Listen with your skin: Aerotak speech perception enhancement system

Donald Derrick[1],[2], *Tom De Rybel*[1], *Greg A. O'beirne*[1], *Jennifer Hay* [1]

[1]University Of Canterbury, New Zealand Institute of Language, Brain & Behaviour (New Zealand)
[2]University of Western Sydney, MARCS Institute (Australia)

donald.derrick@gmail.com, tomderybel@yahoo.com,
gregory.obeirne@canterbury.ac.nz, jen.hay@canterbury.ac.nz

Abstract

Here we introduce Aerotak: A system for audio analysis and perception enhancement that allows speech perceivers to listen with their skin. The current system extracts unvoiced portions of an audio signal representative of turbulent air-flow in speech. It stores the audio signal in the left channel of a stereo audio output, and the air flow signal is stored in the right channel. The stored audio is used to drive a conversion unit that splits the left audio channel into a headphone out (to both ears) and right channel air pump drive signal to a piezoelectric pump that is mounted to the headphones. We have shown, using two-way forced-choice experiments, that the system enhances perception of voiceless stops and voiceless fricatives in noise such that 1 out of every 4 such words that would otherwise be missed will be heard correctly. We are currently conducting experiments on word identification while listening to a short-story, and are completing a stand-alone version of the Aerotak that works with real-time audio and from an embedded system. The short-story research and real-time system will be complete for InterSpeech 2014.

Index Terms: speech perception, aero-tactile integration, embodiment theory, audio perception enhancement

1. Introduction

Aerotak is the practical application of research showing that air puffs touching the skin help with speech perception, much as visual information does. If you put your hand in front of your face and say ba, you will not feel much, but if you say pa, you will feel a small puff of air - aspiration. Gick and Derrick exploited this difference to discover that mechanically produced air-puffs, similar to this aspiration, aligned with the speech signal and directed at the skin near the hand, neck [1], or ankle [2, 3] helped enhance perception of syllables beginning with aspirated stops (‘pa, ‘ta). This enhancement occurs if the air-puff occurs from as early as 50 milliseconds (ms) before and as much as 200 ms after the related acoustic signal [4].

But in all these experiments, the aero-tactile stimuli timing and strength was generated during post-processing, by hand, and based on researcher knowledge of air-flow produced from the lips during speech [5]. In order to use aero-tactile stimulation to help with real-world speech perception enhancement, it was necessary to extract air flow information from an audio signal directly. Fortunately, aperiodic information in the audio signal correlates reasonably well with turbulent air-flow released from the lips during speech. Techniques have been developed over the years to extract aperiodic information from an audio signal to varying degrees of accuracy [6, 7, 8, 9, 10, 11]. However, these were never intended to be used to convert audio information to air flow information, so we needed to design and build a new system for this purpose - Aerotak:

2. Current system

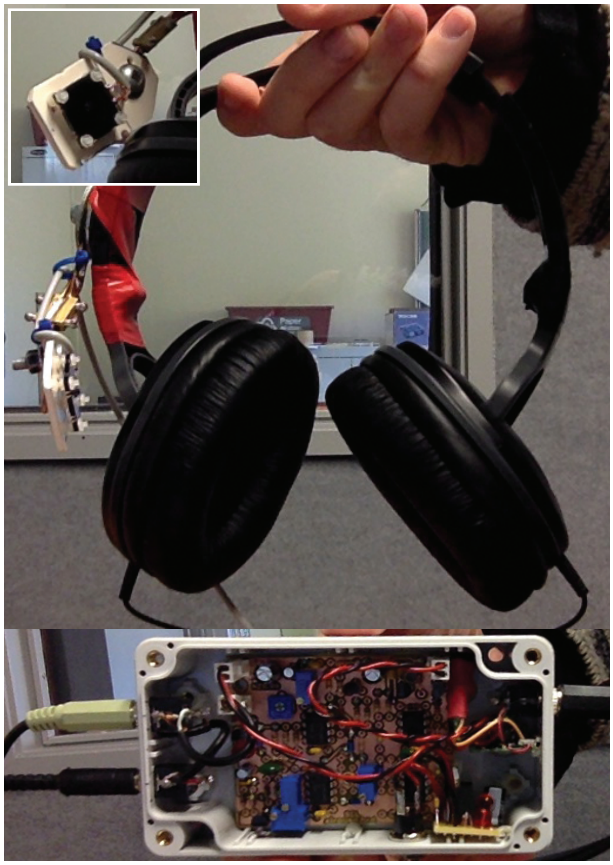

Figure 1: *Aerotak: Speech perception enhancement hardware proof-of-concept*

The current version of Aerotak takes a recorded audio source and extracts a signal representing turbulent air-flow from the lips during speech. The air-flow signal is passed to an air-flow controller that drives an air-flow source directed to the skin of the neck. This *System for audio analysis and perception enhancement* is currently under US patent submission [12].

The current system embodiment uses Octave [13] or Matlab to extract unvoiced portions of an audio signal, storing the audio signal in the left channel of a stereo audio output, and the air

Copyright © 2014 ISCA

flow signal in the right channel. The current system uses stored audio is used to drive a conversion unit that splits the audio into a headphone out (to both ears) and air pump drive signal to a Murata MZB1001T02 piezoelectric pump that is mounted to a set of Panasonic RP-HT265 headphones, as seen in Figure 1.

This extraction of turbulent air-flow information from audio is performed by a classifier that uses both zero-crossing rate and instantaneous frequency information computed from the audio. The Instantaneous frequency is computed using the direct energy separation algorithm (DESA) 1a algorithm [7]. DESA algorithms use both Teager's energy [6] and differential Teager's energy as input. These energy measures take into account the physics of speech production, assigning a heavier weighting to high-frequency utterances versus lower frequency ones. The production of turbulent airflow produces high-frequency, and therefore energetic, components, making this a relevant measure to help separate turbulent airflow information from the remainder of speech sound.

The audio zero-crossing rate is the second indicator used. It is a simple measure that differentiates sound generated from fundamental frequencies from sound generated by turbulent airflow. Strong periodic sound has relatively few zero-crossings per unit of time compared to noise, such as environmental noise and unvoiced portions of speech, which each also have distinct levels from each other. When strong periodic signals are present in the audio, they move the noise in the signal away from the zero base line so that it "rides" the fundamental frequency, resulting in significantly fewer zero-crossings per unit of time.

With these inputs, the classifier uses threshold operations to select the unvoiced portions of the audio signal. This resulting control signal is used to gate a signal that appropriately matches the envelope of the unvoiced portions of the audio. We find using a moving-average-filtered Teager's energy, when scaled by a natural logarithm, provides a suitable base for the gating operation to generate the air pump control signal.

The various inputs for the classifier are also filtered using a moving-average filter, and the final output from the classifier is processed with a median filter to prevent spurious spikes from interfering with the air pump drive signal, and therefore the smooth operation of the air pump.

3. Next version

We are currently developing an embodiment of Aerotak that will accept an audio input from microphone or external audio signal and process it in real-time on an Arduino Due board, along with improvements to the classifier code. This revised system will be presented at the show-and-tell at InterSpeech 2014.

The purpose of this system is to provide a prototype of aero-tactile audio enhancement that can be integrated into embodiments for headphones, emergency radios, hearing aids, and smartphones, as well as any audio device that may be used for audio speech communication. We have already demonstrated that Aerotak can enhance the perception, not just of voiceless stops, but also voiceless fricatives in two forced-choice experiments [14]. We are currently testing in short-storey experiments. Participants listen to a short story presented in blocks of 1-5 sentences and are asked to identify which of two words was used in a sentence context, comparing identification of minimal pairs contrasting 'p' vs 'b', (voiced vs. voiceless stops) 'b' vs 'f' (voiced stops vs. voiceless fricatives), and 'p' vs. 'f' (voiceless stops vs. voiceless fricatives) in noisy environments (0 DB signal-to-noise ratio). This experiment is designed to test aero-tactile enhancement to speech perception in real-world listening conditions using a real-world task - understanding speech in sentences. This experiment will be conducted with normal and hard-of-hearing participants, and pilot results have been promising.

4. Acknowledgements

The authors would like to thank Kieran Stone and Romain Fiasson for data collection. This research was funded by a New Zealand Ministry of Business, Innovation and Employment (MBIE) grant ONT-30003-HVMSSI-UOC for *Aero-tactile Enhancement of Speech Perception*.

5. References

[1] B. Gick and D. Derrick, "Aero-tactile integration in speech perception," *Nature*, vol. 462, pp. 502–504, 26 November 2009, doi:10.1038/nature08572.

[2] D. Derrick and B. Gick, "Full body aero-tactile integration in speech perception," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010) Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*, Makuhari, Chiba, Japan, 26-30 September 2010, pp. 122–125.

[3] ——, "Aerotactile integration from distal skin stimuli," *Multisensory Research*, vol. 26, pp. 405–416, 2013.

[4] B. Gick, Y. Ikegami, and D. Derrick, "The temporal window of audio-tactile integration in speech perception," *Journal of The Acoustical Society of America - Express Letters*, vol. 128, no. 5, pp. EL342–EL346, 2010.

[5] D. Derrick, P. Anderson, B. Gick, and S. Green, "Characteristics of air puffs produced in English 'pa': Experiments and simulations," *Journal of the Acoustical Society of America*, vol. 125, no. 4, pp. 2272–2281, April 2009.

[6] J. F. Kaiser, "On a simple algorithm to calculate the energy of a signal," in *International Conference on Acoustics, Speech, and Signal Processing, 1990 (ICASSP-90)*, 1990, pp. 381–384.

[7] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Transactions on Signal Processing*, vol. 41, no. 10, pp. 3024–3051, 1993.

[8] G. S. Ying, C. D. Mitchell, and L. H. Jamieson, "Endpoint detection of isolated utterances based on a modified Teager energy measurement," in *1993 International Conference On Acoustics, Speech, and Signal Processing (ICASSP-93)*, vol. 2, 1993, pp. 732–735.

[9] O. Deshmukh, C. Y. Espy-Wilson, A. Salomon, and J. Singh, "Use of temporal information: Detection of periodicity, aperiodicity and pitch in speech," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 776–786, September 2005.

[10] P. Zubrychi and A. Petrovsky, "Accurate speech decomposition into periodic and aperiodic components based on discrete harmonic transform," in *15th European Signal Processing Conference (EUSIPCO 2007)*, Poznan, Poland, September 3-7 2007, pp. 2336–2340.

[11] K. Aczél and I. Vajk, "Separation of periodic and aperiodic sound components by employing frequency estimation," in *16th European Signal Processing Conference (EUSIPCO 2008)*, Lausanne, Switzerland, August 25-29 2008.

[12] D. Derrick and T. De Rybel, "System for audio analysis and perception enhancement (us 61/939,974)," 02 2014.

[13] Octave community, "Gnu octave 3.8," 2014, www.gnu.org/software/octave/.

[14] D. Derrick, G. A. O'Beirne, T. De Rybel, and J. Hay, "Aero-tactile integration in fricatives: Converting audio to air flow information for speech perception enhancement," in *Proceedings of the 15th Annual Conference of the Internation Speech Communication Association (INTERSPEECH 2014)*, In Submission.