2pSC7

Aerodynamic Modeling of Coarticulation for Concatenative Speech Synthesis

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

Abstract

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. For example, the vowel in *can* includes a nasalization feature (from the final nasal consonant) not present in the word *cat*. We believe details like this provide the listener with a rich network of informative cues and are key to understanding our astonishing ability to disambiguate meaningful speech sounds from a seemingly infinite range of noisy inputs. Unfortunately, these cues, whether subtle or overt, are generally missing or contradictory in text to speech (TTS) synthesis output.

We present a method of improving concatenative speech synthesis by explicitly modeling coarticulation. The Festival speech synthesis system (Taylor *et al.* 1998) was modified to use airflow data during unit selection. The output of this modified system and the unmodified system were compared in a listening experiment. Results indicate not only that listeners are sensitive to the sub-categorical phonetic differences but that, in general, they prefer speech synthesized from a hybrid acoustic/articulatory model to standard acoustic-only speech synthesis.

Background: Coarticulation

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

- Coarticulation is not ideal for the listener because it introduces variation (Ladefoged 2001, Tatham & Moreton 2006) or because it renders contrasts less distinct (Lindblom 1990).
- Sometimes instances of coarticulation overlap and obliterate featural cues but 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).

Coarticulatory cues influence listeners' decisions:

> *"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*

- *Identification*: Vowel nasalization alone, with no N, elicits N percepts in American English, suggesting that listeners are, indeed, sensitive to vowel nasalization (Beddor *et al.* 2007).
- *Reaction time*: Absence of anticipatory vowel nasalization slows listeners' reaction times in identifying N vs. C (Fowler & Brown 2000).

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 cost of aligning a particular unit with the desired speech output (target cost) and the cost of adjoining the next sound to the most-recently selected unit (join cost):

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

- where t is the target phone in the sequence and
- u represents the unit of sound to be appended.
- These costs are calculated using primarily acoustic information
- Specifically, a distance metric based on MEL frequency cepstral coefficients (MFCCs).

There are two assumptions implicit in a model that generates speech by minimizing acoustic distances between segments:

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.

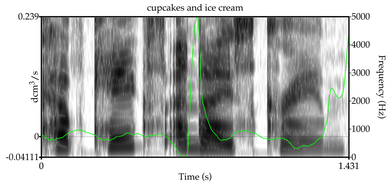
The history of concatenative speech synthesis has, in many respects, been shaped by attempts to deal with coarticulation given these assumptions. Early models attempted to eliminate coarticulation by recording only carefully articulated diphones in heavily controlled articulatory contexts (leading to interpretable but unconvincing speech). More recent systems have moved toward implementing coarticulation indirectly by recording enormous databases of speech and synthesizing utterances by preferring units that were naturally contiguous. **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 presence of 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?

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.

Method

1. 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.

3. Reference Voice Creation

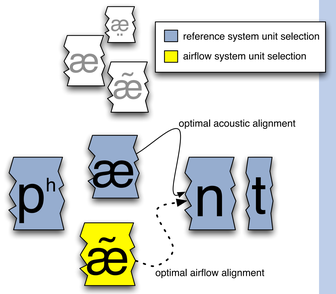
The acoustic recordings were used to create a clunits voice (Black & Taylor 1997) using the Festival open source speech synthesis system. Clunits was chosen both for its conceptual simplicity and its use of phoneme-sized (aka uniphone) segments; the use of diphone or larger units would mask some of the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

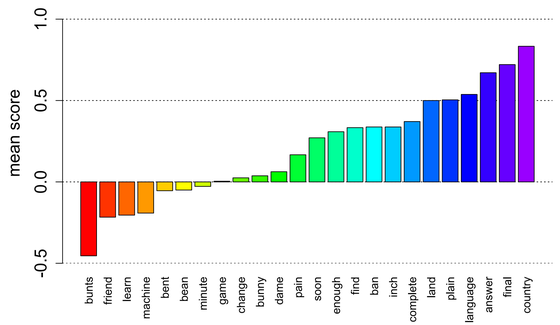
50 words containing nasals that *were not* present in the TIMIT prompts were synthesized using the reference system. An independent listener chose the 25 most natural of these utterances. Finally, airflow-guided versions were synthesized by re-ranking units for the vowel targets to minimize first differences 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.



6. Assessment

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.

Results: Listener Judgments



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, 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. There is a significant main effect for word ($F(23,11503) = 61.608$, $p<0.001$).

Do listeners *prefer* this more accurate synthesis?

On average, **yes** they do. 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, 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. We know that nasalization is a less salient cue for these vowels and more work will be required to see if there really is an affect of vowel quality.

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.

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.

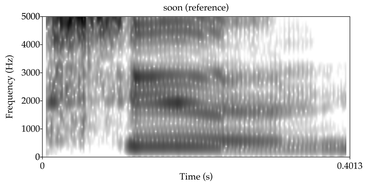
Conclusions

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.

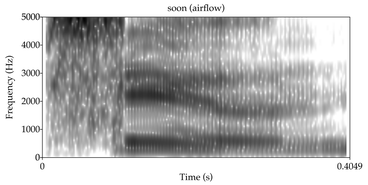
Our findings appear to support the position that coarticulation is useful signal for listeners and not, as many contend, mere distortion of the speech stream.

Stimuli: Spectrograms of soon

soon: reference system



soon: airflow re-ranking



References

Beddor, Patrice Speeter, Anthony Brasher, & Chandan Narayan. 2007. Applying perceptual methods to the study of phonetic variation and sound change. In *Experimental Approaches to Phonology*, ed. by Maria-Josep Sole, Patrice Beddor, & Manjari Ohala, 127–143. Oxford University Press. In Honor of John Ohala.

Black, Alan W, & Paul Taylor. 1997. Automatically clustering similar units for unit selection in speech synthesis. In *in Eurospeech97*, 601–604.

Dahan, D., J. S. Magnuson, M. K. Tanenhaus, & E. M. Hogan. 2001. Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language & Cognitive Processes* 16.507–534.

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, C. A., & J. M. Brown. 2000. Perceptual parsing of acoustic consequences of velum lowering from information for vowels. *Percept and Psychophysics* 62.21–32.

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.

Tatham, M., & K. Moreton. 2006. *Speech Production and Perception*.. Palgrave.

Taylor, P., A. Black, & R. Caley, 1998. The architecture of the the festival speech synthesis system.

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 Stephen P. Abney & Patrice Speeter Beddor for their assistance with this project. Nickolay V. Shmyrev and Alan K. Black of the Festival open source project were helpful at various points during system development; Ivan A. Uemlianin's SpeechCluster tools were indispensable during this project; as 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 frequent feedback and insight.

e-mail: clunis@umich.edu