



Time-domain auditory processing of speech

Alain de Cheveigné
CNRS - Ircam, 1 place Igor Stravinsky, 75004, Paris, France,
cheveign@ircam.fr

Speech patterns develop over time. Temporal phenomena extend from the very short to the very long: segregation, timbre and pitch cues, segmental, prosodic and rhetoric patterns, memory, learning, development and evolution. This paper focuses on the small-scale region, where time-domain descriptions merge with spectral descriptions. Temporal cues are used by the listener to organize the auditory scene, to parse acoustic information, and to assign the relevant fragments to one speaker among others, or among sources of interference. Cues include interaural delays (that vary according to source position), periodicity (of voiced speech), and envelope modulation features such as onsets. Temporal cues may determine pitch (intonation) and timbre (of vowels or consonants), for which it is also common to find accounts based on spectral cues. Frequency and time are closely linked, although some phenomena may be easier to describe in one domain or the other. With a frequency-selective cochlea the ear is equipped to analyze spectral patterns, but from physiology we know that their temporal counterpart is also available for analysis in the auditory nervous system. Synchrony to the acoustic signal degrades as one proceeds from the auditory nerve to the cortex, but accurate temporal patterns are present for processing at nuclei up to the inferior colliculus, and there is abundant evidence for neural "circuitry" specialized for time. Models that explain pitch on the basis of interval statistics between nerve firings are currently popular (although consensus is not complete), and similar models have been proposed for vowel timbre identification. Segregation of competing voices can be explained by models that use time cues, either to label frequency channels (created in the cochlea) as belonging to one voice or another, or to tease apart information within each channel on the basis of the temporal patterns of that channel. An important aspect of these models is that they delay to within the nervous system some of the processing that is often thought to occur at the periphery, within the cochlea. The cochlea then plays the role of a bank of "prefilters" to more central signal processing, rather than of a mere Fourier transformer as usually assumed. If this account is correct, frequency and temporal resolution are not entirely determined by properties of the cochlea.

This paper reviews models of time-domain processing of speech from several points of view. First, it questions whether these models are capable of implementing the expected functions. Parallels with speech technology are useful in this respect, as they give an idea of what is involved in implementing such functions in a working system. Auditory models have often been tried, not always with success, and this may reflect functional weaknesses of the models themselves. Second, it tries to understand where and how time-domain processing might be carried out within the auditory system, and attempts to identify clues among the behavioral data as to the processing that is actually going on. In the process, it addresses questions such as spectrum vs time, inter-event interval measurements vs ongoing correlation measurements, and feature duration vs integration time. [This work was supported in part by the Cognitique Programme of the French Ministry of Research.]