INTERSPEECH 2009 BRIGHTON

Automatically Rating Pronunciation Through Articulatory Phonology

Joseph Tepperman[1], *Louis Goldstein*[2], *Sungbok Lee*[1], *and Shrikanth Narayanan*[1,2]

[1] Signal Analysis and Interpretation Laboratory, University of Southern California, USA

[2] Department of Linguistics, University of Southern California, USA

{tepperma@,louisgol@,sungbokl@,shri@sipi.}usc.edu

10.21437/Interspeech.2009-708

Abstract

Articulatory Phonology's link between cognitive speech planning and the physical realizations of vocal tract constrictions has implications for speech acoustic and duration modeling that should be useful in assigning subjective ratings of pronunciation quality to nonnative speech. In this work, we compare traditional phoneme models used in automatic speech recognition to similar models for articulatory gestural pattern vectors, each with associated duration models. What we find is that, on the CDT corpus, gestural models outperform the phoneme-level baseline in terms of correlation with listener ratings, and in combination phoneme and gestural models outperform either one alone. This also validates previous findings with a similar (but not gesture-based) pseudo-articulatory representation.

Index Terms: pronunciation modeling, nonnative speech, articulatory phonology

1. Introduction

In our past work in [8], we showed that an articulatory representation of speech could be used, along with a more traditional phonetic representation, to automatically discriminate between close phoneme-level errors made by nonnative speakers of English, and also between perceptually confusable English phonemes produced by native speakers - the addition of these pseudo-articulatory models resulted in better performance than with phonetic models alone. Since we did not work with any real articulatory data (e.g., derived from magnetometry or magnetic resonance imaging), we proposed a new mapping from phonemes to articulatory features based on [3, 7], and it was clearly this additional articulatory representation (though artificial) that accounted for the improvement in classification. The use of several streams of Hidden Markov Models representing articulatory configurations seemed especially suited for the task of discriminating between close phonemes, since that representation had more explanatory power than the traditional notion of speech as a sequence of segments - a close error in nonnative production could be represented as a subtle dynamic change across some subset of eight overlapping articulatory streams, rather than as a full-on substitution of an entire phoneme. This also had implications for second-language instruction in that these pseudo-articulatory models could potentially be used to provide feedback to students in concrete anatomical terms - i.e., they could be taught to sound more like a native speaker through specific physical changes to their vocal tract configuration.

However, there were a few obvious limitations to that method. Most importantly, with a set of Hidden Markov Models for articulatory acoustics, we didn't explicitly model the characteristic variants in duration that would probably have improved classification accuracy. Our method of deriving expected articulatory sequences from phonemes, though it accounted for context and was grounded in phonological theory, assumed very rigid and unrealistically quantized positions of the vocal tract constituents. Finally, the eight derived articulatory streams were assumed to move independently and were decoded as such - this was for the sake of computational simplicity, and didn't reflect the anatomical dependencies among organs (e.g., jaw and tongue) nor the phonological synchronization of these organs when producing linguistic units like phonemes.

For these reasons, a reformulation of this problem from an Articulatory Phonological point-of-view seemed in order. The theory of Articulatory Phonology [1] proposes a set of overlapping articulatory gestures as the fundamental units of speech that, through coordinated action, give rise to larger units such as phonemes and syllables. These gestures specify only a constriction location and a degree of constriction for each vocal tract variable (e.g., tongue tip, tongue body) - beyond that they allow for a degree of physical abstraction that suits acoustic-to-articulatory inversion without true articulatory data, as in this task. As implemented in the Task Dynamics Application (TaDA) model [5], each gesture is associated with an underlying timing oscillator that controls the activation and de-activation of an ensemble of gestures, and the phase relations among these oscillators constitute a model of coordinated planning of the tract variables' constriction tasks. From these hypothesized inter-gestural coupling relations, Articulatory Phonology provides some notion of the durational and coarticulatory effects that are characteristic of native speech - this was one component missing from the previous work.

With this new approach we do not intend to replicate our previous work in segment-level error detection, but rather to estimate a word-level rating of pronunciation quality for nonnative speech. Here a pronunciation rating is defined as a real number that is proportional to the subjective degree of nativeness in a speaker's English production. We propose using Bayesian Inference on a continuous Gaussian hidden variable to estimate this rating. The abstract perceptual property we call "nativeness" manifests itself on many simultaneous levels, and we expect that the insights of Articulatory Phonology will allow for automatic ratings better correlated with listener perception than similar ratings derived from phoneme-level acoustic and duration models like those outlined in [6].

2. Articulatory Phonology: Background

The primitive phonological unit in Articulatory Phonology is the articulatory gesture, defined as the dynamic constriction action of a distinct vocal tract organ (or *tract variable*) [1]. These gestures are uniquely specified by constriction degree/location

Table 1: Amount of data used in this study.

native language	*speakers*	*hours*	*words*
English	39	2.54	4569
Arabic	18	1.25	2240

pairs for the lips (LA/LP), tongue tip (TTCD/TTCL), and tongue body (TBCD/TBCL), and by constriction degree descriptions alone for those tract variables that cannot change location: the glottis (GLO) and velum (VEL). For an input sequence of phonemes, the implementation in TaDA [5] will generate the corresponding *gestural score* where "score" is used in the same sense as that of a musical transcription - it is an expected constellation of gesture activations across tract variables, derived from a phoneme-to-constriction mapping, with allophonic exceptions based on position in the syllable.

Gestures can overlap within one tract variable or among several tract variables, signifying the presence of multiple simultaneous constriction efforts. The expected overlap between any two gestures is computed by a coupled oscillator model of inter-gestural coordination. Each gesture is activated and deactivated by an oscillator equation with parameters defined by the tract variable, the type of constriction involved, and the syllabic structure; gestures in a constriction degree/location pair share the same oscillator. Phasal relations among these oscillators determine the sequence in which their gestures are activated and hence their overlap in the gestural score. Currently TaDA only allows for 0, 90, 180, and 360 degrees of relative phasing between oscillators. The coupling of these oscillators is the way in which TaDA models the coordination of gestures into larger linguistic units such as phonemes and syllables.

Though Articulatory Phonology was first formally proposed in 1986, it hasn't been until very recently, with the widespread interest in articulatory modeling of speech, that speech engineers have started to investigate the potential of the articulatory gesture as a viable unit for automatic speech recognition. Two studies have worked with synthetic time-varying physical realizations of gestural scores as generated by TaDA's task dynamic model of inter-articulator coordination. One has automatically estimated these vocal tract time functions based on the corresponding synthetic acoustics [4], and the other has demonstrated automatic estimation of the gestural score based on the synthetic tract variable time functions [10]. As of yet there has been no published work in decoding a gestural score from real speech acoustics, but the idea of inferring articulatory behavior from real acoustics is not new. Studies such as [2] have done this not with gestures but with an articulatory feature-based representation in which each speech frame is described by the discrete values of an ensemble of articulatory *features* (e.g., manner, place, rounding, nasalization, etc.) rather than through Articulatory Phonology's overlapping *gestures*. The main advantages of Articulatory Phonology over articulatory feature-based models are in its level of physical abstraction, its patterns of coordinated temporal overlap among tract variables, and its capacity to connect cognitive aspects of speech planning with the physical realizations of those planned constrictions.

In assigning ratings to nonnative speech, the expected gestural score generated by TaDA is assumed to be that of a native speaker, and so it can serve as a reference against which all incoming test utterances are compared. Dynamic changes in gestural overlap correspond to acoustic changes in real speech, and the couplings among gestures denote relative activations that are characteristic of native-like coarticulatory timing - these relative gestural onsets can be represented in duration models. Here we follow [10] in using the *gestural pattern vector* as the unit for encoding the gestural score in acoustic model form. The gestural pattern vector is the pattern of activation across all variables in a gestural score, at any one instant. Encoding our models this way ensured that synchronous gestures (those sharing an oscillator, or with a 0 degree phase relation) would be decoded simultaneously, an improvement over our previous work in [8] in which the articulatory streams were decoded independently. Nonnative speech is hence rated along two time-varying dimensions: the degree to which the acoustics match a gestural pattern vector's acoustic model trained on native speech, and the degree to which the differences in onsets of the coupled gestures fall within native-like distributions.

Table 2: The number of acoustic and duration models required for each of the words in the CDT vocabulary.

word	*phonemes*	*gestural vectors*	*coupling pairs*
believe	6	13	6
chin	3	7	3
drag	4	10	5
forgetful	9	19	12
go	3	6	3
paper	5	11	6
racetrack	8	16	7
then	3	6	3
thing	3	6	3
typical	7	21	11
understand	9	18	12
wood	3	8	3

3. Speech Corpus

All speech data used in this study come from the CDT corpus. This consists of a small vocabulary of English words (12 total) spoken in isolation by 24 native speakers of Arabic and 39 native speakers of English. The words were designed to elicit varying degrees of English proficiency, some with intentionally difficult consonant clusters (e.g., "racetrack"). Each word was repeated roughly 10 times per speaker (in a random order), and integer pronunciation ratings on a 1 to 7 scale were elicited from 8 native English listeners for all tokens from 18 of the Arabic speakers. The average inter-listener correlation over all nonnative tokens was 0.470, and the average correlation between any listener and the median of the other listeners' ratings was 0.627 - these medians were then taken as the reference set of ratings against which to measure automatic performance. This relatively poor agreement reflects the brevity of the one-word tokens and their scant evidence on which to base subjective scores. Statistics for the size of this data set are given in Table 1, while the number of phonemes, gestures, and gestural couplings for the words in this corpus are given in Table 2.

4. Pronunciation Modeling

This section explains the phoneme and gesture models for native English pronunciation used in this study. These consist of acoustic and duration models, as well as pronunciation rating models in which all acoustic and duration measures are combined to synthesize an overall word-level pronunciation rating.

4.1. Acoustic Models

All acoustic models in this study - whether for phonemes or gestural pattern vectors - were designed as Hidden Markov Models trained on 39-dimensional MFCC feature vectors, with 3 hidden states and 32 Gaussian mixtures per state. The window length was standard (25 msec) and the frame rate was shorter than usual (5 msec) so as to capture very fine changes in gestural overlap. An expected sequence of phonemes for each word came from TaDA's lexicon. Similarly, an expected sequence of gestural pattern vectors was derived from TaDA's gestural score specification for each phoneme sequence (as explained in Section 2): the sequence consisted of a concatenation of all regions in the gestural score for which the gestural pattern vector did not change. Vectors resulting from a gestural overlap less than 2 frames long (at TaDA's default 10 msec frame rate) were discarded from the sequence. Similarly, any intra-variable overlap between a release gesture and the gesture immediately preceding it was ignored. See Figure 1 for an illustration of deriving the sequence of vectors from the gestural score. Among the twelve words in our task vocabulary, there were a total of 28 unique phonemes and 115 unique gestural pattern vectors, 96 of which only appeared in one word - 15 appeared in two words, 1 appeared in three words, 2 appeared in four words, and 1 appeared in five of the twelve words. Note that the gestural score does not simply specify more linguistic units per word (see Table 2) but maps those units to the words with a different distribution than that of phonemes. For example, though "then" and "thing" share no common phonemes, they have two gestural pattern vectors in common due to their similar articulations.

The CDT corpus has no transcribed segmentations on the phoneme level (and certainly not on the gestural pattern vector level) so all acoustic models were trained using an iterative bootstrap procedure like that described in [9]. After Viterbi decoding of the expected sequence using the trained models, a phoneme-level or gestural vector-level pronunciation quality measure based on these models was defined for phoneme or gesture n as the log-likelihood ratio

$$A_n = log\,[P(O_n|M_n)/P(O_n|f_i)] \quad (1)$$

where $P(O_n|M_n)$ is the likelihood of the speech observation given the target segment's HMM, and $P(O_n|f_i)$ is the likelihood of the same observed speech given a generic filler HMM. Two filler models, f_p and f_g, were trained - one for phonemes and one for gestural pattern vectors - based on all training data.

4.2. Duration Models

From the automatic segmentation determined with acoustic models as described in Section 4.1, we could extract duration information from relevant segments. Two different approaches were investigated: one, D_{seg}, in which the durations of all segments (whether they represent phonemes or gestural pattern vectors) were measured relative to native distributions of durations; the second, D_{coupl}, in which only the delay in activation times between all pairs of coupled gestures was compared to native distributions (this applied only when using gestural acoustic models, of course).

Following the method used in [6], the measure of nativeness for segment duration or activation delay n was formally defined as $D_n = P(f(d_n)|M_n)$ where d_n is the n^{th} duration or delay in the sequence, $f(\cdot)$ is the duration normalization function, and M_n is n^{th} segment or activation delay. The probability of the duration was modeled as a Gaussian distribution estimated from all native speech for that particular word. Duration normalization was computed using $f(d_n) = d_n \cdot ROS_w$ where ROS_w is the rate of speaking (in phonemes or gestural vectors per second) for word w.

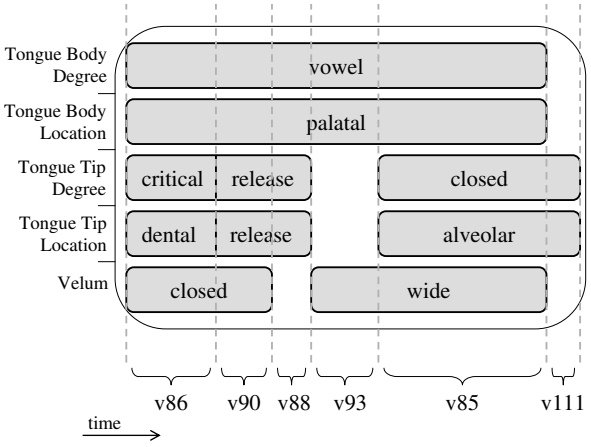

Figure 1: Gestural score for the word "then." The sequence of gestural pattern vectors is shown along the bottom, with each vector assigned an arbitrary number.

4.3. Pronunciation Rating Models

To combine acoustic and duration measures in estimating an overall score, we used a Naive Bayes framework. The hidden node, Q_w, representing the 1 to 7 subjective nativeness score for word w, was modeled as the generative parent of two feature nodes: $A_w = \{A_1, \ldots, A_N\}$ for N acoustic measures, and $D_w = \{D_1, \ldots, D_N\}$ for N duration measures. All three nodes were modeled as linear Gaussian distributions with diagonal covariances. The inferred value of Q_w was calculated as the mean of the marginal distribution of Q_w given the features: $\hat{Q}_w = E[P(Q_w|A_w, D_w)]$. The cardinalities of A_w and D_w were word-dependent, as each word had a unique number of phonemes, gestural pattern vectors, and coupled gestures (as outlined in Table 2); consequently, a unique network had to be trained for each word (using only nonnative examples, to reflect a range of ratings), though the overall structure was identical across all words.

5. Experiments

Experiments were designed to address the following:

- How do phoneme and gestural models compare to each other and to a baseline pronunciation rating?
- For gestural models, were D_{coupl} duration measures (activation delay measures based on hypothesized intergestural coupling) better than gestural D_{seg} measures that were coupling-blind?
- How did phoneme and gestural models perform in combination? Was it better than with phonemes alone?

In all cases, performance was evaluated in terms of correlation with the median of the 8 listeners' scores. The baseline word-level rating was taken as $\frac{1}{N}\sum_{n=1}^{N} A_n$, i.e., the mean of all the phoneme-level acoustic measures in a word, without using Bayesian inference, following [6]. Both phoneme and ges-

Table 3: Correlation coefficients between automatic and median listener ratings. Entries in bold were significantly better than the baseline with $p \leq 0.05$

			Phoneme acoustic models		*Gesture acoustic models*				*listener*
	N	*baseline*	*alone*	*w/* D_{seg}	*alone*	*w/* D_{seg}	*w/* D_{coupl}	*Combined*	*agreement*
all speakers	6809	0.510	0.500	0.517	**0.593**	**0.591**	**0.582**	**0.674**	0.801
Arabic only	2240	0.428	0.463	**0.492**	**0.510**	**0.513**	**0.502**	**0.547**	0.627

tural vector acoustic models were compared, the latter with either coupling duration measures or segment duration measures. All data were trained and tested using a leave-one-speaker-out crossvalidation procedure. Only native English speech was used for training the acoustic and duration models, and only rated Arabic-speaker data was used to train the pronunciation rating model. When combining phoneme and gestural information, the acoustic measures A_w included both phoneme and gestural pattern vector measures, and the duration measures D_w were comprised of both phoneme D_{seg} duration measures and gestural D_{coupl} measures (but not gestural D_{seg} measures).

Correlation results for different models and feature sets are reported in Table 3. The overall automatic scores for each row were simply the concatenation of the automatic scores from each word's individual pronunciation rating model. Entries in bold were significantly better than the baseline with $p \leq 0.05$ using a z-test for the difference in correlation coefficients. Listener agreement with the median of the other 7 listeners is provided as an upper bound on automatic performance. In the "all speakers" case, the native English speakers were artificially assigned the highest pronunciation score, 7.

6. Discussion

According to Table 3, we see a number of general trends. First, improvements over the baseline are in general only ever achieved either through the use of gestural models, or the combination of phoneme and gestural models - phoneme models were not enough to achieve a significant improvement without duration measures. Using the D_{coupl} duration measures was not significantly better or worse than using the D_{seg} measures with the gestural models, and neither one was statistically worse than just using the gesture acoustic measures alone, with no duration measures. In both populations, the best result came from combining both phoneme and gestural acoustic measures along with both phoneme D_{seg} and gesture D_{coupl} duration measures.

All results fall significantly below the inter-listener agreement upper bound. However, the agreement for "all speakers" is artificially inflated since it includes many of the native English examples for which all listeners' scores were assumed to be 7 (and so there could be no disagreement). Because the performance trend is similar in both populations, this indicates that the scoring model is capable both of assigning an appropriate range of scores to nonnative speech, and of giving high scores to native speech. In a word-dependent analysis of the results, the words with the poorest performance overall - "go" and "then" - were also the ones with the lowest inter-listener agreement, and were also among the shortest words in the set and therefore the most difficult to judge due to a dearth of evidence on which to base a rating. However, other short words like "wood" and "thing" did not show this effect, and so perhaps some of the words were more difficult for Arabic speakers to pronounce, or for English-speaking listeners to rate.

7. Conclusion

We have shown the usefulness of the articulatory gesture as a unit for automatically assigning subjective ratings to both nonnative and native English. With gestural acoustic and duration models, we demonstrated improvements both separately and in combination with phoneme models, as in our previous work using pseudo-articulatory features [8]. Gestural vectors do have the disadvantage of demanding many more acoustic models than phonemes; over a larger vocabulary, the number of unique vector models to train might become intractable. Future work is needed to properly incorporate Articulatory Phonology's intergestural coupling as a cue to nativeness - no improvement was seen through its addition in this case. These findings still suggest that Articulatory Phonology is a promising avenue not just for pronunciation rating, but for speech recognition as well.

8. Acknowledgments

This work was supported by NSF grant II-0703048.

9. References

[1] C. P. Browman and L. Goldstein, "Articulatory Phonology: An Overview," in *Phonetica*, 49:155-180, 1992.

[2] K. Hacioglu, B. Pellom, and W. Ward, "Parsing speech into articulatory events," in *Proc. ICASSP04*, Montreal, 2004.

[3] P. Ladefoged, *A Course in Phonetics*. 5th Edition. Boston: Thomson, 2006.

[4] V. Mitra, I. Y. Ozbek, H. Nam, X. Zhou, and C. Y. Espy-Wilson, "From Acoustics to Vocal Tract Time Functions," in *Proc. of ICASSP*, Taipei, 2009.

[5] H. Nam, L. Goldstein, E. Saltzman, and D. Byrd, "TaDA: An enhanced, portable task dynamics model in MATLAB," *J. Acoust. Soc. Amer.*, 115(5,2):2430, 2004.

[6] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic Scoring of Pronunciation Quality," *Speech Communication*, 30(2-3):83-94,1999.

[7] M. Richardson, J. Bilmes, and C. Diorio, Hidden-Articulator Markov Models for speech recognition, *Speech Communications*, vol. 41, no. 2, October 2003.

[8] J. Tepperman and S. Narayanan, "Using articulatory representations to detect segmental errors in nonnative pronunciation," in *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):8-22, Jan. 2008.

[9] S. Young et al. *The HTK Book.* [Online]. Available: http://htk.eng.cam.ac.uk/, 2002.

[10] X. Zhuang, H. Nam, M. Hasegawa-Johnson, L. Goldstein, and E. Saltzman, "The Entropy of the Articulatory Phonological Code: Recognizing Gestures from Tract Variables," in *Proc. of Interspeech*, Brisbane, 2008.