Manipulation of Consonants in Natural Speech

Jont Allen and Feipeng Li
Beckman Inst., Univ of IL, Urbana IL

September 25, 2009

Abstract

Starting in the 1920s, researchers at AT&T Research characterized speech perception. Until 1950, this work was done by a large group working under Harvey Fletcher, which resulted in the *articulation index*, an important tool able to predict average speech scores. In the 1950s a dedicated group of researchers at Haskins Labs in NYC attempted to extend these ideas, and then again at MIT under the direction of Ken Stevens, further work was done, on trying to identify the reliable speech cues. Most of this work after 1950 was not successful in finding speech cues, therefore today many consider it impossible. That is, many believe that there is no direct unique mapping from the time-frequency plane to consonant and vowel recognition. For example it has been claimed that context is necessary to successfully identify nonsense consonant-vowels. In fact this is not the case. The post 1950 work mostly used synthetic speech. This was a major flaw with all these studies. Also only average results were studied, again a major flaw.

In 2007 we carefully measured the consonant error for 20 talkers speaking 16 different consonants, in two types of variable noise. For many consonants, the human performance is well above chance at -20 dB SNR, and at 0 dB SNR, the score is close to 100% for most sounds. The error patterns for individual sounds are quite different from the average. Vowels preform very differently than consonants. The lesson learned is to carefully study token inhomogeneity.

The present work is a natural extension of these 1950 studies, but this time we have been successful and have determined the mapping. Using

1. extensive psychoacoustic methods,
2. working with a large data-base
3. of recorded speech sounds, with
4. the newly developed techniques that
5. use a model of the auditory system to
6. predict audible cues in noise, all
7. with a large number of listeners to evaluate the induced confusions,

we have precisely identified the acoustic cues for individual utterances and for a large number of consonants.

This paper explores the potential use of this new knowledge about perceptual cues of consonant sounds in speech processing. These cues provide deep insight into why Fletcher's articulation index is successful in predicting average "nonsense" speech syllables. Our analysis of a large number of nonsense Consonant-Vowel syllables from the LDC database reveals that natural speech, especially stop consonants, often contain *conflicting speech cues* that are characteristic of confusable sounds. Through the manipulation of these acoustic cues, one phone (a consonant or vowel sound) is be morphed into another. Meaningful sentences can be morphed into nonsense, or a sentence with a very different meaning. The resulting morphed speech is natural-sounding human speech. These techniques are robust to noise: a weak sound, easily masked by noise, can be converted into a strong one. Results of speech perception experiments on feature-enhanced /ka/ and /ga/ show that any modification of speech cues significantly changes, and can even improve the score in noise, for both normal and hearing-impaired listeners. The implications for ASR will be discussed.