Aero-tactile integration in fricatives: Converting audio to air flow information for speech perception enhancement

Donald Derrick[1,2], *Greg A. O'Beirne*[1], *Tom De Rybel*[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, gregory.obeirne@canterbury.ac.nz,
tomderybel@yahoo.com, jen.hay@canterbury.ac.nz

Abstract

We follow up on research demonstrating that aero-tactile information can enhance or interfere with accurate auditory perception among uninformed and untrained perceivers [1, 2, 3]. We computationally extract aperiodic information from auditory recordings of speech, which represents turbulent air-flow produced from the lips [4, 5]. This extracted signal is used to drive a piezoelectric air-pump producing air-flow to the right temple simultaneous with presentation of auditory recordings. Using forced-choice experiments, we replicate previous results with stops, finding enhanced perception of /pa/ in /pa/ vs. /ba/ pairs, and /ta/ in /ta/ vs. /da/ pairs [1, 6, 2, 3]. We also found enhanced perception of /fa/ in /ba/ vs. /fa/ pairs, and /sha/ in /da/ vs. /sha/ pairs, demonstrating that air flow during fricative production contacting the skin can also enhance speech perception. The results show that aero-tactile information can be extracted from the audio signal and used to enhance speech perception of a large class of speech sounds found in many languages of the world.

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

1. Introduction

It has been known for some time that visual information can enhance speech perception [7]. Providing that information is relatively easy today with video-conferencing and voice/video over IP. Recent research has shown that aero-tactile information can also enhance speech perception [1, 6, 2, 3]. That is, the air puffs released from the lips during the production of voiceless stops can be replicated through machinery and directed towards the skin of a speech perceiver simultaneously with the relevant audio signal in order to enhance accurate speech perception.

However, 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 [4]. In order to make audio aero-tactile integration useful for real-world applications, it is necessary to extract air flow information from an audio signal directly. Techniques have been developed over the years to extract aperiodic information from an audio signal to varying degrees of accuracy [8, 9, 10, 11, 12, 13]. Because these were never intended to be used to convert audio information to air flow information, we developed a system for this purpose.

1.1. Air flow system

This system uses Octave [14] to extract unvoiced portions of an audio signal, storing the audio signal in the left channel of a stereo audio output, and the air flow signal in the right channel. The 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.

The extraction of the unvoiced portions of the audio is performed by a classifier using both zero-crossing rate and instantaneous frequency information computed from the audio.

Instantaneous frequency was computed using the direct energy separation algorithm (DESA) 1a algorithm [9], which itself uses both Teager's energy [8] and differential Teager's energy as input. These energy measures take into account the physics of speech production, and assign a heavier weighting to high-frequency utterances versus lower frequency ones. The production of unvoiced utterances requires many high-frequency, thus energetic, components, making this a relevant measure to help separate voiced from unvoiced utterances.

The zero-crossing rate is the second indicator used. It is a simple measure that differentiates strong fundamentals from unvoiced utterances. Strong fundamentals have 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 fundamentals are present in the audio, they "lift" the noise in the signal away from the zero base line so that it "rides" the fundamental, resulting in significantly fewer zero-crossings per unit of time.

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

The various inputs for the classifier were also filtered using a moving-average filter, and the final output from the classifier was processed with a median filter to prevent spurious spikes from polluting the air pump drive signal.

1.2. Distinguishing sound classes

In addition to needing to be able to extract air flow information from audio, in order for such a system to be useful in real-world speech recognition, it is necessary to demonstrate that it works for a larger class of sounds than just voiced vs. voiceless stops.

As a result, we designed a forced-choice battery of 8 experiments. These experiments tested: (1) /pa/ vs. /ba/, and (2) /ta/ vs. /da/ to compare the effects of air-flow on distinguishing between voiced and unvoiced stops, to replicate the results on *Aero-tactile integration in speech perception* [1]. The next two were: (3) /fa/ vs. /ba/, and (4) /ʃa/ vs. /da/, to test the effects of air flow on distinguishing between voiced stops and voiceless fricatives. The fifth experiment: (5) /va/ vs. /ba/, was designed to test the effects of air flow on distinguishing between voiced stops and voiced fricatives. The sixth experiment: (6) /d͡ʒa/ vs. /da/, was designed to test the effects of air flow on distinguishing voiced stops and voiced affricates. The seventh experiment: (7) /t ͡ʃa/ vs. /ta/, was designed to test the effects of air flow on distinguishing between voiceless stops and voiceless affricates. The eighth and last experiment: (8) /t͡ʃa/ vs. /ʃa/, was designed to test the effects of air flow on distinguishing between voiceless fricatives and voiceless affricates. Taken together, each experiment was designed to test if air flow would help in distinguishing ever smaller differences in speech sounds.

2. Methods

In order to create the stimuli for this experiment, four native speakers of New Zealand English were recorded in a soundproof booth using a Sennheiser MKH-416 microphone attached to a Sound Devices USB-Pre 2 microphone amplifier fed into a PC. These speakers produced 12 repetitions each of stimuli for use in 8 forced-choice experiments. Stimuli were presented in randomized order for the speakers to read aloud. For each speaker and for each token, 4 tokens were selected based on subjective audible clarity to be used as stimuli for these 8 experiments.

2.1. Psychometric experiment

To generate speech noise for each speaker, 36 recordings of their speech tokens were randomly superimposed 10,000 times within a 10 second looped sound file using an automated process. Noise created using this method results in a noise spectrum that is virtually identical to the long-term spectrum of the speech tokens from that speaker [15, 16], and ensures that the SNR's of the stimuli presented in the experiment were the same for each of the five speakers. Speech tokens and the noise samples for all speakers were adjusted to the same A-weighted sound level prior to mixing at different SNR's.

The headphones used in the experiment were placed on a Brel & Kjr Type 4128 Head and Torso Simulator (HATS) connected to a Brel & Kjr 7539 5/1-ch. Input/Output Controller Module (Brel & Kjr, Nrum, Denmark). The 1-second average A-weighted sound level of the samples was measured using the Brel & Kjr PULSE 11.1 noise and vibration analysis platform to confirm their output level. Using this information, output was set to an average (mean) of 75 dB for all tokens.

Participants were each asked to provide psychometric data for 4 of the 8 forced-choice experiments. For each of the 4 experiments they provided data for, they listened to 4 unique tokens of each of 2 syllables, produced by 4 speakers, at SNR's from -20 to 0 dB in increments of 2.5. This provided a total of 4x4x4x2x9, or 1152 tokens. As a result, the psychometric data collections took 1 hour each.

Sixteen (16) participants in total were used, providing 8 participants' worth of data for each of the 8 psychometric curves. Each participant was seated in a sound attenuated room with headphones. Using an experiment designed in PsychoPy [19, 20], they were presented with audio stimuli, and asked to press computer keys to indicate which of two words they heard based on the forced-choice designs described above.The 80% accuracy level was measured using an SI function based on generalized linear models, and the SNR results are provided in Table 1.

Table 1: *Table of experiments and Signal-to-Noise Ratio (SNR)*

experiment #	paradigm	SNR
1	/pa/ vs. /ba/	-5.5
2	/ta/ vs. /da/	-12.5
3	/fa/ vs. /ba/	-10.6
4	/ʃa/ vs. /da/	-13.0
5	/va/ vs. /ba/	-2.0
6	/d͡ʒa/ vs. /da/	-9.0
7	/t͡ʃa/ vs. /ta/	-16.0
8	/t͡ʃa/ vs. /ʃa/	-3.5

2.2. Air flow extraction

These stimuli syllables were also passed through our air flow extraction algorithm to generate a signal for driving a system to present air flow to the skin of participants simultaneous with audio stimuli. Because we used an early version of this air flow extraction system, the final product was examined, in order to determine how successful the system was, and to assess whether any manual repairs to the generated airflow signal were required. Minor manual repairs were conducted. These repairs were only necessary for some portions of the burst and aspiration produced for the syllable /pa/ due to the low amplitude nature of the audio portion for New Zealand English /p/ - no other outputs were modified.

2.3. Forced-choice experiment

The experiments were programmed in PsychoPy [19, 20] such that for each forced-choice experiment, the participant heard 16 tokens of each syllable without air flow, and 16 tokens of each syllable with air flow generated from the underlying sound file, for a total of 64 tokens. Each participant completed all 8 experiments, for a total 512 tokens, taking 30 minutes.

Twenty-four (24) participants were recorded. Each participant signed consent forms and was asked three questions about their hearing 1) "Do you have difficulty with your hearing?" (yes/no), 2) "Do you have difficulty following TV programs at a volume others find acceptable?" (no, slight, moderate, great), and 3) "Do you have difficulty having a conversation with several people in a group or in a noisy environment?" [17, 18]. Each of the participants answered "no" to all three questions, a requirement of participation to capture perceivers with normal hearing only. Five participants reported hearing issues based on the three questions listed above, so these participants were excluded. This left 19 participants, 18-40 years of age, 7 males and 12 females. All but two were raised in a monolingual English environment - one learned some French as a child, the other some Cantonese.

Participants were seated in a sound-attenuated room, were given the experiment preamble, and signed consent forms. They then had the headphones containing the air-flow device placed on their heads. During the experiment, participants were asked to identify which of two syllables they heard.

2.4. Statistical tests

For each experiment, the stimuli that contained underlying air-flow information were tested for perceptual enhancement effects. Therefore, for experiment 1-6, only the tokens with aspiration or frication were used in the statistical models, but for experiment 7 and 8, both tokens were used in the statistical models. Two statistical tests were completed for each of the forced-choice experiments. The first was a standard linear model comparing whether the answer provided by participants matched the auditory signal by whether air flow was simultaneously presented to the right temple. This model does not take into account any random effects. The second was a linear mixed-effects model taking into account subject variability seen in equation 1:

$$correct \sim flow + (1 + flow|participant) + (1|speaker) \quad (1)$$

Here, correct represents whether the answer matched the auditory signal or not, flow represents whether air-flow was presented to the skin of the right temple of the participants or not, and speaker represents a code distinguishing between the four speakers used for data in each experiment. This test accounts for both subject variability and variable comprehensibility of the speaker's voices (and noise masking) used in the experiments.

3. Results

For the statistical tests of each individual experiment, all results are based on 19 participants, with 608 observations for each experiment.

3.1. Experiments 1 and 2: Voiceless stops vs. voiced stops

For experiment 1, in which speakers were asked to identify /pa/ in a forced-choice /pa/ vs. /ba/ paradigm, the results of the standard linear model were significant (T = 2.681, p = 0.008). The results from the linear mixed-effects model were also significant (Z = 2.969, p = 0.003). The enhancement graph can be seen in Figure 1.

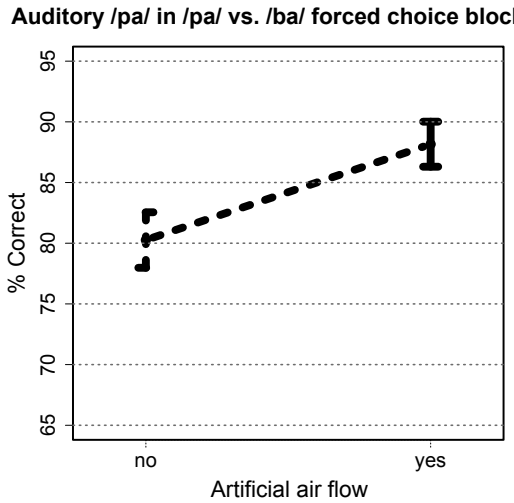

Figure 1: *Identification accuracy of auditory /pa/ in forced-choice /pa/ vs. /ba/ experiment - with and without air flow*

For experiment 2, in which speakers were asked to identify /ta/ in a forced-choice /ta/ vs. /da/ paradigm, the results of the standard linear model were significant (T = 2.102, p = 0.036). The results from the linear mixed-effects model were also significant (Z = 2.216, p = 0.027). The enhancement graph can be seen in Figure 2.

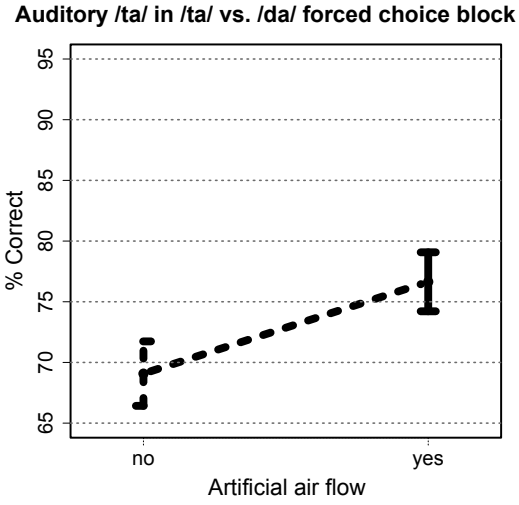

Figure 2: *Identification accuracy of auditory /ta/ in forced-choice /ta/ vs. /da/ experiment - with and without air flow*

3.2. Experiments 3 and 4: Voiceless fricatives vs. voiced stops

For experiment 3, in which speakers were asked to identify /fa/ in a forced-choice /fa/ vs. /ba/ paradigm, the results of the standard linear model were significant (T = 2.157, p = 0.0314). The results for the linear mixed-effects test were marginally significant (Z = 1.838, p = 0.066). The enhancement graph can be seen in Figure 3.

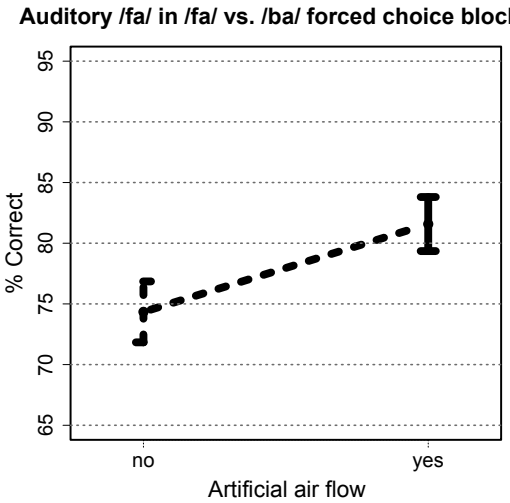

Figure 3: *Identification accuracy of auditory /fa/ in forced-choice /fa/ vs. /ba/ experiment - with and without air flow*

For experiment 4, in which speakers were asked to identify /ʃa/ in a forced-choice /ʃa/ vs. /da/ paradigm, the results of the standard linear model were significant (T = 2.44, p = 0.015). The results from the linear mixed-effects model were also significant (Z = 2.451, p = 0.014). The enhancement graph can be seen in Figure 4.

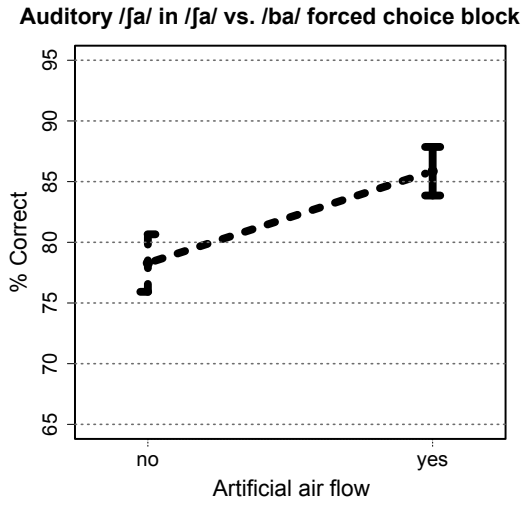

Figure 4: *Identification accuracy of auditory /ʃa/ in forced-choice /ʃa/ vs. /da/ experiment - with and without air flow*

3.3. Experiment 5: Voiced fricatives vs voiced stops

For experiment 5 (/ba/ vs. /va/), the results trended towards air-flow enhancement of accurate perception /va/, as seen in Figure 5, but the results of the standard linear model and the linear mixed-effects model were not significant.

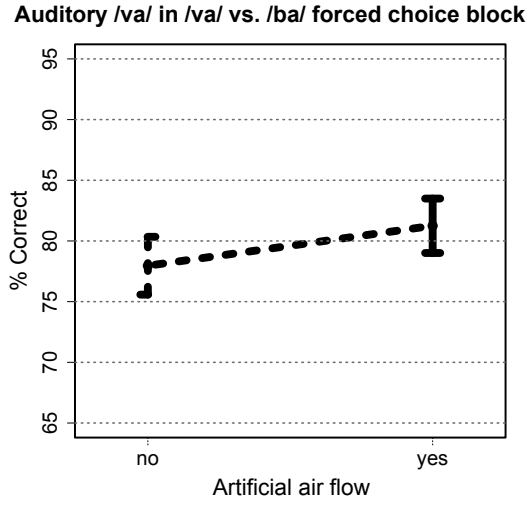

Figure 5: *Identification accuracy of auditory /ʃa/ in forced-choice /ʃa/ vs. /da/ experiment - with and without air flow*

3.4. Experiment 6-8

For experiment 6, /d͡ʒa/ vs. /da/ were not significant - accuracy was almost 95% for identifying d͡ʒa/ correctly regardless of whether there was air flow or not. That is, the audio signal was saturated and responses were already as accurate as they could be even before air flow enhancement. For experiments 7 and 8 the results were not significant despite accurate SNR's.

4. Discussion

The results demonstrate that it is possible to extract information representing air-flow produced from the lips during speech from the audio signal, use that information to drive a small piezoelectric pump to produce air-flow to the skin of the right temple, and with that air-flow information enhance perception of stops and fricatives in New Zealand English. The innovation achieved here has been submitted as a patent application [5], and is currently undergoing technology transfer.

The audio extraction system as it existed at the time of running this experiment worked for most sounds, though there were some inconsistencies in the data concerning the rather softly produced New Zealand English /p/ - some tokens needed hand correction of the extracted air flow data.

The results of using air flow information extracted from the audio signal were largely successful. Air flow directed to the skin enhanced perception of voiceless stops in a forced-choice decision task between voiceless and voiced stops, replicating previously reported results [1]. In addition, air flow directed to the skin enhanced perception of voiceless fricatives in a forced-choice decision task between voiceless fricatives and voiced stops, demonstrating that air-flow as well as air-puff information can enhance audio perception.

For the task of identifying voiced fricatives in a forced-choice decision task between voiced stops and voiced fricatives, the data trended in the expected direction for experiment 5 (/va/ vs. /pa/), but was not significant. The psychometric experiment failed to isolate the correct SNR for experiment 6 (/d͡ʒa/ vs. /da/). Participants were able to identify the tokens nearly perfectly regardless of air flow, and so the results of experiment 5 and 6 were inconclusive. For experiments 7 and 8, the results showed quite clearly that the distinction between the air-flow information from 7) voiceless stops vs. voiceless affricates, and 8) voiceless fricatives vs. voiceless affricates is simply too subtle for aero-tactile enhancement of perception in forced-choice decision tasks. Air flow did not degrade accurate audio perception, but it did not enhance it either.

The results show that aero-tactile speech perception enhancement systems can significantly enhance speech perception in noisy environments for a large class of speech sounds. The next step is to test for aero-tactile enhancement to speech perception during real-world tasks such speech comprehension of a short story. We are currently conducting this research. We also have plans to expand research to Hindi and Thai, which have more variety in stop contrasts suitable for such experiments.

5. Acknowledgements

Thanks to Scott Lloyd for his technical assistance, and Kieran Stone for data collection. 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*.

6. 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] 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.

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

[6] 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.

[7] W. H. Sumby and I. Pollack, "Visual Contribution to Speech Intelligibility in Noise," *Journal of the Acoustical Society of America*, vol. 26, pp. 212–215, 1954.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] O. community, "Gnu octave 3.8," 2014, www.gnu.org/software/octave/.

[15] S. Jansen, H. Luts, K. C. Wagener, B. Frachet, and J. Wouters, "The french digit triplet test: A hearing screening tool for speech intelligibility in noise," *International Journal of Audiology*, vol. 49, no. 5, pp. 378–387, 2010.

[16] C. Smits, T. S. Kapteyn, and T. Houtgast, "Development and validation of an automatic speech-in-noise screening test by telephone," *Internation Journal of Audiology*, vol. 43, no. 1, pp. 15–28, 2004.

[17] J. W. Pierce, "PsychoPy - Psychophysics software in Python," *Journal of Neuroscience Methods*, vol. 162, no. 1-2, pp. 8–13, 2007.

[18] ——, "Generating stimuli for neuroscience using psychopy," *Frontiers in Neuroinformation*, 2009.

[19] S. Gatehouse and W. Noble, "The speech, spatial and qualities of hearing scale (ssq)." *International Journal of audiology*, vol. 43, no. 2, pp. 85–93, 2004.

[20] W. Noble, "Identifying normal and non-normal hearing: Methods and paradoxes," November 2nd 2011, wARC talk, MARCS Auditory Laboratory.