A computer model of perceptual compensation for reverberation

Amy Beeston and Guy J. Brown

a.beeston@dcs.shef.ac.uk | g.brown@dcs.shef.ac.uk

Speech and Hearing Group, Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello, Sheffield S1 4DP

Abstract

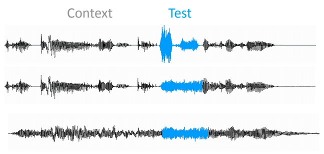
Experiments at Reading Auditory Laboratory have shown that humans exhibit perceptual constancy for hearing just as they do for seeing. Reverberation has an adverse affect on speech identification, however the effect is small for humans compared with machines. Implementing a constancy mechanism in machine 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 a 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 recognise 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 word] 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.

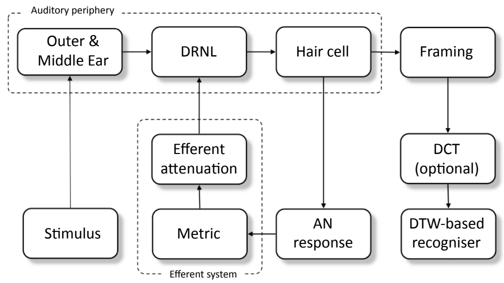


- 'sir/stir' category boundary recorded
- Test word only reverberated, category boundary shifts
- Context also reverberated, category boundary shifts back

- The *category boundary*, where 'sir/stir' perception flips, shifts in response to the quality of preceding sound.
- We ask whether the auditory efferent system could play a role, and if compensation could be characterised as a restoration of dynamic range?
- The efferent system can exert a suppressive influence on the basilar membrane, and has been implicated in control of dynamic range (Guinan & Gifford, 1988).

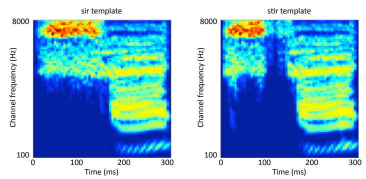
Methodology

- 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).

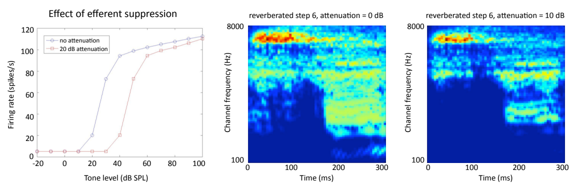


The model is configured to run with 80 frequency channels whose best frequencies lie in the (log-spaced) range 100 Hz to 8 kHz, with the stimuli presented at 56 dB SPL. The model is implemented in MATLAB, and includes a framework for running simulations on Sheffield's computing grid. Features for recognition were either the AN firing rate (computed at 5 ms intervals over 20 ms window) or the discrete cosine-transformed firing rate (15 coefficients, not including the first).

- 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 dry, unreverberated continuum.

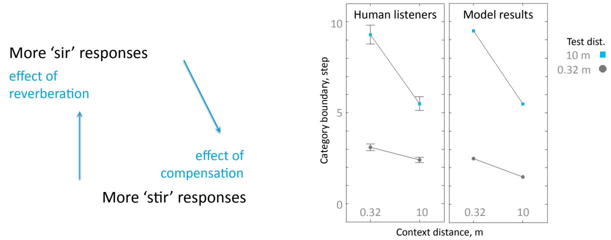


- Efferent suppression is modelled by attenuation in the nonlinear path of the DRNL, as described by Ferry and Meddis (2007).
- Suppression increases as the context reverberation increases.
- This helps to recover the dip in the temporal envelope following the 't' closure in 'stir'.



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).



Human listener results from Watkins (2005) Fig. 4(b) on page 258 show means and standard errors of category boundaries over a possible range from -0.5 to 9.5 to show the proportion of 'sir' responses minus 0.5. Model results are deterministic and have category boundaries in the same range, quantised to the nearest .5.

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/G009805/1) entitled 'Perceptual constancy in real-room 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

Ferry, RT & Meddis, R (2007). A computer model of medial efferent suppression in the mammalian auditory system. *Journal of the Acoustical Society of America* 122 (6), 3519-3526.

Ghitza, O (2007). Using auditory feedback and rhythmicity for diphone discrimination of degraded speech. *Proceedings of the 16th International Congress of Phonetic Sciences.*

Guinan, JJ & Gifford, ML (1988). Effects of electrical stimulation of efferent olivocochlear neurons on cat auditory-nerve fibers. III. Tuning curves and thresholds at CF. *Hearing Research* 31, 29-46.

Meddis, R (2006). Auditory-nerve first-spike latency and auditory absolute threshold: A computer model. *Journal of the Acoustical Society of America* 119 (1), 406-417.

Meddis, R, O'Mard, LP, & Lopez-Poveda, EA (2001). A computational algorithm for computing nonlinear auditory frequency selectivity. *Journal of the Acoustical Society of America* 109 (6), 2852-2861.

Watkins, AJ (2005). Perceptual compensation for effects of reverberation in speech identification. *Journal of the Acoustical Society of America* 118 (1), 249-262.