Aero-tactile integration in Mandarin

Donald Derrick[1], Matthias Heyne[2], Greg O'Beirne[1], and Jennifer Hay[1]

[1]University of Canterbury, [2]Boston University
donald.derrick@canterbury.ac.nz

ABSTRACT

Previous research has shown that audio-aligned air puffs applied to the skin can enhance the perception of speech audio [12]. In this study, we applied dynamically varying air flow during two-way forced-choice identification of Mandarin words, comparing them to results of a study on English which showed perceptual enhancement for both stops and fricatives [6]. Two differences emerged: Psychometric testing identified the 80% accuracy signal-to-noise ratio for Mandarin words to be at -1.1 dB SNR, compared to -9.0 for English nonsense syllables. In addition, in Mandarin, aero-tactile stimuli only enhanced classification of voiceless stops, whereas it enhanced classification of voiceless stops and fricatives in English. These differences may partially result from the interaction of high conditional acoustic entropy in Mandarin compared to English [24] and air flow – that is, the Mandarin syllables had to be played with more preserved acoustic information, weakening the potential effect of air flow.

Keywords: Speech Perception, Speech Acoustics, Laboratory Phonology, Multimodal Phonetics

1. INTRODUCTION

Various studies have shown that visual information can enhance [29, 26, 17] or interfere [19, 18] with accurate speech perception, even at a young age [5, 28]. More recently, studies have shown that aero-tactile information can similarly enhance the perception of speech audio [10, 12, 13]. This enhancement follows from early notions of tactile perception of speech production by Alcorn [1]; cf. [14]. As with audio-visual speech [20], the benefit of airflow depends on temporal alignment [21, 30, 11] and even extends to visual-tactile stimuli presented without an audio signal [4]. More specifically, this line of research has demonstrated that 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, leading to the improved discrimination of such sounds in a two-way forced-choice paradigm.

These results were extended by including fricatives and affricates from English [9], and using a system that produces an air flow continuum rather than a binary flow versus no-flow paradigm. Doing so required two methodological improvements: a method for obtaining accurate representations of dynamic air flow in speech, and a system that could produce continuously varying artificial air flow. This study also applies dynamically varying air flow to the participants' right temple (see [8, 6]).

However, unlike the English study [9], where dynamic air flow was calculated from the audio signal, in this study of Mandarin Chinese, air-flow was directly estimated from the speech signal. Both procedures represent an improvement on prior research, where air flow was manually determined post-hoc on the basis of researcher knowledge [12].

This study also represents the first application of the aero-tactile enhancement of speech perception in a non-Western language (Mandarin), providing multi-lingual evidence for aero-tactile enhancement in speech perception.

1.1. Hypotheses

Phonetic and phonological observations regarding our own recordings on air flow intensity during consonant production in Mandarin lead us to formulate the following hypotheses about two-forced choice discrimination of stimuli pairs with varying air flow values.

Hypothesis 1: Air flow enhances two-alternative forced choice (2AFC) discrimination of Mandarin words as long as there is a measurable difference in the average speech air flow rates between the two choices.

Hypothesis 2: Enhancement will be proportional to the size of the difference in the air flow rates between the two choices.

2. METHODS

2.1. Recording of stimuli

In order to create the stimuli for this experiment, two native speakers of Mandarin were recorded in a sound-attenuated booth using a Sennheiser MKH-416 microphone attached to a Sound Devices USB-Pre 2 microphone amplifier fed into a PC. Both speakers were asked to produce twelve repetitions of

each stimulus presented on a computer screen. Four tokens (two from each speaker) were then selected for use in the perception experiment, based on subjective audible clarity.

2.2. Stimuli

During audio recording, a ping pong ball mounted on a carbon fibre rod attached to a number of strain gauges was placed directly in front of the speakers' lips to allow simultaneous and direct measurement of air flow while they were producing the relevant stimuli. The outputs of the ping pong puff device (PPP; cf. [7]) were fed into a computer as an amplitude-modulated sine wave and recorded in Audacity [2]. This sine wave was subsequently demodulated using an envelope detector and low pass filtered at 100 Hz using an algorithm implemented in Octave [23]. In a final step, all air flow signals were manually checked, and signal artefacts unrelated to speech air flow were deleted.

Table 1: Mandarin word pair experiments.

#	Paradigm	Han	Pinyin	SNR (dB)
1	[pa] vs. [p^{h}a]	八 趴	bā pā	-0.25
2	[ka] vs. [k^{h}a]	嘎 咖	gā kā	1.75
3	[ta] vs. [t^{h}a]	搭 他	dā tā	2
4	[ta] vs. [t͡sa]	搭 紮	dā zā	-0.5
5	[ta] vs. [t͡s^{h}a]	搭 擦	dā cā	-4.5
6	[pa] vs. [fa]	八 发	bā fā	-0.25
7	[ta] vs. [sa]	搭 撒	dā sā	-2
8	[t^{h}a] vs. [t͡sa]	他 紮	tā zā	-2.25
9	[t^{h}a] vs. [t͡s^{h}a]	他 擦	tā cā	-4
10	[t͡sa] vs. [t͡s^{h}a]	紮 擦	zā cā	-0.25
11	[t͡ʂa] vs. [t͡ʂ^{h}a]	扎 差	zhā chā	-1.5

To generate speech noise for each speaker, the recordings of their speech tokens were randomly superimposed 10,000 times within a 10 second looped sound file using an automated process. The resulting noise spectrum is virtually identical to the long-term spectrum of the speech tokens from that speaker [28], ensuring the SNR's of the experiment stimuli are all the same.

The experiment headphones were placed on a Brüel & Kjær Type 4128 Head and Torso Simulator connected to a Brüel & Kjær 7539 5/1-ch. Input/Output Controller Module (Brüel & Kjær, Nærum, Denmark). The 1-second average A-weighted sound level of the samples was measured using the Brüel & Kjær 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.

Air flow stimuli were generated using a piezoelectric air-pump. The pump has a 30 ms 5/95% rise time, produces about 15 cm H_2O maximum pressure, and 0.8 liters/minutes of air flow, or $1/12^{th}$ the maximum in conversational speech. To compensate for the low air flow, the pump head was placed about 2.5 cm away from the left temple, making the air flow contact on the skin more appropriate to that which could happen in close-contact speech.

2.3 Psychometric tuning

Fourteen participants, (12 female, 2 male), were used to provide psychometric tuning data for each of the 11 Mandarin 2AFC experiments listed in Table 1. Participants were seated in a sound-attenuated room and wore *Extreme Isolation EX-29* headphones (Direct Sound Headphones, Fenton, MO). They were presented with the audio stimuli through an experiment designed in PsychoPy [25], and asked to press computer keys to indicate which of two words they heard. For each of the Mandarin experiments listed in Table 1, two adaptive staircases were interleaved with random pair-ordering. Participants listened to speech-in-noise with SNRs that went up two dB when answered incorrectly, and down 0.5 dB when answered correctly, allowing for tracking of an 80% identification accuracy [16]. A minimum of ten reversals for each of the two syllable types were recorded, and averages were computed from the last five reversals. The average of the signal-to-noise ratios in decibels (SNR dB) for the two words were then used for the main experiment. These SNRs are shown in the last column of Table 1.

2.4. 2AFC experiments

We used an adaptive air flow production system [8, 6] to apply air flow to the participants' temples, depending on stimulus pair and air flow condition (see below for details). The air flow directly measured using the ping pong puff device was used to control the air flow system's piezoelectric pump mounted to *Extreme Isolation EX-29* headphones.

Stimuli presentation and response capture technique were identical to those used for psychometric tuning, and identical to that used for the previous study on English fricatives [9]. For each 2-way forced-choice experiment, the participant heard sixteen tokens of each syllable without air flow, and sixteen tokens of each syllable with air flow generated from the underlying sound file, for a total of 64 tokens. Each participant completed all eleven experiments, for a total 704 tokens, lasting about 40 minutes.

We collected data for 28 participants (20 females, 8 males). All of the participants were native monolingual Mandarin speakers except for one that was a bilingual Mandarin and English speaker, one bilingual Mandarin and Cantonese speaker, and one trilingual Mandarin, English, and Malay speaker. Following three questions about their hearing (see [22]), all reported normal hearing except for one with a slight difficulty listening to TV and three with slight difficulty listening to conversations in noisy environments.

2.5. Statistical Analysis

Generalized linear mixed-effects models (GLMM) were first run for each experiment to see if there was a significant enhancement in token identification accuracy for the audio + air flow condition compared to the audio only condition. However, because it is difficult to interpret a series of individual tests, and to avoid Bonferroni-correction-style errors in analysis, we also ran a model covering all of the experiments. We assigned a measure of the mean difference in air flow between the two word choices in each experiment, based on the air flow recordings obtained from our stimuli sources. This allowed each experiment to be placed on a continuous scale of energy difference. The model is shown in Formula 1:

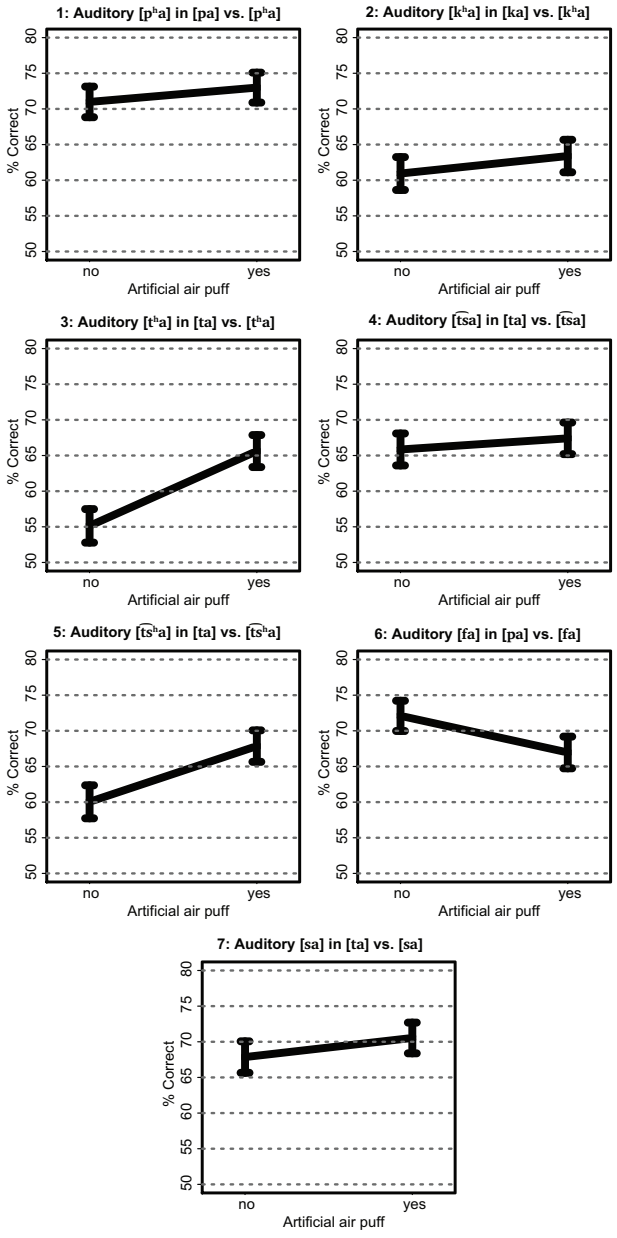
(1) correct answer ~ air flow * energy difference + (1 + (air flow * energy difference) | participant) + (1 | speaker)

Where *correct answer* refers to identification of the correct stimuli based on the acoustic signal, *air flow* refers to the presence or absence of the application of artificial air flow to the skin, *energy difference* represents a direct measurement of the mean difference in air flow delivered to the skin between the stimuli used for two choices in the 2AFC experiment (i.e. [pa] vs. [p^{h}a]). The fixed effects identify the interaction of air flow and energy difference. The sub-formula *(1 + (air flow * energy difference) | participant)* represents the full-factor random effect for the main effect for each participant. *Speaker* represents the ID of the speaker who provided the acoustic stimuli, and the sub-formula *(1 | speaker)* is a simple intercept for the difference in intelligibility for each speaker. We also attempted an analysis of non-linear effect of energy difference in the model using *rcs* [30], but found that doing so did not significantly improve the model.

3. RESULTS

The individual interaction plots for each of the 11 experiments are shown in Figures 1 and 2. Figure 2 contains experiments 8-11, where both choices have air flow in the underlying consonant, so both have valid comparisons between auditory only and auditory + airflow conditions. None of the results were statistically significant except for audio + tactile enhancement in experiments 3 ($Z = 3.346$, $p < 0.001$) and 5 ($Z = 2.602$, $p = 0.009$), and audio + tactile interference in experiment 9 ($Z = -3.148$, $p = 0.002$), and 10 ($Z = -3.176$, $p = 0.002$).

Figure 1: Interaction plot: Aero-tactile perceptual enhancement by experiment (1-7).



The GLMM shows that participants were significantly less accurate at identifying the audio

correctly in the audio + air flow condition than in the audio-only condition. However, this trend was reversed when the energy difference in the air flow between the two choices was high enough (see Table 2). The interaction effect, as obtained from the statistical model's predictions [3], can be seen in Figure 3.

Figure 2: Interaction plot: Aero-tactile perceptual enhancement by experiment (8-11).

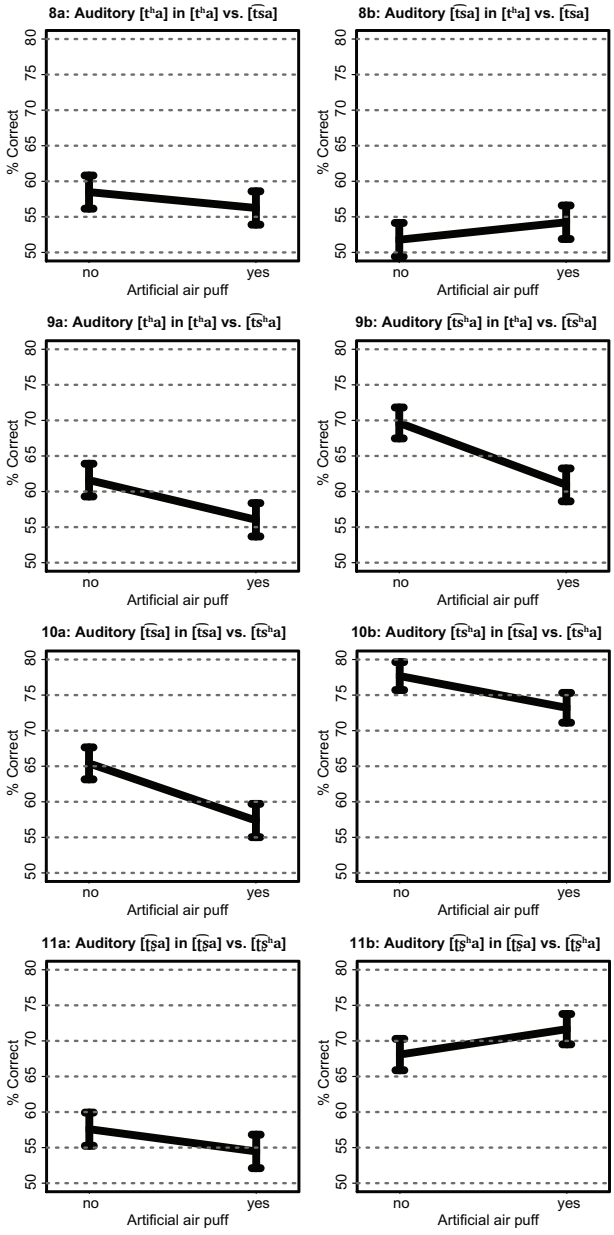
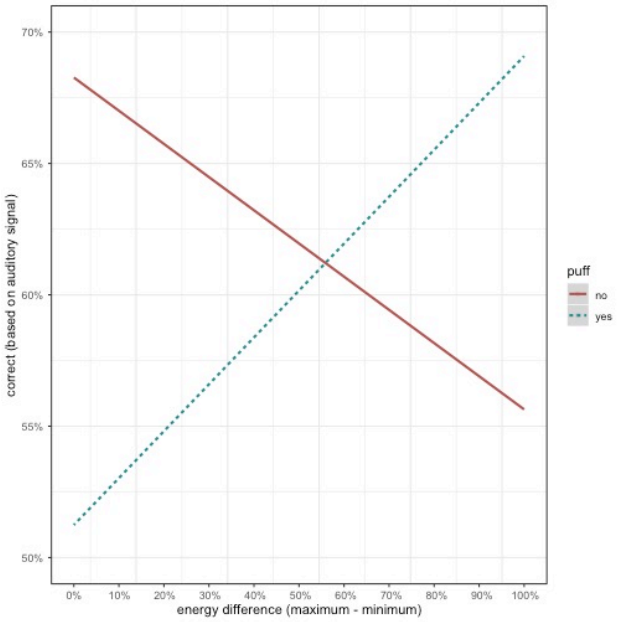


Table 2: Table of fixed effects outcomes for model formula 1 (* = significant, $\alpha = 0.05$).

Fixed	Estimate	StErr	z-value	p-score
(Intercept)	0.717	0.178	4.030	<0.001 *
air flow (yes)	-0.252	0.119	-2.115	0.0344 *
energy diff	-5.137	6.894	-0.745	0.456
flow:energy	12.39	5.736	2.161	0.0307 *

Figure 3: Interaction plot: Aero-tactile perceptual enhancement by energy difference from LMER.



4. DISCUSSION

The results of each experiment show that in Mandarin, air flow only enhanced speech perception for two experiments, both involving voiced vs. voiceless stops. Examining airflow energy differences directly shows that for Mandarin, air flow enhances classification of 2AFC tasks as long as there is a sufficient difference between the expected air flows of the paired stimuli. However, if the air flows are not sufficiently different, they interfere with audio speech perception. These results are similar to those found for English [9], but air flow differences between stimuli need to be of larger magnitude in Mandarin to provide a discrimination benefit to the listener.

These results are in agreement with Oh's [24] analysis of phonological complexity in Mandarin: Mandarin has twice the amount of *conditional entropy* as English, calculated as "the average amount of information taking contextual information into account" (p. 8) based on bits per linguistic units, among monosyllabic tokens (Figure 2.10, p. 63), once tone is accounted for. As a result, Mandarin required increased audio clarity, on average -1.1 dB SNR compared to English's -9 dB SNR [20]. This higher clarity reduced the potential for a benefit of air flow, thus requiring a greater distinction between air flow rates of the two choices for each 2AFC experiment compared to what was required for English.

7. REFERENCES

[1] Alcorn, S. 1932. The Tadoma method. *Volta Review*, 34, 195-198.

[2] Audacity team 2014. Audacity ®: Free Audio Editor and Recorder, version 2.0.6 [Computer program]. *Retrieved April 20th 2014 from http://audacity.sourceforge.net/"*

[3] Baayen, R. H. 2013. languageR: Data sets and functions with *Analyzing Linguistic Data: A practical introduction to statistics*. R package version 1.4.1. https://CRAN.R-project.org/package=languageR

[4] Bicevskis, K., Derrick, D., Gick, B. 2016. Visual-tactile integration in speech perception: Evidence for modality neutral speech primitives. *Journal of the Acoustical Society of America*, 140(5), 3531-3539.

[5] Burnham, D., Dodd, B. 1996. Auditory-visual speech perception as a direct process: The McGurk effect in infants and across languages. In Stork, D. G., Hennecke, M. E. (Eds.), *Speechreading by humans and machines* (pp. 103-114). Berlin: Springer.

[6] Derrick, D., De Rybel, T. 2015. System for audio analysis and perception enhancement. *PCT patent number* WO 2015/122785 A1.

[7] Derrick, D., De Rybel, T., Fiasson, R. 2015. Recording and reproducing speech airflow outside the mouth. *Canadian Acoustics*, 43(3), 102–103.

[8] Derrick, D., De Rybel, T., O'Beirne, G. A., Hay, J. 2014. Listen with Your Skin: Aerotak Speech Perception Enhancement System. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014)*, 1484-1485.

[9] Derrick, D., O'Beirne, G. A., De Rybel, T., Hay, J. 2014. Aero-tactile integration in fricatives: Converting audio to air flow information for speech perception enhancement. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014),* 2580–2584, Singapore.

[10] Derrick, D., Gick, B. 2013. Aerotactile integration from distal skin stimuli. *Multisensory research*, 26(5), 405-416.

[11] Gick, B., Ikegami, Y., & Derrick, D. 2010. The temporal window of audio-tactile integration in speech perception. *Journal of the Acoustical Society of America*, 128(5), EL342-EL346.

[12] Gick, B., Derrick, D. 2009. Aero-tactile integration in speech perception. *Nature* 462(7272), 502–504.

[13] Gick, B., Jóhannsdóttir, K. M., Gibraiel, D., & Mühlbauer, J. 2008. Tactile enhancement of auditory and visual speech perception in untrained perceivers. *Journal of the Acoustical Society of America*, 123(4), EL72-EL76.

[14] Fowler, C. A., Dekle, D. J. 1991. Listening with eye and hand: cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 17(3), 816.

[15] Harrell, F. E. 2018. rms: Regression Modeling Strategies, R package version 5.1-2. https://CRAN.R-project.org/package=rms

[16] Kaernbach, C. 1991. Simple adaptive testing with the weighted up-down method. *Attention, Perception, & Psychophysics*, 49(3), 227-229.

[17] Macleod, A., Summerfield, Q. 1990. A procedure for measuring auditory and audiovisual speech-reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use. *British Journal of Audiology*, 24(1), 29–43.

[18] Massaro, D. W., Cohen, M. M., Gesi, A., Heredia, R. 1993. Bimodal speech perception: An examination across languages. *Journal of Phonetics*, 21, 445-478.

[19] McGurk, H., MacDonald, J. 1976. Hearing lips and seeing voices. *Nature*, 264, 746–748.

[20] Munhall, K. G., Gribble, P., Sacco, L., Ward, M. 1996. Temporal constraints on the McGurk effect. *Perception & Psychophysics*, 58(3), 351-362.

[21] Navarra, J., Vatakis, A., Zampini, M., Soto-Faraco, S., Humphreys, W., Spence, C. 2005. Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration. *Cognitive Brain Research,* 25(2), 499-507.

[22] Noble, W. 2011. *Identifying normal and non-normal hearing: Methods and paradoxes.* wARC talk, MARCS Auditory Laboratory, Sydney, Australia.

[23] Octave community 2014. Gnu octave 3.8. <www.gnu.org/software/octave/>.

[24] Oh, Yoon Mi. 2015. *Linguistic Complexity and Information: Quantitative Approaches*. PhD dissertation, University of Lyon, Lyon, France.

[25] Peirce, J. W. 2007. PsychoPy Psychophysics software in Python. *Journal of Neuroscience Methods*, 162(1-2), 8-13.

[26] Reisberg, D., Mclean, J., Goldfield, A. 1987. Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. E. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip–reading* (pp. 97–114). Hillsdale, NJ: Lawrence Erlbaum.

[27] Rosenblum, L. D., Schmuckler, M. A., Johnson, J. A. 1997. The McGurk effect in infants. *Attention, Perception, and Psychophysics*, 59(3), 347–357.

[28] Smits, C., Kapteyn, T. S., Houtgast, T. 2004. Development and validation of an automatic speech-in-noise screening test by telephone. *International Journal of Audiology*, 43(1), 15–28.

[29] Sumby, W. H., Pollack, I. 1954. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America* 26(2), 212–215.

[30] Van Wassenhove, V., Grant, K. W., Poeppel, D. 2007. Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, 45(3), 598-607.