Session 39#20

Aerodynamic Modeling of Coarticulation for Concatenative Speech Synthesis

Kevin B. McGowan
Department of Linguistics, University of Michigan, Ann Arbor

Introduction

We know from decades of speech perception research that listeners can perceive and use a wide array of fine-grained phonetic details, including the detailed coarticulatory influences that nearby sounds have on each other, when perceiving speech.

I believe coarticulation provides the listener with a network of informative cues and is key to understanding our ability to disambiguate meaningful speech sounds from apparently noisy inputs. These coarticulatory cues are often missing or contradictory in text to speech (TTS) synthesis output.

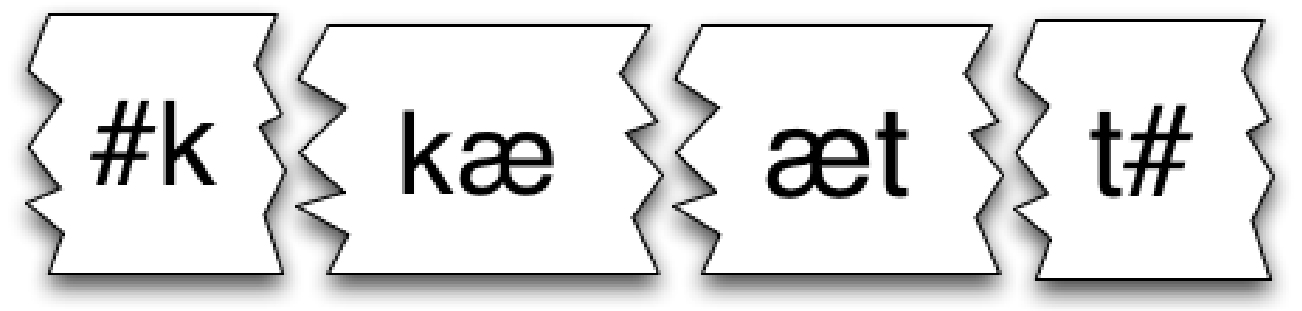
> *"If synthetic speech is to be listened to for long periods with the intention of getting the content straight, the synthesis must be more than interpretable. It must be accurate in ways that the person doing the synthesis cannot hear directly." — Whalen 1984*

Background: Coarticulation

There is consensus in speech perception research that coarticulatory information affects listener judgments, but researchers disagree on the perceptual usefulness of the information:

- Coarticulation introduces problematic variation (Ladefoged 2001) or renders contrasts less distinct (Lindblom 1990).
- Some instances of coarticulation overlap and obliterate featural cues; other instances enhance the perceptual saliency of neighboring features. (Stevens & Keyser 2008).
- Coarticulation is useful information that aids listeners in tracking the speakers' articulatory gestures (Fowler 1996).

Background: Concatenative Speech Synthesis



Concatenative synthesis works by stringing together sound units chosen from a large database of recorded speech. These units are chosen to minimize two **acoustic metrics**: the **join cost** of aligning a particular unit with the desired speech output and the **join cost** of adjoining the next sound to the most-recently selected unit:

$$C(t_1^n, u_1^n) = argmin(\sum_{i=1}^{n} C^{target}(t_i, u_i) + \sum_{i=2}^{n} C^{join}(u_{i-1}, u_i))$$

- *t* is the target phone in the sequence and
- *u* represents the unit of sound to be appended.

Implicit Assumptions:

1. There are invariants in the speech stream that identify segments (so, for example, a segment from one utterance can be used to produce the percept of that segment in another context).
2. These invariants are acoustic.

Early diphone synthesizers attempted to eliminate coarticulation mismatches by recording only carefully articulated diphones in heavily controlled articulatory contexts (leading to interpretable but unconvincing speech). Unit Selection systems use enormous databases of speech and synthesize utterances by preferring units that were contiguous in the source recordings. One of the greatest limitations of these systems is the jarring juxtaposition of perfectly natural-sounding speech (using contiguous units from the database) with mis-matched units from another part of the database —**we believe the solution to this problem lies in modeling coarticulation directly.**

Theoretical Goals

We take the position that coarticulation is *signal* rather than noise and serves to facilitate listeners' perception of speech (including synthesized speech). The primary goal of the present project is to develop a principled join cost calculation that explicitly takes coarticulation into account when selecting acoustic units.

- Baseline: Is accurate coarticulatory information perceptually useful in synthesized speech?
- Do listeners *prefer* this more accurate synthesis?
- Is airflow a useful and efficient means of automatically labeling a speech synthesis database with fine-grained coarticulatory detail?

Method

1. Database Recording

A native speaker of a Southeastern-Michigan dialect of English read the 452 sentence 'phonetically-balanced' portion of the TIMIT database (Fisher *et al.* 1986) in a sound booth from prompts displayed on an LCD screen. The speaker was not a professional voice actor (contra to recommendations in the synthesis literature).

2. Airflow Data Collection

The same speaker then re-recorded these prompts while attached to the EVA2 pneumotachograph for both oral and nasal airflow data collection. The speaker had a silicon tube inserted in one nostril and wore a flexible silicone mask to capture nasal and oral airflow respectively. The silicone mask necessarily distorts the acoustic signal —requiring the recording of separate databases (see discussion).

To maximize utterance similarity between the acoustic and aerodynamic recordings, the TIMIT prompts for these recordings were delivered by playing-back the original acoustic recordings over headphones.

Voice talent and EVA2 airflow system.

3. Reference Voice Creation

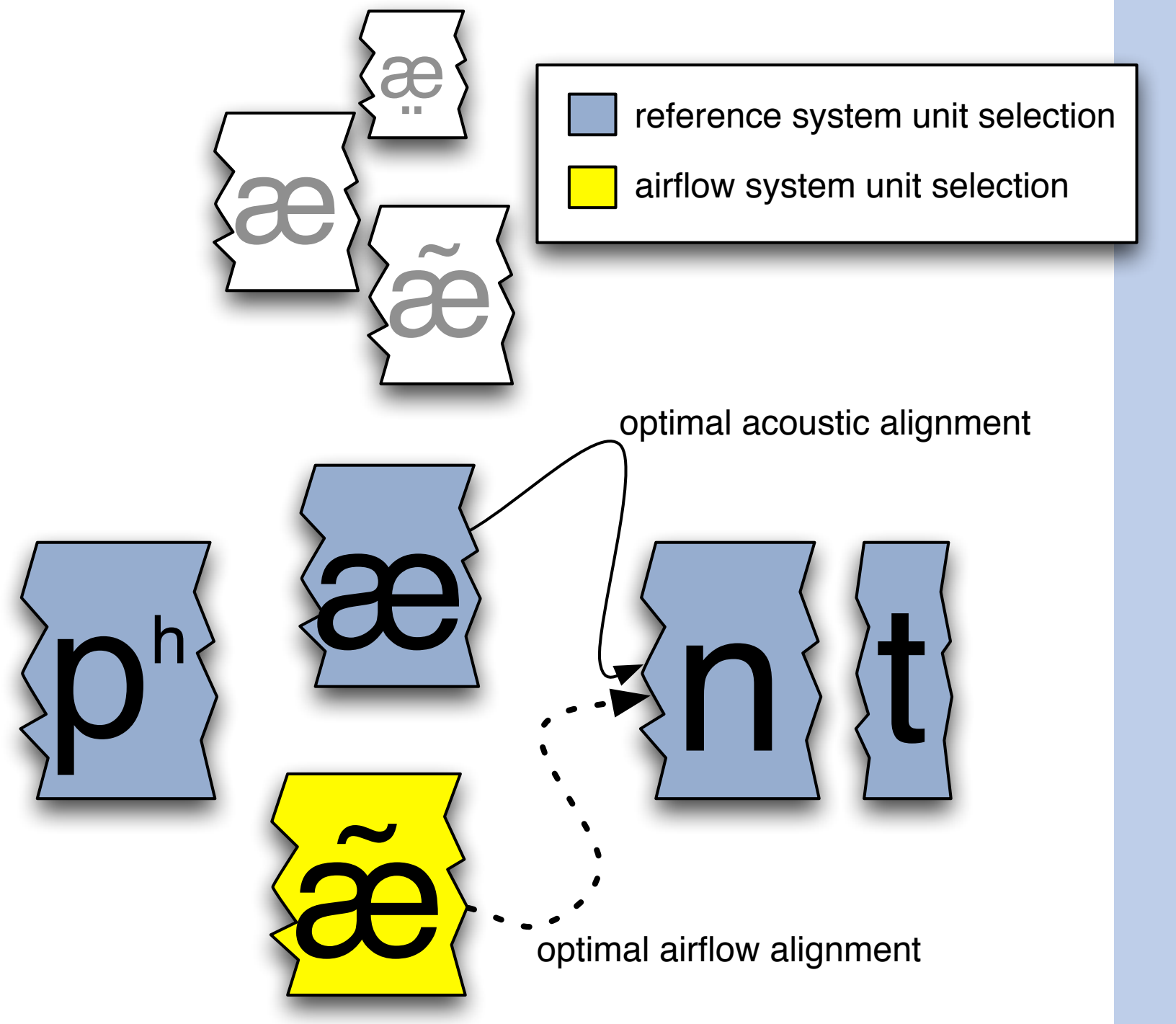
The acoustic recordings were used to create a clunits voice (Black & Taylor 1997) using the Festival open source speech synthesis system. Clunits was chosen both for its conceptual simplicity and its use of phoneme-sized (aka uniphone) segments; the use of diphone or larger units would mask some of the improvements possible with an airflow-guided system.

4. Airflow Database Labeling

To label units in the clunits speech database using airflow data, both the acoustic and airflow databases were force-aligned to the TIMIT prompts to produce segment-level labels with a 5-state left-right HMM with no skips and a 'silences' model allowing self-loops (Young *et al.* 2009). Many segmentation problems in both databases were hand-corrected.

5. Stimulus Generation

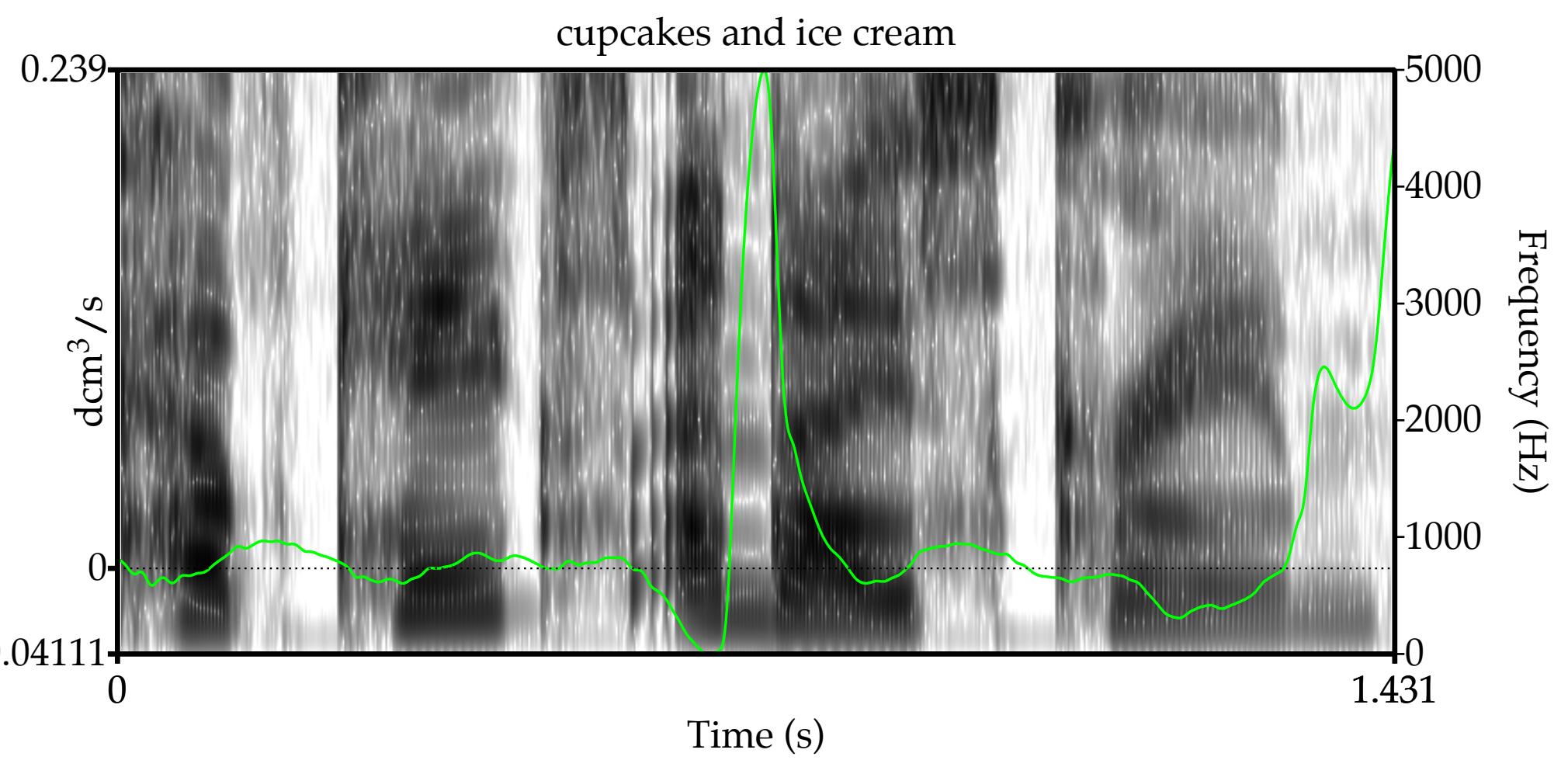
50 words containing nasals that *were not* present in the TIMIT prompts were synthesized using the reference system. An independent listener chose the 25 most natural of these utterances. Finally, airflow-guided versions were synthesized by re-ranking units for the vowel targets to minimize first differences in the raw nasal airflow traces between candidate vowel units and the Festival-selected consonant units. Consonants were held constant across both airflow-guided and reference stimuli. A final list of required units (including re-ranked vowels) was synthesized using a modified version of the Festival software.


reference system unit selection
airflow system unit selection
æ̤
æ
æ̃
optimal acoustic alignment
pʰ
æ
n
t
æ̃
optimal airflow alignment

6. Assessment: Listening Experiment

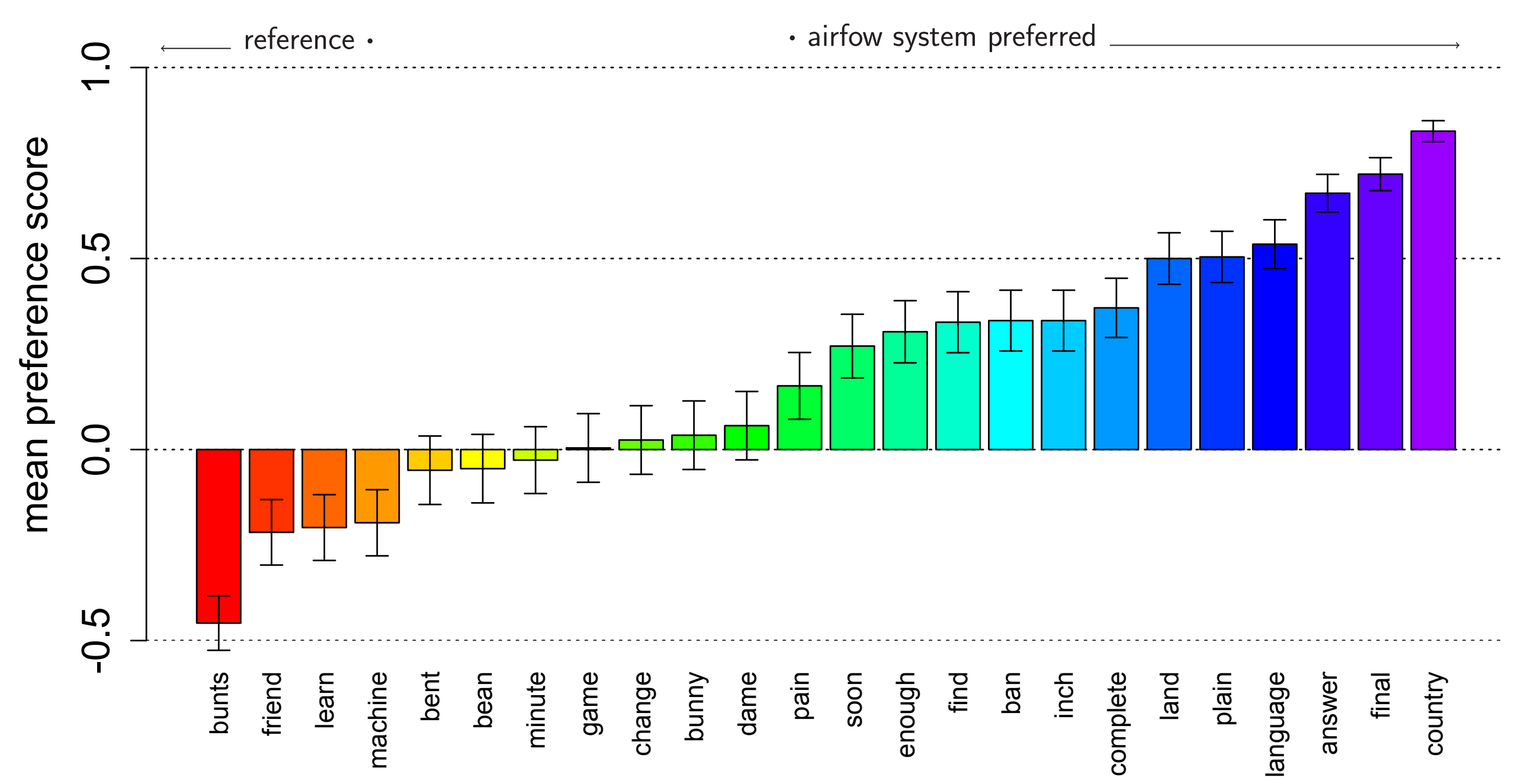
Synthesis quality was assessed via listening experiment. 20 undergraduates at the University of Michigan saw the text of each stimulus 24 times, in random order while hearing both the airflow-guided and reference stimuli. Participants were instructed to indicate via response pad whether the first or second utterance sounded "more natural". Presentations were balanced for first/second order. One stimulus pair (*against*) was withheld for use as a practice item.

Sample Nasal Airflow Tracing (RMS)

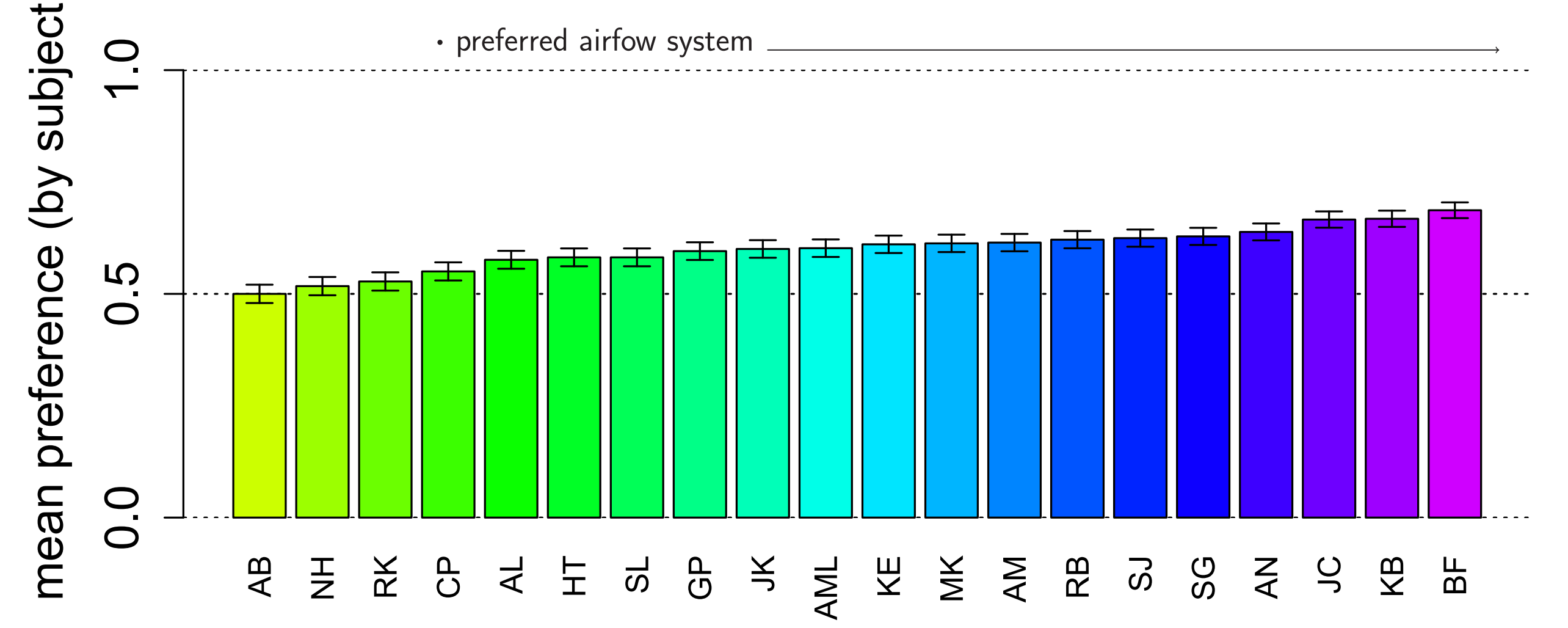


A spectrogram of the words "cupcakes and ice cream" extracted from one of the TIMIT prompts and aligned with RMS nasal airflow data.

Results: Listener Judgments by Item

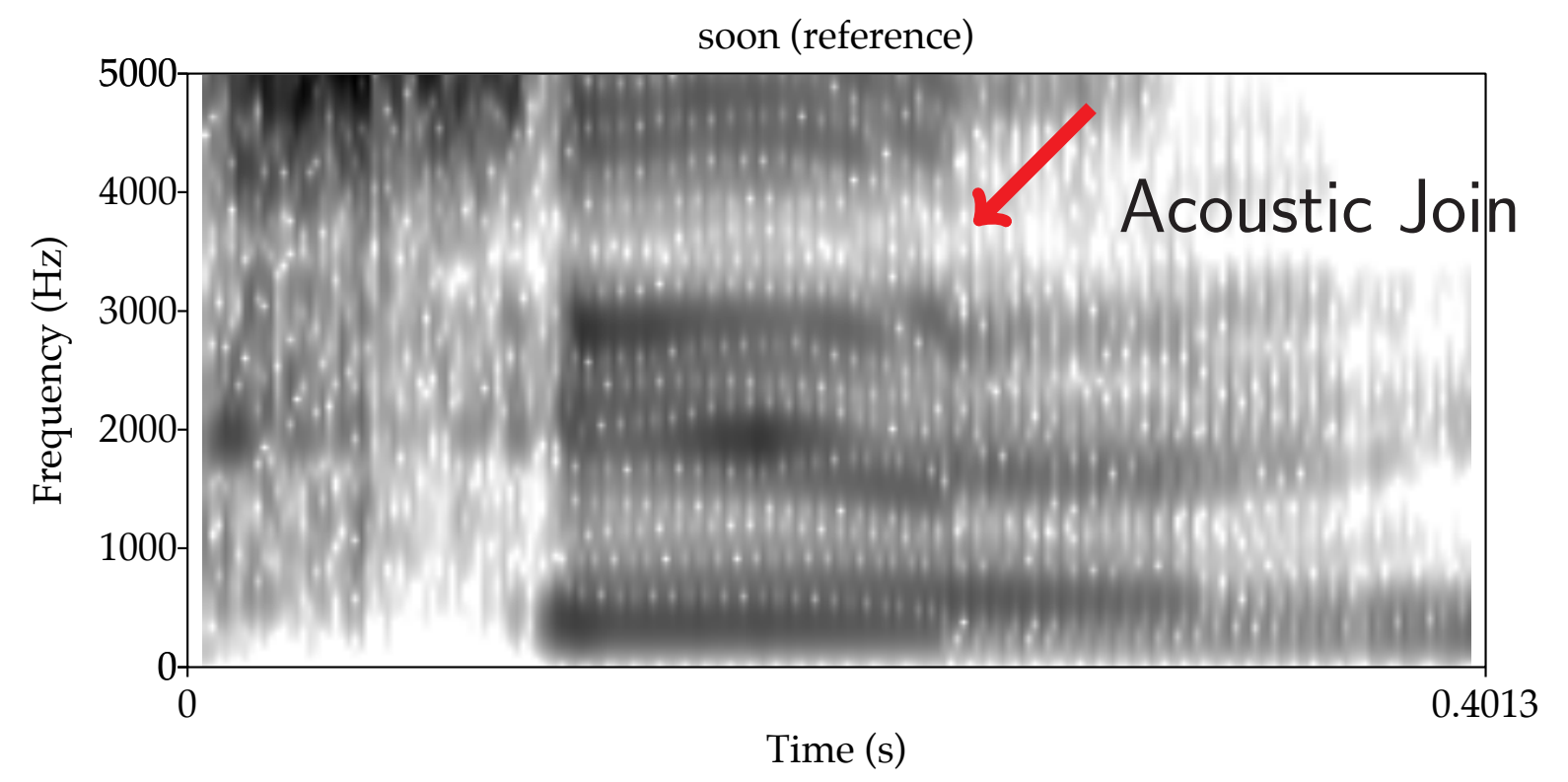


Results: Listener Judgments

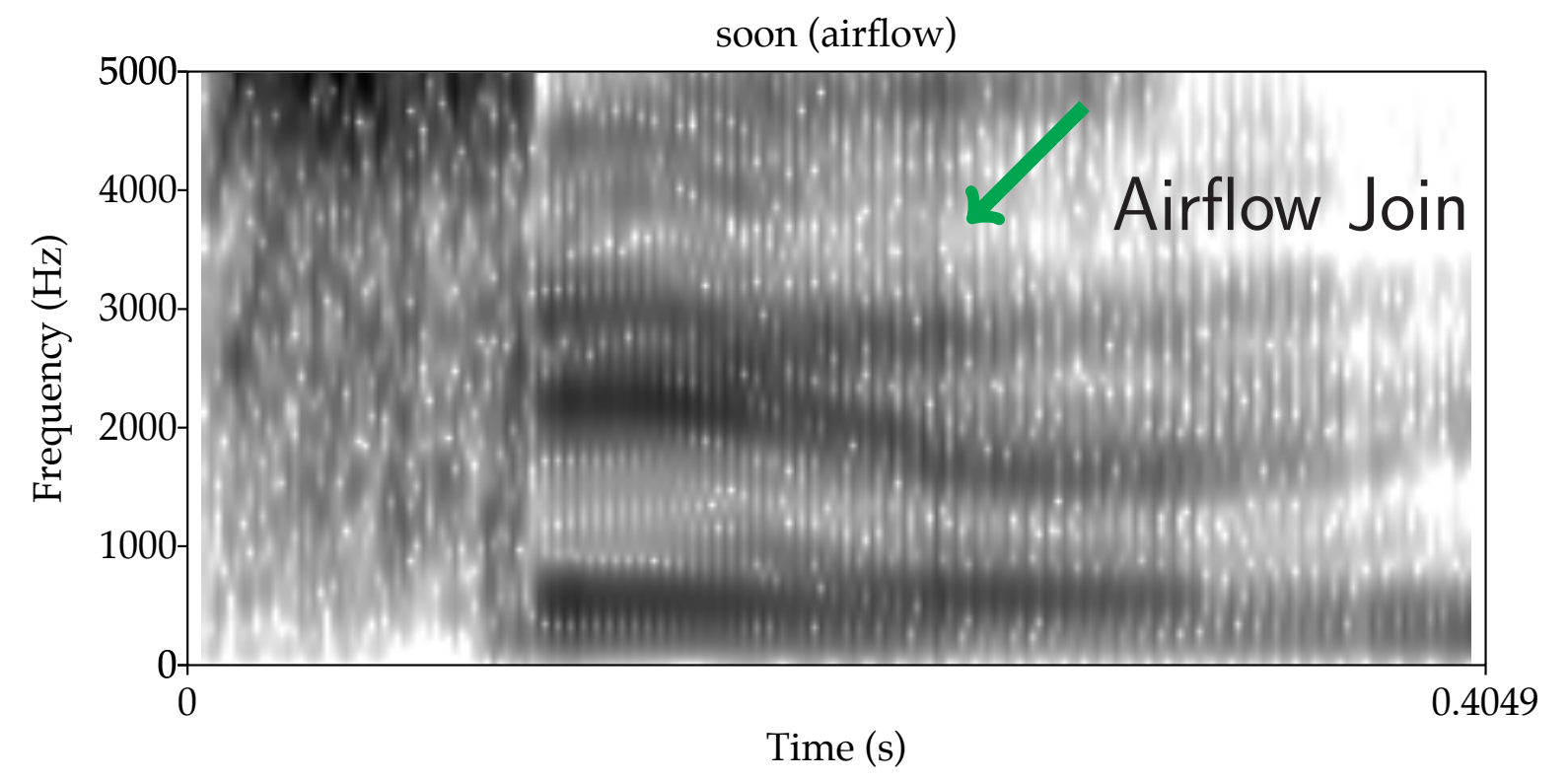


Example Stimuli: Spectrograms of soon

soon: reference system



soon: airflow re-ranked vowel



Discussion

Baseline: Is presence of accurate coarticulatory information perceptually useful in synthesized speech?

Yes. Though participants performed at chance levels for 7 of the 24 items according to a logistic regression model, they performed significantly differently from chance on the other 17. A repeated measures analysis of variance on the response data from the synthesis comparison confirms what the plot above implies.

Do listeners prefer this more accurate synthesis?

On average, **yes** they do. 17 of the 20 participants had mean preference scores significantly above chance (preferring the airflow-guided synthesis). Also, the mean response (on a scale of 0 to 1) was 0.6. Participants preferred the airflow re-ranked utterance in 13 of the 24 pairs (there is a significant main effect for each word at the .01 or .001 level for every word *except* bent, bean, minute (first syllable), game, change, bunny and dame (participants performed at chance on these items). Interestingly, in all but one of these items for which the manipulation was ineffectual, the manipulated unit was a mid or high front vowel. Unfortunately, the stimuli were not well balanced for vowel quality; additional work is required on this point.

Is airflow a useful and efficient means of automatically labeling a speech synthesis database with fine-grained coarticulatory detail?

In this experiment it absolutely was not, but it easily could have been. Most of the difficulty with this project (and, very possibly, several of the poorer results) relate to having to record the speech database twice, segment it twice, and assign airflow values from one database to acoustic units in a parallel database. All of these difficulties were due to the use of the oral airflow mask (which renders the audio recordings unusable for synthesis). As the oral airflow data turned out to be unnecessary for the reranking, we could easily have recorded the TIMIT prompts only once while collecting both nasal airflow data and a clean, usable acoustic signal.

Conclusions

The collection of nasal airflow data with a pneumotachograph is easy, non-invasive, the head is free to move and the subject is relatively comfortable (particularly when compared with electromagnetic articulography, velotrace or even the relatively non-invasive ultrasound). Without an oral airflow mask to perturb the audio signal, the method reported here has great promise as an efficient and useful means of modeling coarticulation for concatenative speech synthesis.

The data reported here are consistent with the position that the ultimate goal of speech synthesis should be to maximize similarity to an idealized *percept* and not, as seems to be the general understanding, to maximize similarity to an idealized utterance. Finally, these findings appear to support the position that coarticulation is useful signal for listeners and not, as many contend, mere distortion of the speech stream.

Selected References

Black, Alan W, & Paul Taylor. 1997. Automatically clustering similar units for unit selection in speech synthesis. In *Eurospeech*, 601–604.

Fisher, William M., George R. Doddington, & Kathleen M. Goudie-Marshall. 1986. The DARPA speech recognition research database: Specifications and status. In *Proceedings of DARPA Workshop on Speech Recognition*, 93–99.

Fowler, Carol A. 1996. Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America* 99.1730–1741.

Ladefoged, Peter. 2001. *A Course in Phonetics*. New York: Harcourt Brace Jovanovich, Inc., fourth edition.

Lindblom, B. 1990. Explaining phonetic variation: a sketch of the H&H theory. In *Speech production and speech modeling*, ed. by W. Hardcastle & A. Marchal. Dordrecht: Kluwer.

Stevens, K. N., & S. J. Keyser. 2008. Quantal theory, enhancement and overlap. *Journal of Phonetics* . in press.

Whalen, Doug H. 1984. Subcategorical phonetic mismatches slow phonetic judgments. *Perception and Psychophysics* 35.49–64.

Young, S., D. Kershaw, J. Odell, D. Ollason, V. Valtchev, & P. Woodland. 2009. *The HTK Book Version 3.4*. Cambridge University Press.

Acknowledgements

I would like to thank my advisors Pam Beddor & Steve Abney for their assistance with this project. Nickolay V. Shmyrev and Alan K. Black of the Festival open source project were extremely helpful; Ivan A. Uemlianin's SpeechCluster tools were indispensable; as, of course, was Paul Boersma's Praat. Finally, I'd like to thank the Phonetics/Phonology and Computational Linguistics discussion groups at the University of Michigan —particularly Anthony Brasher, Susan Lin and Terrence Szymanski— for their copious feedback and insight.

Department of Linguistics, University of Michigan, Ann Arbor, MI
e-mail: clunis@umich.edu
http://umich.edu/~clunis/