HIDDEN-ARTICULATOR MARKOV MODELS FOR PRONUNCIATION EVALUATION

Joseph Tepperman and Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory
Viterbi School of Engineering
University of Southern California
http://sail.usc.edu/
tepperma@usc.edu, shri@sipi.usc.edu

ABSTRACT

The design of a robust language-learning system, intended to help students practice a foreign language along with a machine tutor, must provide for localization of common pronunciation errors. This paper presents a new technique for unsupervised detection of phone-level mispronunciations, created with language-learning applications in mind. Our method uses multiple Hidden-Articulator Markov Models to asynchronously classify acoustic events in various articulatory domains. It requires no human input besides a pronunciation dictionary for all words in the end system's vocabulary, and has been shown to perform as well as a human tutor would, given the same task. For the majority of systematic mispronunciations investigated in this study, precision in detecting the presence of an error exceeded the 70% inter-annotator agreement reported by our test corpus.

1. INTRODUCTION

Correct phone-level pronunciation is crucial for effective communication in any language, and therefore should be of high significance to students of a foreign language. Systematic phone insertion, deletion, or substitution can potentially alter the perceived meaning of what was spoken, or at the very least hinder a speaker's intelligibility in her new tongue. If such a student is to make any use of a language learning system designed to automatically assess her pronunciation (like the ones proposed in [9] and [12]), then that system must be capable of identifying and correcting pronunciation mistakes on the phone level at least as well as a human tutor would.

Oftentimes an easily predictable or commonly recurring mispronunciation may be a function of the phonological rules and phonetic structure of the speaker's native language [1]. For example, native speakers of German learning English as a second language are likely to substitute /s/ for /z/ in such words as "dessert" or "warnings," because in German the character 's' is often pronounced unvoiced as the phoneme /s/ in these contexts. In the beads-on-a-string model of speech, this particular substitution is equivalent to any other, even one that would not normally occur in English or German. However, with a model of speech as a time-series of articulatory events – asynchronous jaw, tongue, lip, velum, and vocal cord movements – both /s/ and /z/ represent very similar physical configurations, in fact so similar that their only distinction is that /z/ is voiced and /s/ is not [10,13]. For purposes of evaluating non-native speech, articulatory feature models would seem ideal for localizing this type of close error in the physiological domain and providing useful feedback to a second-language student in the realm of speech production. And prior knowledge of these expected, recurring mispronunciations can make a language-learning tool more effective in targeting the most important and difficult errors a non-native speaker might make.

Articulatory models of speech have been used to improve accuracy in such tasks as speaker verification [11], general ASR [13], and spectrally-impoverished or whispery speech recognition [8,10]. Typically, automatic detection or correction of pronunciation errors on the phone level relies on acoustic or prosodic analysis within a previously-trained set of phoneme models [3]. The use of articulatory models in this paper is motivated by the desire to generalize the pronunciation evaluation task while minimizing the human input required so that it can be made as automatic as possible. Assuming ready access to prior knowledge of the machine prompts to be spoken by the student (and possibly some background information about a registered user's native language, the phones and coarticulations expected to be most difficult, etc.), the method proposed requires no human supervision other than a dictionary of phoneme-level transcriptions for each

feature	*classes*	*# classes*
Jaw	Nearly Closed, Neutral, Slightly Lowered, Lowered	4
Lip Separation	Closed, Slightly Apart, Apart, Wide Apart	4
Lip Width	Rounded or Protruded, Slightly Rounded, Neutral	3
Tongue Body 1	Back, Slightly Back, Neutral, Slightly Forward, Forward	5
Tongue Body 2	Low and Flat, Mid, Mid-High, High	4
Tongue Tip	Low, Neutral, Touching Teeth, Near Alveolar Ridge, Touching Ridge	5
Velic Aperture	Closed, Open	2
Voicing	Unvoiced, Voiced	2

Table 1. Articulatory feature space. Note the gradual physical progression among classes within a given feature.

utterance and a mapping to their corresponding canonical articulations. This study can be thought of as an extension of our previous work in syllable stress detection [15] – the next step in automatically detecting and modeling mispronunciations on multiple time scales.

The data used in these experiments was compiled by the University of Leeds in their ISLE corpus [1]. These recordings consist of 46 adult Intermediate British English learners who are native speakers of either Italian or German – 23 of each. Utterance prompts were complete sentences designed to highlight specific difficulties English learners typically encounter in pronouncing single phone pairs, phone clusters, and primary stress pairs. The recordings were automatically segmented by a forced-aligner, then these transcriptions were augmented on the phone level by a team of five linguists to reflect each speaker's pronunciation (though no effort was made to correct discrepancies in the automatic segmentation times).

2. HIDDEN-ARTICULATOR MARKOV MODELS

2.1. Previous work

Articulatory models are defined by partitioning a given phone set into a number of overlapping sub-classes, each one representative of a specific position within a particular articulatory "property" or "feature." Because each articulatory gesture is shared among multiple phones, articulatory models do not necessarily require training data as extensive or balanced as that of phonemic models [13]. A Hidden-Articulator Markov Model (or HAMM) is defined as a standard Hidden Markov Model in which the "hidden" state is one possible articulatory configuration - for example "velic aperture closed" or "stop burst" - which is probabilistically linked to a sequence of observations. In this case, the state sequence is quite literally hidden because we cannot observe an articulation directly without the aid of, for instance, an Electromagnetic Articulograph, as done in [6]. All we know is all the transcribers of the ISLE corpus knew: what was perceived, and what production mechanism most likely generated it (though this box is especially black when working with nonnative speakers of English, as we cannot necessarily make the same articulatory assumptions we would for a native speaker).

Various articulatory class configurations have been proposed, depending on the application. The assignments defined in [10] have served as a useful model in the experiments reported in [7] and [11], but are perhaps too abstract to be of much use in pronunciation evaluation or language learning applications. After all, if a Japanese student of English effects the classic mispronunciation of substituting /r/ for /l/, is an automatic language tutor going to tell her, "Try it again, this time more approximant and retroflex, less lateral and coronal"? No, because such terminology is perhaps meaningless to a student not well-versed in linguistics (and even to one who is, it might not clearly indicate the best method of reconfiguring her vocal tract). Additionally, in this paradigm the model configurations for certain pairs of oft-substituted vowels are not always distinct. A good example is the substitution of /iy/ for /ih/ frequently made by Italian learners of English – [10]'s feature space renders them both as high, front vowels, without distinction.

2.2. Choice of models

Chosen because of their representation in concrete physical terms, the models used in this study (enumerated in Table 1) are based primarily on the Hidden-Articulator Markov Models proposed in [13] (which, in turn, have their linguistic basis in [5]), but with some important differences. First of all, rather than training each entire articulatory configuration as a separate state (all phoneme mappings plus other physically possible articulations and transition states among them), we built a separate Hidden-Articulator Markov Model for each of the eight features (Jaw, Lip Separation, etc.), and classified every one separately. This allowed for a degree of asynchrony among the articulators, so that the results might mimic the overlapping behavior of a true vocal tract's constituent parts [14]. It also permitted the generation of independent "language models" specific to each feature, and simplified the training and testing process, since in this quantization scheme no feature has more than five states (or "classes"), compared to the several thousand states trained in [13]'s previous work.

These simplifications rest on the assumption of independent motion among these eight articulator streams, which in a different study might not be valid – it allows for results that could potentially violate the fundamental

German				Italian			
difficult phone	*Common error*	*err freq*	*% of all errs*	*difficult phone*	*common error*	*err freq*	*% of all errs*
/z/	/s/	18.38	8.37	/uh/	/uw/	32.10	1.16
/ax/	/uh/	3.33	4.56	/ih/	/iy/	28.51	9.64
/v/	/f/	10.56	3.65	/ah/	/ax/	22.77	3.78
/w/	/v/	7.13	2.87	/ax/	/oh/	7.22	2.42
/uw/	/uh/	2.98	0.84	/t/	/t/ + /ax/	6.72	3.67
/ah/	/ax/	15.93	9.18	/ng/	/ng/ + /g/	20.27	1.29
/t/	/deletion/	2.19	4.09	/er/	/eh/ + /r/	3.02	0.40

Table 2. Phones of interest, their common mispronunciations, the frequency of this unique error, and the error's frequency as a percentage of all errors in the ISLE corpus, according to the human transcriptions.

physical constraints of the human vocal tract (e.g. dependencies between the jaw position and lip separation, tongue tip and tongue body, etc.). But the point of this project was not to build an articulator-based speech recognizer or even a general phoneme recognizer. Rather, we intended to demonstrate clustering of correct and incorrect pronunciations in articulatory feature space, regardless of the accuracy in recognizing an individual articulatory feature.

Perhaps quantizing the Lip Width feature into "rounded," "slightly rounded," and "neutral" positions seems a bit arbitrary and rigid, but the important thing is that such a partitioning scheme can divide a phone set into perceptually meaningful classes (from the perspective of acoustic observations), and can be used to effectively distinguish between a canonical articulation and its most common systematic mispronunciation, regardless of the assumptions made about dependencies among the output streams. In fact, such clustering (explained below) should perform better under an assumption of independence, since disallowing physically impossible articulatory configurations will seriously limit the representation of fine pronunciation distinctions within articulatory feature space. If the results point toward a physically unlikely or noncanonical articulation, it probably signifies the presence of a pronunciation error, and that is exactly what we intend to detect.

3. EXPERIMENTS AND RESULTS

3.1. Feature expansion and model training

Using the hand-corrected phone-level transcriptions from the ISLE corpus, we began the training procedure by mapping every phoneme sequence to its canonical series of articulatory configurations as specified in [13], with minor modifications to suit the ISLE phone set. For training purposes, phonemes with more than one valid articulation (such as /er/ with its variable Tongue Tip) were distilled to just one pronunciation – a reasonable compromise since many other examples of an excluded articulation class might be found elsewhere, in other phones (if we assume all realizations of /er/ to have Tongue Tip at Low rather than Neutral, an instance of Neutral Tongue Tip can be found in a phone like /p/ or /ng/). Though the transcribers did correct all context-dependent phone substitutions, the ISLE corpus wasn't transcribed for articulation. So in the case of a phone with multiple valid articulations, there's no way of knowing which one was actually used by the speaker. But our phoneme-articulator mappings did preserve finer-scale sequences such as a stop consonant's closure and release. And these intra-phone sequences, though also not hand-segmented in the creation of the ISLE data, were included in our training stage. The training data here came specifically from the BLOCK D recordings of the ISLE's native German and Italian speakers, which comprised roughly 80 sentence-length utterances per speaker.

As described in the last section, we used this articulatory expansion of the phone-level transcriptions to train 8 separate classifiers, each one with a different number of articulation classes. In addition to the classes listed in Table 1, each HAMM classifier included silence and background models, all with three left-right HMM states per model and four mixtures per state. These model parameters were initialized using a flat-start procedure and updated with embedded Baum-Welch re-estimation. All the observation data were represented with 39-dimensional MFCC vectors (the first 13 cepstral coefficients, normalized, plus their corresponding delta and acceleration coefficients) sampled at 10 ms intervals using a 25 ms Hamming window. Everything so far was done using HTK [2].

3.2. Articulation recognition and feature vector extraction

With these models and a bigram "language model" to constrain the physical possibilities of articulatory transitions within a given feature (e.g. Tongue Body 2

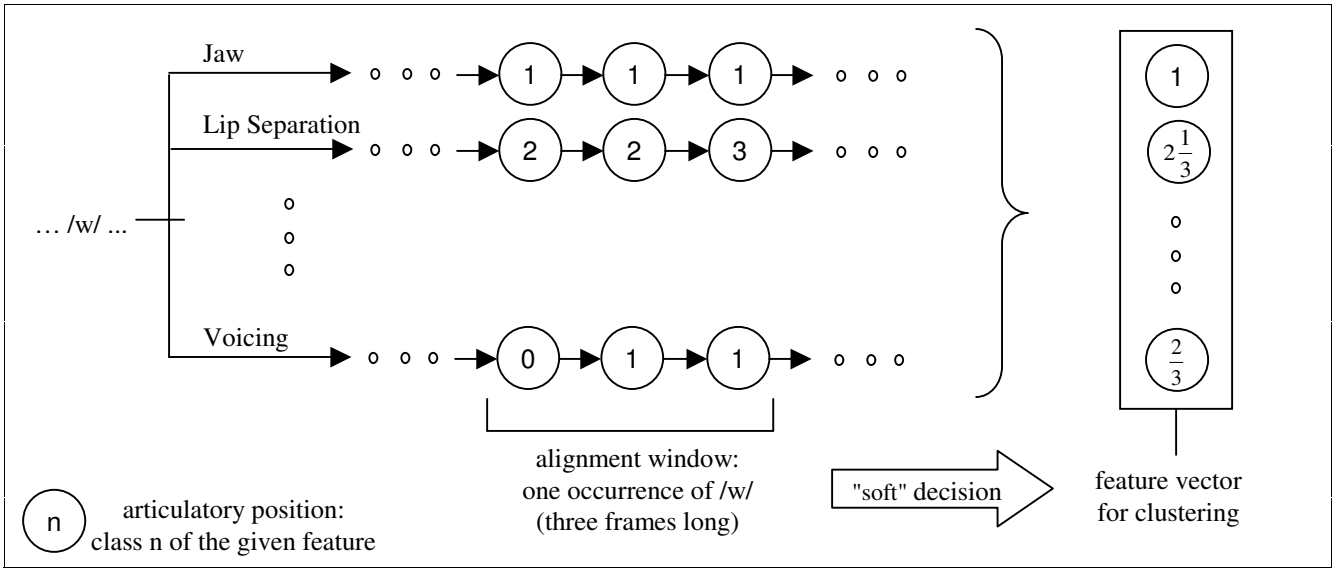

Figure 1. An example of decoding 8 HAMM streams and then extracting the 8-dimensional articulatory feature vector from them. This occurrence of /w/ will be clustered with others so as to distinguish it from the cluster of /v/ occurrences.

cannot pass from Mid to High without first going through Mid-High), we performed Viterbi decoding for articulation recognition on the BLOCK E, F, and G recordings from the same 46 ISLE speakers (a total of about 90 complete sentences per speaker). Given the assumption of independent motion among the output streams, the uncertain nature of our articulatory mapping, and the end goal of this study, the specifics of these results are not important – suffice it to say that the accuracy varied dramatically from one classifier to another (on average between 60 and 70%).

But holding to the hypothesis that, given reasonable results, these eight classifiers should be able to distinguish a specific correct pronunciation from its most common systematic error, we set upon locating within these multi-streamed results some meaningful feature vectors corresponding to the pronunciations of interest. [1] lists the most difficult English phones for the German and Italian speakers in the ISLE corpus, along with their associated most common mispronunciations. Those used in our analysis are listed in Table 2. Note that /ah/ was not omitted, even though its most common mispronunciation, /ax/, has the same canonical articulation under the mapping derived from [13].

Starting with the forced-alignment segmentation of the ISLE data, we matched up occurrences of the difficult phones (pronounced correctly or not) to the eight streams of our recognition results. These occurrences varied by phone, but usually numbered on the order of 1000. Since forced-alignment segmentation times (provided along with the ISLE recordings, and purposely left uncorrected by the annotators) are sometimes erroneous on the phone level, and since the decoding results would probably overlap asynchronously with the forced alignment times, we executed a "soft" decision scheme that averaged all articulatory results within the alignment segmentation interval, allowing for resulting vectors lying "between" the previously-defined quantization positions. This was done by initially representing each class as an integer number, in a sequence consistent with the physical progression of classes within a given feature (see Table 1), similar to what was done in [13]. For example, the four Jaw Position classes were assigned: 0, Nearly Closed; 1, Neutral; 2, Slightly Lowered; 3, Lowered. So, for an occurrence of /w/, if during its alignment interval (let's say eight frames in this example) the Jaw results were 0 for five frames and 1 for the remaining three frames, the soft decision scheme would output 0.375 as the overall Jaw position result for that particular occurrence. A similar illustration can be found in Figure 1.

After matching these occurrences to their articulation results using the above decision scheme, we assigned each occurrence to one of two classes for supervised clustering based on the ISLE transcriptions: correct pronunciation, or its most common expected error. With the MATLAB PRTools nearest-mean classifier [4], we used these eight-dimensional "soft" decision vectors (one for each occurrence) to generate a separate binary classification rule for each difficult phone and its corresponding error, empirically deriving the best relative sizes of a random partitioning of the vectors into training and test sets (though there were far fewer errors than correct pronunciations, we maintained equal priors in each of these sets).

Performance of this final classification procedure is reported in Table 3.

German				Italian			
phoneme - error	*Precision*	*recall*	*accuracy*	*phoneme - error*	*precision*	*recall*	*accuracy*
/z/ - /s/	67.59	79.32	57.46	/uh/ - /uw/	74.34	72.51	57.91
/ax/ - /uh/	72.76	78.98	61.04	/ih/ - /iy/	72.18	72.37	56.40
/v/ - /f/	74.40	80.70	63.24	/ah/ - /ax/	79.67	83.71	68.98
/w/ - /v/	76.97	82.74	66.21	/ax/ - /oh/	69.52	75.69	56.81
/uw/ - /uh/	80.57	77.07	64.49	/t/ - /t/ + /ax/	71.70	73.02	56.59
/ah/ - /ax/	72.38	78.68	60.47	/ng/ - /ng/ + /g/	76.38	72.39	59.14
/t/ - /deletion/	71.48	76.96	58.42	/er/ - /eh/ + /r/	70.88	75.39	57.64

Table 3. Percentage results by phoneme, each averaged over three random training and test set partitions.

4. DISCUSSION

[1] reported an inter-annotator agreement of "at best" 70% when simply detecting the location of a pronunciation error (but not deciding what the error is). This agreement rating was presented somewhat vaguely, so it is not clear if their error localization measure compensated for missed detections or false alarms. But the precision and recall ratings in Table 3 all lie close to or better than 70%, and some of the accuracy results get close as well. These results all come from unsupervised classification of human-tagged speech, so results around 70% indicate that our method performs as well as a human annotator would. Since this method is intended to evaluate pronunciation in students of a foreign language, a small number of false alarms is tolerable, as that will only err on the side of requiring the student to practice her pronunciation more.

Of course, [1]'s results were averaged over all possible phonemes and errors, and for each native language we only considered seven, with each one's most common error. But the phonemes and systematic errors investigated here account for roughly 20-30% of all phone-level pronunciation errors in the ISLE corpus. Moreover, they encompass different types of mispronunciations on various articulatory levels, with consistent results. Canonical articulations for /ax/ and /uh/ differ in the Jaw, Lip Width, and Tongue Body features, whereas /uw/ and /uh/ differ only in terms of Lip Separation and Lip Width, /v/ and /f/ differ only in Voicing, and so on. So, our results can be thought of as representative of this method's potential performance on the corpus at large.

Considering the relative sparseness of the incorrectly-pronounced data (see the *err freq* column in Table 2), the results for detection of each of these phoneme's most common errors are consistent between the two native languages and among all phonemes. Even /ah/ and /ax/, though in training mapped to the very same articulatory configuration, could be distinguished with performance comparable to the other, more dramatic misarticulations. The slight overall decline in performance for the native Italian speakers (as compared to the Germans) can be attributed to a difference in the level of experience between the two populations [1]. The native Italian speakers in this corpus were less familiar with English than the native German speakers, and everything, even the correct pronunciations, they tended to pronounce with a heavy Italian accent. As a result, for the Italian speakers there was a shorter "distance" (in articulatory feature space) between any pair of canonical and mispronounced phones. Additionally, German bears a closer phonological relationship to English than Italian does, so the German speakers could usually articulate the subtle difference in /uw/ and /uh/ (for instance) more adeptly than the Italian speakers, therefore making the German mispronunciation more easily classifiable. But the results in Table 3 suggest that the method proposed here, to be useful in language-learning contexts, is not dependent on any particular native language.

What about other phone-level errors besides the most common unique mispronunciation? In the native German section of the ISLE corpus, /uh/ is substituted for /ax/ almost as often as /oh/, /uw/, /ae/, and /eh/. In distinguishing between a correctly pronounced /ax/ and any error at all, we found the results to be the same as in distinguishing between only /ax/ and /uh/. So this method has the potential to be incorporated into more general pronunciation evaluation tasks, or even phoneme recognition. Given robust models of canonical and non-native phones, articulatory features could be applied to any generic pronunciation evaluation task, regardless of the languages or mispronunciations involved.

For the /z/ and /v/ substitutions (both of which differed from their most common error only in the Voicing feature), we obtained comparable results by simply measuring periodicity and f0 values within the regions of interest and excluding all other articulatory information, instead of modeling the voicing feature as an HAMM time series. Nonetheless, our results demonstrate the power of articulatory features alone to distinguish between these close phonemes.

This method of locating an expected pronunciation

mistake represents a preliminary step in the complex pronunciation evaluation task. Once the presence of a recurring error has been identified, a robust language learning system should provide physically meaningful instruction to the student as to a proper articulatory reconfiguration ("keep your lips rounded," for instance).

Because we have assumed prior knowledge of the student's native language and its associated systematic mispronunciations, and because we have restricted this classification problem to those difficult phones and their one most common error, this type of feedback can be given without necessarily precise results in the recognition of the articulatory gestures themselves – the erroneous pronunciations seem to cluster in articulatory feature space regardless. To further generalize this method that it may be applied to all types of errors (and not just the most common ones), future researchers in this area would be best served to implement a more advanced graphical model – a Hidden-Articulator Dynamic Bayesian Network – that would allow for dependencies among the recognized states and compensate for the coarticulatory constraints of a true speech production system.

5. ACKNOWLEDGEMENTS

This work was supported in part by grants from the National Science Foundation through a CAREER and an IERI award.

6. REFERENCES

[1] E. Atwell, P. Howarth, C. Souter, "The ISLE Corpus: Italian and German Spoken Learner's English," *ICAME Journal*, Vol. 27, pp. 5-18, 2003.

[2] Cambridge University, *HTK 3.2*, http://htk.eng.cam.ac.uk/, 2002.

[3] R. Delmonte, M. Petrea, C. Bacalu, "SLIM Prosodic Module for Learning Activities in a Foreign Language," *Proc. ESCA*, Eurospeech97, Rhodes, Vol. 2, pp. 669-672.

[4] R.P.W. Duin, *PRTools 3.1.7*, Delft University of Technology, the Netherlands, http://www.prtools.com/, 2002.

[5] H.T. Edwards, *Applied Phonetics: The Sounds of American English*, Second Edition, Singular, San Diego.

[6] A. Gutkin and S. King, "Detection of Symbolic Gestural Events in Articulatory Data for Use in Structural Representations of Continuous Speech," *Proc. ICASSP'05*, Philadelphia, 2005.

[7] K. Hacioglu, B. Pellom, and W. Ward, "Parsing Speech into Articulatory Events," *Proc. ICASSP'04*, Montreal, 2004.

[8] S.-C. Jou, T. Schultz, and A. Waibel, "Whispery Speech Recognition Using Adapted Articulatory Features," *Proc. ICASSP'05*, Philadelphia, 2005.

[9] A. Kazemzadeh, H. You, M. Iseli, B. Jones, X. Cui, M. Heritage, P. Price, E. Anderson, S. Narayanan, and A. Alwan, "Tball data collection: the making of a young children's speech corpus," in *Proc. of EUROSPEECH*, Interspeech, Lisbon, Portugal, 2005.

[10] K. Kirchhoff, "Robust Speech Recognition Using Articulatory Information," PhD thesis, University of Bielefeld, 1999.

[11] K.Y. Leung, M.W. Mak, and S.Y. Kung, "Applying Articulatory Features to Telephone-based Speaker Verification," *Proc. ICASSP'04*, Montreal, 2004.

[12] N. Mote, A. Sethy, J. Silva, S. Narayanan, and L. Johnson, "Detection and modeling of learner speech errors: The case of Arabic tactical language training for American English speakers," in *Proceedings of InStil*, Venice, Italy, July 2004.

[13] M. Richardson, J. Bilmes, and C. Diorio, "Hidden-Articulator Markov Models for Speech Recognition," *ASR2000*.

[14] J. Sun and L. Deng, "An overlapping-feature-based phonological model incorporating linguistic constraints: Applications to speech recognition," J. Acoust. Soc. Am., Vol. 111, No. 2, 1086-1101 (2002).

[15] J. Tepperman and S. Narayanan, "Automatic Syllable Stress Detection for Pronunciation Evaluation of Language Learners," *Proc. ICASSP'05*, Philadelphia, 2005.