AUTOMATIC DETECTION OF VOICE ONSET TIME CONTRASTS FOR USE IN PRONUNCIATION ASSESSMENT

Abe Kazemzadeh[1], *Joseph Tepperman*[1], *Jorge Silva*[1], *Hong You*[2], *Sungbok Lee*[1], *Abeer Alwan*[2], *and Shrikanth Narayanan*[1]

[1]Speech Analysis and Interpretation Laboratory, University of Southern California
[2]Speech Processing and Auditory Perception Laboratory, University of California, Los Angeles

ABSTRACT

This study examines methods for recognizing different classes of phones from accented speech based on voice onset time (VOT). These methods are tested on data from the Tball corpus of Los Angeles-area elementary school children [1]. The methods proposed and tested are: 1) to train models based on standard English VOT contrasts and then extract the VOT characteristics of the phones by measuring the duration of phone-level and sub-phone-level alignments, 2) to train phone models with explicit aspiration, and 3) to train different models for different phoneme classes of VOT times. Error rates of 23-53% for different phone classes are reported for the first method, 5-57% for the second method, and 0-36% for the third. The results show that different methods work better on different phone classes. We interpret these results in relation to past research on VOT, explain possible uses for these findings, and propose directions for future research.

Index Terms: recognition of speech variation, pronunciation variation, voice onset time.

1 Introduction

Voice Onset Time (VOT) is defined as the length of time between the release of closure of a stop and the onset of voicing [2]. Stop consonants are produced with a closure of the vocal tract at a specific place which is known as the place of articulation. In English, the stop classes of /p,b/, /t,d/, and /k,g/ are produced at the labial (lips), alveolar (gum ridge), and velar (soft palate) places of articulation, respectively. During the closure, there is a build up of sub-laryngeal pressure. When the closure is released, there is a transient burst of air, then some frication due to turbulence at the place of articulation, and possibly aspiration noise from turbulence caused by the glottis being in an open or spread position [3]. Voicing may begin before, at the moment of, or after the release of closure. When the start of voicing comes after the release of closure, the VOT has a positive value, when the start of voicing is coincident with the release of closure VOT is zero, and when the start of voicing precedes the release of closure (i.e., there is a voicing bar during the closure) for a stop, the VOT takes on a negative value (the distance from the release backward to the beginning of the voicing bar).

VOT is an important characteristic of stop consonants that plays a large role in perceptual discrimination of phonemes of the same place of articulation [4]. It is also at play in word segmentation, stress related phenomena, and dialectal and accented variations in speech patterns, [5, 6]. For example, in American English, voiceless stops have a long VOT with aspiration when at the beginning of a word and when in a simple onset of a stressed syllable, but have a shorter, unaspirated VOT when embedded in consonant clusters or when in an onset of an unstressed syllable.

For most languages, within a given place of articulation stops are differentiated by their laryngeal setting and its timing with respect to oral closure. In particular, voice onset times and aspiration are common contrastive laryngeal features. Within a given language, these can be represented by the features voiced/voiceless or aspirated/unaspirated. However, when distinguishing the phones of different languages, it is necessary to consider voicing as a continuum represented by VOT.

In our research, this point is illustrated by comparing the stop sequences of English and Spanish. In English the classes of voiced and voiceless stops correspond to stops with near zero VOT (voicing beginning at the same time that closure is released) and stops with a positive VOT of 50-80 ms (a delay in the onset of voicing with respect to release of closure) and aspiration [2]. In Spanish however, the classes of voiced and voiceless stops correspond to stops that have a negative VOT of -40 ms (voicing begins 40 ms before the closure is released) and stops that have near zero VOT [5]. Thus, the class of voiced stops in English bears similarity to the class of voiceless stops in Spanish with respect to VOT. Hypothetically, in the case of accented speech these sounds may be mispronounced. In this study, we examine methods in automatic speech recognition that may be used to detect these differences.

This work is part of the TBall project [1], which aims to build standards-based assessment of early literacy in diverse

student populations. In building such a system to accept or reject an example of a child's read speech [7], we must consider variations that may occur due to children's speech with possibly a non-native accent [8].

2 Data

The data used in this study comes from the Tball corpus [1]. This corpus contains speech from children of ages 5-8 being tested on oral reading skills. The children come from a variety of different language backgrounds, the majority of whom are native Spanish speakers learning English (69%). A subset of 5740 single word utterances were used. These utterances were transcribed with an enhanced set of phone symbols that detailed pronunciation variations and which could be mapped to the CMU dictionary pronunciation symbols. The added symbols included dental versions of alveolar stops, unaspirated voiceless stops, prevoiced stops (i.e., negative VOT), lispy /s/, and a trill, among others [1]. These additional symbols were intended to capture speech tendencies of our target population, namely children who may have accented speech.

Although the additional phones provided more information on variant pronunciations, they also created sparseness issues for training. However, this problem was partially relieved by not requiring absolute transcriber agreement and by merging confusable phone pairs (e.g., the symbols for a short VOT /t/ and a dental /t/ were merged). The problem of sparsity prevented using a baseline approach of simply training phone models for discriminating the different variant phone classes, but the enhanced transcriptions provided enough examples to evaluate our methods of distinguishing the variant phone classes.

3 Method

To differentiate pronunciation variations of stop classes we focused on distinguishing long VOT aspirated voiceless stops /p^h, t^h, k^h/ (the standard English pronunciation) from short VOT stops /p, t, k/ (accented variants) in word initial position. We used hidden Markov model (HMM) based approaches for these classification experiments. These experiments were carried out using the Cambridge Hidden Markov Toolkit (HTK). In addition to these different approaches, the effect of frame rate in parameterizing the speech waveform was analyzed to see if a higher resolution improved the classification accuracy. Besides these settings, the other HMM parameters stayed the same. The phone models were 3-state HMMs with 4 mixtures for each of the 39 features. These features were the energy and 12 cepstral coefficients, and the delta and acceleration figures for each. The training data included 5686 utterances. Examples of short VOT phone variants included instances that were transcribed as such, or ones that occurred after a word initial /s/ or before an unstressed syllable. The training data was bootstrapped with 275 manually aligned utterances. The test data included 800 utterances.

We tested three different methods. The baseline method (1a) involved measuring the duration of the stop and using this information to determine a threshold that would separate the long VOT stop from the short VOT stop. A slight improvement of this baseline was to examine the duration of the third (final) HMM state instead of the whole phone model (1b). A second approach we tried was to explicitly include aspiration as an HMM state (2). The third method (3) we tested was to train separate models for each phone variant and then use the likelihood of each model to select the best one.

The intuition behind method (1a) is simply that a stop with a long VOT will have a longer recognized duration than one with a short VOT. To test this, phone-level forced alignments were used to measure the duration of the phone of interest. The transcription symbols for phone variants were collapsed to the standard English transcription symbols and the durations of these phones were measured. Different thresholds were tested to determine the best cut in separating the original long and short VOT stop classes (see figure 1).

A slight improvement of this baseline approach was to measure the duration of individual states from the HMM phone model (method 1b). This refined the baseline method by focusing on states of the model that correspond to the region between the release of closure and the onset of voicing. Manual analysis of the forced alignments and plots of the duration distribution of the different HMM states confirmed that the third state corresponded best to the amount of VOT of a given phone.

Another approach that we tested was to explicitly include aspiration in the transcriptions when training the recognizer (2). To do this we inserted the /h/ phoneme in the transcriptions that contained aspirated stops. Word initially this insertion depended on the enhanced transcriptions and word internally the aspiration symbol was inserted if the stop was located in the onset of a stressed syllable and not in a consonant cluster. The dictionaries were expanded to include both the presence and absence of the aspiration symbol. The dictionary entries were used to decode the appropriate words of the test data. If the recognizer output contained the aspiration symbol after the first stop, then the these stops were considered to be long VOT stops.

The final method (3) that we tested was to use the forced alignments to segment the test speech waveforms so that only the stop consonants of interest were left. With the separate classes, long and short VOT models were trained for each phone. Then the segmented clips were evaluated with the trained models and the likelihood of both long and short VOT stops were compared with respect to each other. Whichever had the higher probability determined the classification of each phone.

To evaluate the results of classifying the voiceless stop phones into either the common pronunciation (long VOT) or the accented pronunciation (short VOT), we measured the error rate for each class separately. This was necessary because there were much fewer instances of the non-standard pronunciations. Had a total error rate been used, error rates as low as 5% could be attained by simply classifying every phone as one of the long VOT class.

When using thresholds in methods (1a) and (1b), the point of equal error rate for both classes (when the error rate measurement for each class is equal) was used. We used this because adjusting the threshold in either direction tilted the error rate to one class or the other. This point represents giving the performance of each class equal weight.

4 Results

The baseline results of using a threshold on phone duration to detect short VOT pronunciation variants can be seen in figure 1.

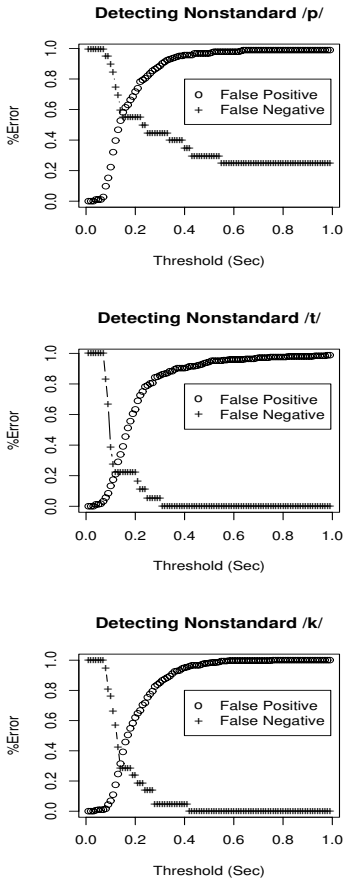

Fig. 1. Baseline method, showing the trade-offs in setting using different thresholds. The results can be seen as % error for classification or rate of false positives and negatives in detecting the accented form).

The point of equal error rate (when the error for each class is equal) for /k/ is 29% and for /t/ is 23%, but for /p/ it is 55%. This can be seen as reflecting the observation that VOT times tend to be longer for articulations that are more posterior as a general tendency across languages [9], (i.e., the VOT of k>t>p).

Using the duration of the third (final) HMM state duration gave better results than the baseline for /p/ and /t/ (19% and 20%) but worse for /k/ (48%). Other combinations of states produced results similar to or worse than the baseline. The shapes of the graphs for this method were similar to the previous method, but even steeper. The precision of determining the threshold seemed to be limited by the 10 ms frame rate of the speech recognizer front-end, but retraining the recognizer at higher frame rates did not improve the results.

Using an aspiration HMM gave good performance for detecting short VOT variants of /p/ and /t/ (5% and 11% error, respectively), but gave worse results for the short VOT /k/ (57%). For detecting the long VOT /k/ this method performed well (17% error), but not for long VOT /p/ and /t/ (36% and 38% error, respectively). This could be due to the /k/ in Spanish being more aspirated and having a longer VOT, making it more similar to the English /k/.

Using the method of comparing model probabilities gave good results for all the phone classes except for the short VOT /p/ (36% error). Because of the few instances of short VOT, the same phones used in the training were used in the testing, so these results could indicate over-fit models. The poor performance of the /p/ could be due to the fact that bilabial stops tend to have less frication noise in general, which may be a salient factor in the the HMM models for the /t/ and /k/.

The results are summarized in the following table 1.

Results Summary (Percent Error)						
	/p/		/t/		/k/	
	acc'd	nat.	acc'd	nat.	acc'd	nat.
alignment duration (model)	55%		23%		29%	
alignment duration (3rd state)	19%		20%		48%	
aspiration model	5%	36%	11%	38%	57%	17%
model prob. comparison	36%	4%	0%	5%	0%	6%

Table 1. Results summary of different stops and VOT characteristics: accented–short VOT (acc'd) and native–long VOT (nat.). The top 2 rows use thresholds set at an equal error rate (as can be seen in figure 1 where the lines cross).

5 Discussion

One of the benefits of using a duration-based thresholding approach is that it can be done with forced alignment, so it can be seen as a by product of the recognizer. However, this assumes that the word is correctly recognized, which is not always the case, especially when dealing with non-standard pronunciation. Another problem is that the frame rates used in parameterizing speech for recognition do not provide adequate resolution when comparing phone durations (or sub-phone states).

The method of using an inserted aspiration model was easily implementable by modifying the label files and dictionary. It circumvented the problem of sparse data by using the fact that stops in some contexts are similar to the non-standard phone variants and by using the /h/ phoneme, which was trained in other contexts as well.

The method of using likelihood ratios is an approach that works within the framework of HMMs by having separate models for each phone variant. The drawback of this approach is that it is often the case that there may be too few instances of a variant pronunciation to adequately train separate models for each phone. The results reported for this method may be too optimistic because the limited amount of data required testing on the training data.

In general, the results show that different methods used work better for certain phones but worse for others. This is most likely due to the fact that the stops of different places of articulation have different VOT characteristics.

One step for future work would be to set up an inter-transcriber agreement task that focuses on VOT, since transcriber agreement figures may be too general to assess the ability to discriminate these particular phonetic variations, as in [4]. Moreover, it would be desirable to relate the VOT characteristics with subjective measures of accentedness in naive and expert listeners.

Knowledge of VOT characteristics could be useful in many speech processing tasks. In a pronunciation assessment task like [7], the VOT characteristics of stops may be one feature used in a more comprehensive system that also relies on other recognized features. Knowledge about a speaker's language background could also be used as an indicator to trigger certain phonetic rules that apply to regular patterns in other aspects of pronunciation variation [8], resulting in models that better fit the speaker.

VOT differences may also be seen as an individual speaker's variation, or as a stress-related phenomenon, so potentials exist to make use of VOT distinctions for speaker recognition and prosodic information extraction. Additionally, it may be that VOT also has developmental factors, with children hypothetically having more variation than adults due to articulator timing and coordination issues, as well as psycholinguistic issues such as the notion of a critical period for language learning.

6 Conclusion

In this study several ways of classifying stop phones based on VOT characteristics were examined. Using durations from forced alignments resulted in a performance of 20-30% error rates, depending on whether the duration measured was the whole phone model or the third HMM state. Using an explicit model of aspiration, low error rates were achieved for detecting short VOT variants of /p/ and /t/ as well as the long VOT (standard) variant of /k/. The method of comparing the model probabilities of the short and long VOT classes produced good results in all of the cases except for short VOT /p/.

7 Acknowledgments

We would like to thank the NSF for its support of this work. An early version of work was presented as an abstract at the 2005 Fall meeting of the Acoustical Society of America.

References

[1] Abe Kazemzadeh, Hong You, Markus Iseli, Barbara Jones, Xiaodong Cui, Margaret Heritage, Patti Price, Elaine Anderson, Shrikanth Narayanan, and Abeer Alwan, "Tball data collection: the making of a young children's speech corpus," .

[2] Peter Ladefoged, *A Course in Phonetics*, Heinle & Heinle, Boston, 2001.

[3] Kenneth N. Stevens, *Acoustic Phonetics*, MIT Press, Cambridge, MA, 2000.

[4] Jintao Jiang, Marcia Chen, and Abeer Alwan, "On the perception of voicing in syllable-initial plosives in noise," *Journal of the Acoustical Society of America*, vol. 119, no. 2, Febuary 2006.

[5] Leigh Lisker and Arthur S. Abramson, "A cross language study of voicing in initial stops," *Word*, vol. 20, pp. 384–422, 1964.

[6] Leigh Lisker and Arthur S. Abramson, "Some effects of context on voice onset time in english stops," *Language and Speech*, vol. 10, pp. 1–28, 1967.

[7] Joseph Tepperman, Jorge Silva, Abe Kazemzadeh, Hong You, Sungbok Lee, Abeer Alwan, and Shrikanth Narayanan, "Pronunciation verification of children's speech for automatic literacy assessment," in *ICSLP*, Pittsburgh, 2006, submitted.

[8] Hong You, Abeer Alwan, Abe Kazemzadeh, and Shrikanth Narayanan, "Pronunciation variation of Spanish-accented English spoken by young children," in *Proceedings of Eurospeech*, Lisbon, 2006.

[9] Peter Ladefoged, *Phonetic Data Analysis*, Blackwell Publishing, Oxford, 2003.