USE OF TRAJECTORY MODELS FOR AUTOMATIC ACCENT CLASSIFICATION

Pongtep Angkititrakul and John H.L. Hansen[†]

Robust Speech Processing Group
Center for Spoken Language Research (CSLR)
University of Colorado at Boulder, Boulder, CO, 80302, U.S.A
{angkitit,jhlh}@cslr.colorado.edu, web: http://cslr.colorado.edu

Abstract

This paper describes a proposed automatic language accent identification system based on phoneme class trajectory models. Our focus is to preserve discriminant information of the spectral evolution that belong to each accent. Here, we describe two classification schemes based on stochastic trajectory models; supervised and unsupervised classification. For supervised classification, we assume text of spoken words are known and integrate this into the classification scheme. Unsupervised classification uses a Multi-Trajectory Template, which represents the global temporal evolution of each accent. No prior text knowledge of the input speech is required for the unsupervised scheme. We also conduct human-perceptual accent classification experiments for comparison automatic system performance. The experiments are conducted on 3 foreign accents (Chinese, Thai,and Turkish) with native American English. Our experimental evaluation shows that supervised classification outperforms unsupervised classification by 11.5%. In general, supervised classification performance increases to 80% correct accent discrimination as we increase the phoneme sequence to 11 accent-sensitive phonemes.

1. Introduction

The field of automatic accent classification is a challenging and interesting research area, since the manner in which a primary(L1) language accent is conveyed during the production of speech in a second language(L2) will depend on both languages, and perhaps, the speakers. The ability to separate accent sensitive traits from speaker sensitive traits is an overriding goal for our accent classification work. The ability to achieve reliable accent classification performance offers potential knowledge to improve speech and speaker recognition systems by directing alternate pronunciation dictionaries, phoneme/word models, or modification of mixture weights in HMM or GMM classifiers [10, 13]. A number of previous studies have considered accent or dialect classification of American English [2, 3, 10]. Some studies have considered accent classification based on formant locations, duration features, pitch profile histograms, energy contour histograms, cepstral features, temporal structure (voice-onset time, stop release time), and isolated or monophone based HMMs [2, 9, 4].

The accent classification system presented in this paper employs a Stochastic Trajectory Model (STM)[6] on individual phoneme classes across a speech utterance, so that we can better capture the spectral/temporal structure over the duration of the phone[12]. For supervised classification, prior text knowledge and forced alignment are combined to generate a sequence of phoneme tokens of the utterance. Accent likelihood scores are generated for each phoneme segment in the utterance and accumulated to produce an overall accent discrimination score. In general, the process of accent classification has been typically used in front-end analysis for accent-specific automatic speech recognition. Such front-end processing would not be responsible for understanding the input of the speech sequence. It is therefore not necessary to decode the speech segment into a string of possible words as is common in speech recognition. Unsupervised classification based on our STM scheme is also proposed to support such applications. The Multi-Trajectory Template(MTT) will consist of mixtures of trajectories, which are trained from all phoneme classes of the same accented speech. The MTT therefore represents the global trajectory movements of all phones for each accent. It is believed that the hypothesized accent will be from the MTT which coincides with the major trajectory components seen in the recognition stage.

Accent classification is a challenging problem since there is no clear boundaries between accent classes (i.e., a speaker may have a mild to heavy accent). In order to study the classification performance, we also conduct a human perception study to compare with automatic algorithm discrimination results. In this paper, we conduct experiments using isolated words from CU-Accent corpus [14]. We concentrate on discriminating between native American English versus Chinese, Thai, and Turkish accents. This paper is organized as follows; Section 2 describes the proposed accent classification schemes. In Section 3, experimental evaluations are conducted using isolated words. The paper ends with conclusions and suggestions for future work.

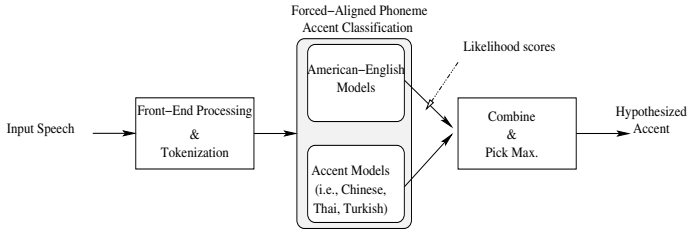
2. Accent Classification Algorithms

The two proposed automatic accent classification schemes use only acoustic information, without requiring any high-level knowledge from a language model. Figure 1 show the block diagrams of our basic accent classification algorithms.

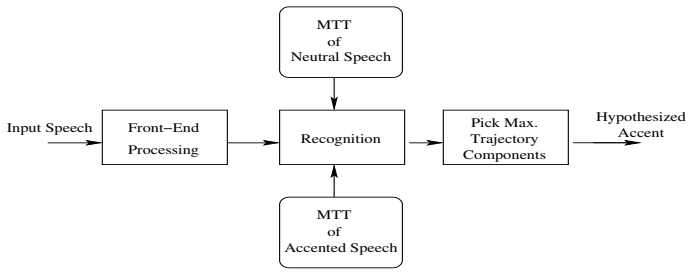
2.1. Stochastic Trajectory Model (STM)

Let $\mathbf{X}$ be a sequence of Q points: $\mathbf{X} = (\mathbf{x_0}, \mathbf{x_1}, \ldots, \mathbf{x_{q-1}})$, where each point is a D-dimensional vector in a speech production space. $\mathbf{X}$ is obtained by re-sampling a sequence of d frames according to a linear time scale. Here, we assumed each phoneme symbol is associated with a set of stochastic generators of trajectories, and a phone model may be viewed as a mixture

[†]This work was supported by the U.S. Air Force Research Laboratory, Rome NY, under contract number F30602-01-1-0511.



(a)Supervised accent classification algorithm.



(b)Unsupervised accent classification algorithm.

Figure 1: *Basic Block diagram for accent classification.*

of trajectory models. A similar concept was originally proposed as a source generator framework for modeling speech production in emotional/stressful condition [7, 8]. Here, the probability density function(pdf) of a segment $\mathbf{X}$, given a duration of d and the phoneme symbol s, can be written as:

$$p(\mathbf{X}|d,s) = \sum_{t_k \in T_s} p(\mathbf{X}|t_k,d,s) Pr(t_k|s) \qquad (1)$$

where T_s is the set of trajectory components associated with s. $Pr(t_k|s)$ is the probability of trajectory t_k given phoneme s, with the constraint $\sum_{k \in T_s} Pr(t_k|s) = 1, \forall s$. $p(\mathbf{X}|t_k,d,s)$ is the pdf of the vector sequence $\mathbf{X}$, given that we know the component trajectory t_k, d and s. The distribution assigned to each of the Q sample points on a trajectory is characterized by a multivariate Gaussian distribution with a mean vector $\mathbf{m}^s_{k,i}$, and covariance matrix $\mathbf{\Sigma}^s_{k,i}$. Using an assumption of frame-independent trajectories, the pdf is modeled as:

$$p(\mathbf{X}|t_k,d,s) = \prod_{i=0}^{Q-1} Gaussian(\mathbf{X}; \mathbf{m}^s_{k,i}, \mathbf{\Sigma}^s_{k,i}). \qquad (2)$$

In order to estimate a model, we need to consider segments representing the same phonetic units that are of different durations, and therefore we will normalize or re-scale all segments to have a fixed duration. The fixed duration length can be thought of as the *underlying spectral trajectory* of $\mathbf{X}$, and $\mathbf{X}$ is the realization of the fixed duration due to variations in speaking rate.

2.2. Supervised Classification

At the training stage, a sequence of Mel-cepstral coefficients are tokenized into individual phoneme segments by using forced alignment and prior text knowledge. A set of 37 context-independent phone models are trained for each accent under consideration. Here, we assume gender dependent models in order to reduce speaker variety due to genders. In practice, using a Gaussian Mixture Model to decide speaker gender is quite effective and almost always correct. During testing, a sequence of Mel-cepstral coefficients are also tokenized into individual phoneme segments by using forced alignment and prior text knowledge. The accent log-likelihood score of each phoneme is calculated from the accent dependent trajectory model corresponding to that phoneme token. The hypothesized accent is the accent which gives the maximum accent likelihood score for each utterance.

The log-likelihood score computed from each STM model is defined as:

$$L_{STM}(\mathbf{X}|s) = log(p(\mathbf{X}|d,s)) + \beta \cdot log(p(d|s)), \qquad (3)$$

normalized by the number of trajectory points Q as the likelihood score from each phone model. Here, $p(d|s)$ is the duration(in frames) distribution of each accent phoneme, modeled using a Γ-distribution. β is the adjustable duration weight.

2.3. Unsupervised Classification

The unsupervised classification scheme is based on the underlying assumption that speech spoken with different accents will occupy a distinct set of trajectories in the feature space, namely the Multi-Trajectory Template(MTT). The MTT consists of mixtures of trajectories, when each mixture represents a global trajectory movement of one acoustic class or a set of acoustic classes. The MTT can be interpreted as a soft representation of the various trajectories of acoustic classes that make up the sounds of a specific accent. Analogous to GMM for a speaker recognition task, we can interpret the GMM as a soft representation of the various acoustic classes that make up the sounds of the speaker; each component density can be thought of as the distribution of possible vectors associated with each of the acoustic classes, each class representing possibly one speech sound or a set of speech sounds.

During training, similar to supervised classification, a sequence of Mel-cepstral coefficients are tokenized into individual phoneme segments by using forced alignment and prior text knowledge. The MTT is trained from all phoneme classes using the LBG clustering technique in a similar fashion to STM training [1] but we ignore the duration distribution. The number of clusters or mixtures, M, should be equal or larger than the number of phoneme classes in the task. For our experiment, we used $M = 50$. During recognition, an input speech utterance is decoded using three MTTs trained from English speech, an accent speech style, and a silence model. The formulation of an utterance recognition is based on maximizing the probability of the utterance with a sequence of trajectory segments.

Let an utterance u be represented by $L(u)$ trajectory segments, $u \triangleq a_1, \ldots, a_h, \ldots, a_{L(u)}$, where a_h is a trajectory component from an MTT. Let the last time slot number of of the hth trajectory segment a_h be an unknown constant n_h. A segmentation of u is represented by a set of slot numbers $\{n_h\}$, where $n_0 = 0, n_{L(u)} = N$, and N is the number of frames in the utterance. Utterance recognition then consists of searching for the set of slot numbers that maximize $\Theta(u|O)$ [6], as follows:

$$\begin{aligned} \Theta(u|O) &= \max_{n_1,\ldots,n_{L(u)}} Pr(u|n_1,\ldots n_{L(u)}, O) \qquad (4) \\ &= \max_{n_1,\ldots,n_{L(u)}} Pr(a_1,\ldots,a_{L(u)}|n_1,\ldots n_{L(u)}, O) \\ &= \max_{n_1,\ldots,n_{L(u)}} \prod_{h=1}^{L(u)} Pr(a_h|\mathbf{X}_{\frac{n_{h-1}+n_h+1}{2}}, n_h - n_{h-1}), \end{aligned}$$

where $\mathbf{X}_{\frac{n_{h-1}+n_h+1}{2}}$ is a sequence of Q vectors centered at time slot $\{n_{h-1}+n_h+1\}/2$, and $n_h - n_{h-1}$ is frame duration, similar

to d in Eq. 1. After decoding, the MTT which contributes the majority of the trajectory components in the utterance is used to decide the speech accent.

3. Experimental Results

This section describes our CU-Accent database, experiments on accent identification, and evaluation results. In this study, we are interested in pairwise classification between native American English (**En**) versus 3 foreign accents; Chinese(**Ch**), Thai(**Th**), and Turkish(**Tu**). Due to the limited number of speakers, 10 male speakers/accent, we performed 5 experiments for each setup by using a Round-Robin procedure (i.e., training on 8 speakers and testing on the remaining 2 speakers, then repeat). The final results are the average of 5 experiments using all open test tokens. All experiments were conducted on isolated word speech, with all models based on phonemes.

3.1. CU-Accent Speech Corpus

The CU-Accent speech corpus[14] was organized and collected at CSLR for algorithm formulation in acoustic-phonetic studies in automatic accent classification and speaker recognition. The speech was collected across a telephone channel, digitized at 8000 Hz and stored in 16-bit linear PCM format. The database contains 8 accents with 4 major accents having more than 10 speakers per gender. The corpus contains 5 tokens of 23 isolated words, 4 short sentences in both English and speaker's native language, and one minute of spontaneous speech on any topic of interest to the speaker. Each subject provided detailed written information on their language background, age, and occupation. Each speaker's L1 language was used as their true accent for (L2) American English.

3.2. Human Perception

This section briefly describes a human perception study which was conducted on 2 native American Speakers, and 6 speakers who use English as their second language. Using a single-wall sound booth, meeting ASHA standards for ambient noise, a formal listener evaluation was performed. Each listener was asked to classify a list of speakers for accent type between native American and accented speech using a set of randomly selected words of 5 speakers from each accent. The listener evaluation was performed in three phases: human accent classification based on (i) 1 word, (ii) 2 words, and (iii) 3 words. Each listener was able to listen to each test token multiple times before making their decision. No listener had any history of hearing loss, and each listener was able to adjust the volume to a comfortable level. On the average, each isolated word contained 4 phonemes (i.e., human accent classification performance using on average 4, 8, and 12 phoneme strings). Fig. 2 summarizes human accent classification performance. We see that after 8 phoneme(2 words) levels off to between 87-90% accent classification for Chinese and Thai accent. Turkish performance showed more variability.

3.3. Experiments

We conducted experiments on speech from forty speakers for three accent classes(Chinese, Thai, and Turkish) and neutral class(English). The isolated-word speech was parameterized with 12 MFCCs and normalized log energy. For supervised classification, in both training and test stages, speech was segmented automatically using forced alignment. A single aligner

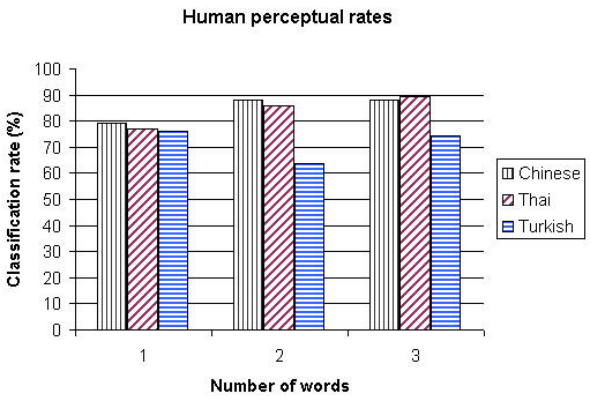


Figure 2: *Human perceptual rates.*

based on Viterbi algorithm was used for consistent and portability of the system. The acoustic models of the aligner were trained from another database. STM-based classification used 5 sample points and 2 trajectory mixtures. For comparison, we trained a conventional HMM based classifier in a supervised manner. Each HMM model consisted of 5 emitting states, with each state having 2 Gaussian mixtures. For unsupervised classification, one MTT with 50 mixtures of trajectories was trained for each accent during the training stage. Table 1 shows the percent correct accent classification rates for supervised and unsupervised schemes at one word test of speech, compared to human classification rates. From the table, STM performed slightly better than HMM. For all methods, classification rates of Chinese/English and Thai/English were comparable, while the classification rate for Turkish/English was the lowest.

	Ch/En	Th/En	Tu/En
Human	79.38	76.88	76.25
HMM	69.73	69.45	63.69
STM	**70.61**	**71.76**	**65.91**
MTT	60.69	58.27	54.77

Table 1: *Accent classification rates at (isolated)word level.*

From our empirical studies, several phoneme classes(i.e., vowels, diphthongs) had more accent discriminability than the others. Next, we consider accent classification experiments that employ only perceptually important accent phone in the speech signal. Supervised classification was performed on entire phoneme sequences of isolated words, but accent likelihood scores were only computed from the accent-sensitive phoneme set (i.e., /iy, ih, eh, ae, aa, er, axr, ax, uw, uh, ay, ay, aw, ey, w, l, r, y/) [1]. Fig. 3 shows the accent classification performance employing STM and HMM models using a decision with single versus phoneme sequence string (i.e., 1, 5, 11, and 17 phonemes used per decision). Increasing the number of phonemes was obtained from random concatenation of different isolated words. As expected, as the number of phonemes increases, both classification schemes performed better in a similar fashion to human perceptual rates. Accent classification rates of Chinese/English and Thai/English were consistently better than Turkish/English in both HMM and STM.

For completeness, Figure 4 shows the DET curves [11] of classification performance between English and Thai accents at

[1] see [1] for a detailed comparison of accent classification over individual phonemes.

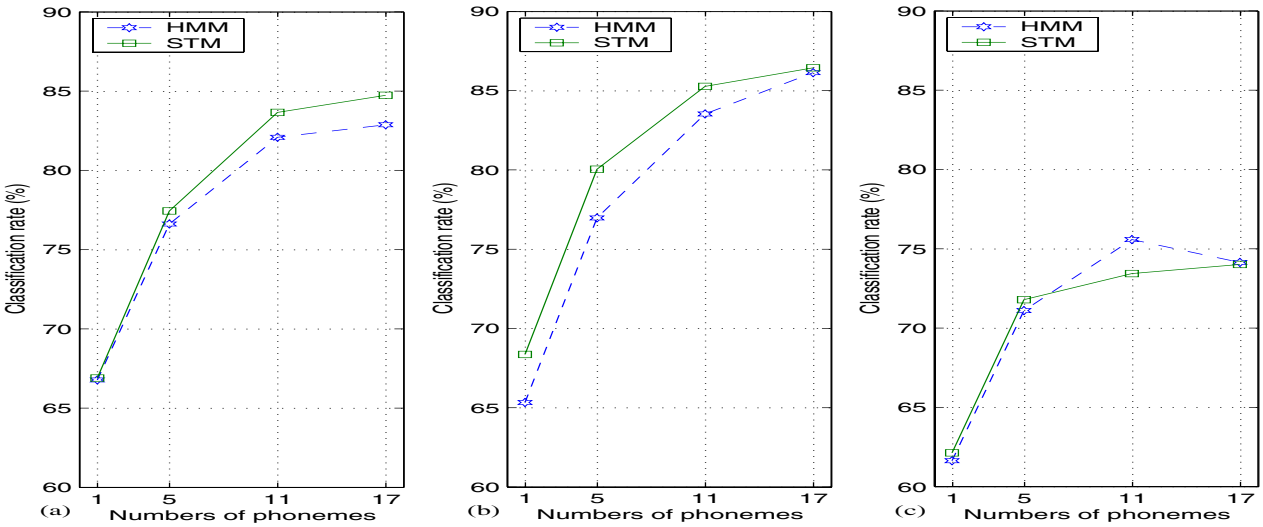


Figure 3: *Accent classification rates vs number of phonemes for (a) Chinese/English, (b) Thai/English(center),and (c) Turkish/English.*

1, 5 and 11 phonemes. The closer line to the lower left corner shows better detection performance. STM again performed slightly better than HMM based accent classification.

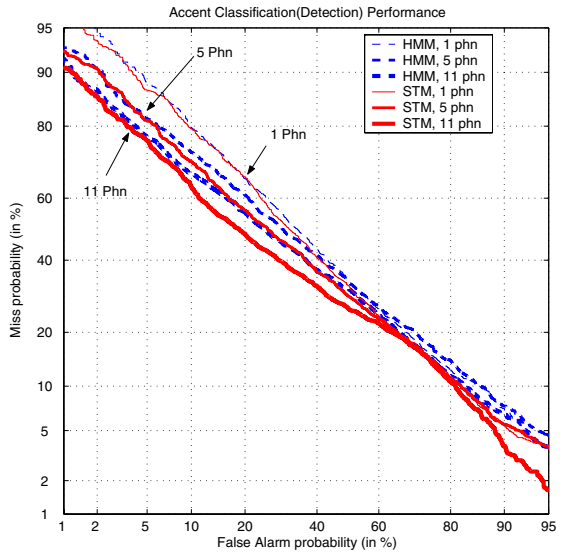


Figure 4: *Accent detection performance for classifying Thai accent against American English using HMM and STM models with different phoneme lengths(1,5,11), in the decision process.*

4. Conclusions and Future Work

In this paper, we presented an accent classification method that employs trajectory models in an effort to capture more specific spectral structure over the phoneme duration. Our phoneme-based automatic accent classification algorithm and evaluations illustrated that useful accent discriminant information is preserved in the accent trajectory models in both supervised and unsupervised schemes. From our study, supervised STM trajectory based classification outperformed unsupervised trajectory based classification. STM always performed slightly better than HMM based classification. In general, unsupervised classification is more flexible for practical accent classification applications. Finally, we saw that STM performed 9,5, and 10% below human accent classification for Chinese, Thai, and Turkish accents. Since speakers may not consistently display accent content across speech, human classification performance should serve as a reasonable approximation to ground truth for these speakers.

Our future work will focus on diphone segments and projection space analysis which increase accent discrimination. We will also consider performance trade-offs using spontaneous speech, and again compare performance with human accent classification.

5. References

[1] P. Angkititrakul, J.H.L. Hansen, "Stochastic Trajectory Model Analysis for Accent Classification," in *Proc. ICSLP'02*, 493-496, 2002.

[2] L. Arslan, J.H.L. Hansen, "Language Accent Classification in American English," *Speech Communications*, vol. 18(4), pp. 353-367, July 1996.

[3] J.E. Flege, "Factors affecting degree of perceived foreign accent in English sentences," *J. Acoust. Soc. Amer.*, Vol 84, No. 6, 70-79, 1988.

[4] P.J. Ghesquiere, D.V. Compernolle, "Flemish Accent Identification based on Formant and Duration Features," in *ICASSP'02*, 749-752, 2002.

[5] T. Gleason, M. Zissman, "Composite Background Models and Score Standardization for Language Identification Systems," in *ICASSP'01*, May 2001.

[6] Y. Gong, "Stochastic Trajectory Modeling and Sentence Searching for Continuous Speech Recognition," *IEEE Trans. Speech and Audio Proc.*, 5(1): 33-44, 1997.

[7] J. H. L. Hansen, "Adaptive source generator compensation and enhancement for speech recognition in noisy stressful environments," in *ICASSP'93*, 95-98, Apr 1993.

[8] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Communications*, vol. 20, pp. 151-173, November 1996.

[9] K. Kumpf and R.W. King, "Automatic accent classification of foreign accented Australian English speech," in *Proc. ICSLP'96*, 1740-1743.1996.

[10] M. Lincoln, et al, "A Comparison of Two Unsupervised Approaches to Accent Identification," *ICSLP'98*, 1998.

[11] A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, "The DET curve in assessment of detection task performance," in *ICASSP'97*, 1895-1898, September 1997.

[12] M. Ostendorf, and S. Roukos, "A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition," *IEEE Trans. Acoustic, Speech, and Signal Proc.*, 37(12):1857-1869, 1989.

[13] C. Teixeira, I. Trancoso, A. Serralheiro, "Accent Identification," in *ICSLP'96*, 1996.

[14] http://cslr.colorado.edu/beginweb/accent