Abstract

Experiments at Reading Auditory Laboratory have shown that humans exhibit perceptual constancy for hearing just as they do for seeing. Reverberation has an adverse effect on speech identification, however the effect is small for humans compared with machines. Implementing a constancy mechanism in machine listening may benefit the development of artificial listening devices.

Watkins (2005) has demonstrated perceptual compensation whereby the preceding context of a speech sound influences its identity. Test words drawn from the continuum between ‘sir’ and ‘stir’ were embedded in a spoken phrase. When the test word alone was reverberated the word ‘sir’ was reported more often than when no reverberation was present. However when the context and test word were reverberated in the same way, more steps were heard again as ‘stir’.

A computer model is described which simulates listeners’ performance in this task, based on the model of efferent suppression described by Ferry and Meddis (2007). In the model, the amount of efferent suppression increases when the context reverberation increases. The model provides a qualitative match to listeners’ performance in Watkins’ experiment. Early results are presented and extensions to the model introduced.

Background

• Our eventual aim is to improve the performance of machine listening systems by incorporating human-like processing into them.

• The present study aims to build a computer model that replicates the performance of human subjects in specific perceptual experiments.

• Perceptual constancy allows us to recognize an object or quality as constant under different conditions: we ‘account for’ our surroundings while listening.

• In speech perception, vowels or consonants are still perceived as constant categories despite considerable acoustic distortions introduced by real-room reverberation.

• Our current focus is compensation for effects of reverberation in the ‘sir/stir’ continuum (Watkins, 2005). Gradual imposition of the temporal envelope of ‘stir’ creates the impression of a stop consonant ‘t’ in ‘sir’. These test words are embedded in a spoken phrase ‘OK next you’ll get test words to click on’.

• In real-room reverberation, reflections fill the temporal gap of the ‘t’ in ‘stir’ making its amplitude envelope similar to that of a ‘sir’ utterance as the dynamic range is reduced owing to decay-tails that obliterate sharp offsets.

Methodology

• This pilot study is inspired by the work of Ferry and Meddis (2007) and Ghitza (2007).

• An efferent attenuation step is added to an existing auditory model, the dual-resonance-nonlinear (DRNL) filterbank, originally proposed by Meddis, O’Mard and Lopez-Poveda (2001) with parameters set to represent human listeners (Meddis, 2006).

• We use a simple template-based speech recogniser based on dynamic time warping (DTW) and cosine distance to compare the auditory nerve (AN) firing rate whilst the target words are being ‘heard’.

• The results reported here use templates for ‘sir’ and ‘stir’ from the extreme ends of the continuous.

• We characterise the efferent system as a resistive-capacitive circuit that reduces the frequency warping of the DTW based recogniser.

Results

• By hand-tuning the efferent attenuation level, the model provides a qualitative match to listeners’ performance in Watkins’ experiment.

• When the test word only is reverberated (less efferent attenuation, 9-12 dB) the category boundary shifts upwards (more ‘sir’ responses).

• When the context and test word are reverberated in the same way (more efferent attenuation, 21-23 dB) the category boundary shifts downwards (more ‘stir’ responses).

Conclusions

• Results from the model are consistent with the proposal that the efferent system could play a role in perceptual compensation for the effects of reverberation.

• In principle a good fit to listener data can be obtained if the amount of efferent attenuation applied to the test word is inversely proportional to the dynamic range of the context.

Ongoing Work

• Reverberation metrics are being evaluated in order to implement the model as a closed-loop feedback system: the amount of reverberation in the context (judged with a sliding time window) determines the amount of efferent attenuation.

• Within-channel mechanisms will be addressed in order to reflect the frequency-dependency of the efferent system (Guinan & Gifford 1988).

• The model is currently being tested with listening contexts that are time-reversed in speech and/or reverberation direction, and will subsequently be applied to diverse listening situations measured for human listeners at the Reading Auditory Laboratory.

Acknowledgements

This work is supported by an EPSRC grant (EP/E030602/1) entitled ‘Perceptual constancy in real-world listening by humans and machines’ and is undertaken in collaboration with the Reading Auditory Laboratory, thanks to Ray Meddis and Robert Ferry of Essex University for the DRNL program code.

References


