



P2-7

Magic time interval 250 ms plus-minus 100 ms

Hynek Hermansky
OGI School of Oregon Health & Sciences University, Portland, Oregon, USA
and
International Computer Science Institute, Berkeley, California, USA
hynek@ece.ogi.edu

Acoustic phonetics is for the past 50 years dominated by the spectrogram and its emphasis on the instantaneous short-term spectral envelope. The short-term spectrum of speech requires analysis over relatively stationary short segments of speech (of the order of centiseconds). The majority of automatic speech recognition (ASR) techniques attempt to classify speech from such snapshots of the signal. This paper reviews recent efforts to employ longer time spans (of the order of a syllable) in ASR.

A critical time interval of about a quarter of a second has been observed in many psychophysical phenomena such as forward masking, perception of gaps, growth of loudness, detection of constant energy stimuli, and auditory saltation. This correlates with maximum sensitivity to modulations around 4 Hz, i.e. 250 ms length of one period of the modulating waveform. Cortical receptive fields in many mammals span over a quarter of a second in temporal direction. Some of this evidence is summarized in Hermansky (1998).

Next, we report on our studies of hand-labelled speech. We observed that within-class variability of the spectrum of a phoneme increases gradually from the centre of the phoneme and reaches a maximal plateau at a distance of more than 100 ms from the centre of the phoneme, indicating that coarticulation extends over 200 ms (Kajarekar *et al.* 1999, Yang *et al.* 2000).

Then, we report on our attempts to improve the accuracy of the automatic speech recogniser. Such attempts yielded band-pass filters for filtering a temporal evolution of the short-term spectral envelope with time constants spanning about 200 ms (Hermansky 1998).

Finally, we present a new ASR technique that attempts to classify temporal trajectories of frequency-localized short-term spectral features. This technique is inherently more robust in the presence of non-linguistic factors such as linear distortions and noise (Hermansky and Sharma 1998).

Overall, we argue for the need to access portions of the speech signal of at least syllable-length for successful recognition of underlying phoneme classes.

Hermansky, H. (1998) Should recognisers have ears? *Speech Communication* 25, 3-27.

Hermansky, H., and Sharma, S. (1998) TRAPS classifiers of temporal patterns. Sydney, Australia. *Proceedings of ICSLP'98*, 3, 1003-1006.

Kajarekar, S., Malayath, N., and Hermansky, H. (1999) Analysis of source of variability in speech. Budapest, Hungary. *Proceedings of Eurospeech99* 1, 343-346.

Yang, H.H., Sharma, S., van Vuuren, S., and Hermansky, H. (2000) Relevance of time-frequency features for phonetic and speaker-channel classification. *Speech Communication* 31, 35-50.