of /na/

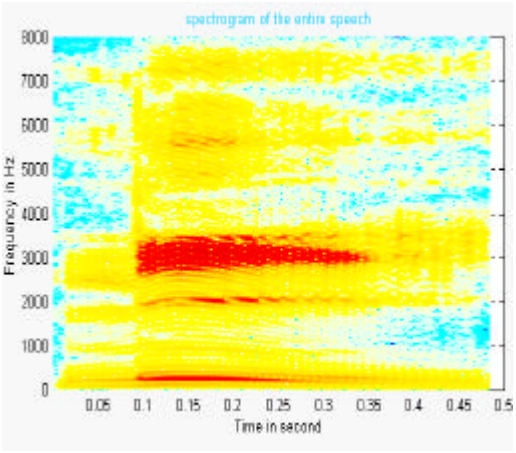

Fig 3b. Spectrogram of /ni/

Fig 3. Spectrograms of two syllables

vowel context	entire CV	murmur removed	transition removed
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.

	clean	SNR=15dB	SNR=5dB	SNR=0dB
MFCC	89	88	67	34
transition-sensitive MFCC	100	95	81	70

Table 2. Recognition accuracy (in percent) for an MFCC front-end and a transition-sensitive MFCC ASR system.

(especially F2) transitions are more pronounced 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.

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.

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 additive noise is lower than it is 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.

3. RECOGNITION EXPERIMENTS

3.1 ASR System

A Hidden Markov Model (HMM)-based automatic speech recognition (ASR) system was constructed; the task was to identify the nasal in the syllables as the signal-to-noise ratio varied. Training was done with half of the tokens, and testing was done with the other half. Training only used clean tokens, while testing included both clean and noisy signals (generated by adding digitally noise to the speech tokens). Inspired by the results of the perceptual experiments, modifications to a baseline ASR system were made. The modifications allowed the front end to place a larger weight on formant transition regions by tracking differences in the energy coefficients and in the MFCC amplitudes in successive short frames of the speech signal. If these differences are large, then more frames are analyzed in that segment. This is in contrast with the way ASR systems typically analyze speech, whereby frames are uniformly-sampled in time. In addition, our method enhances the position of spectral peaks in each frame by employing a technique described in [8]. Both the Mel-Frequency Cepstral Coefficients (MFCCs), and their first and second derivatives (deltas and delta deltas) were used in the evaluations.

3.2 Results

Table 2 summarizes recognition results for the /CV/ syllables as the SNR varies for both our system (transition-sensitive MFCCs) and that of the commonly-used MFCC front-end. Notice the significant improvement in recognition accuracy of our system, especially at low SNRs.

4. 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

This work was supported in part by NIH R29DC02033. Thanks to Brian Strope and James Hant for insightful discussions. We also thank our subjects for their cooperation.

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., Strope, 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] Strope, 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.