Advances in Phone-Based Modeling for Automatic Accent Classification

Pongtep Angkititrakul, *Member, IEEE,* and John H. L. Hansen, *Senior Member, IEEE*

***Abstract*—It is suggested that algorithms capable of estimating and characterizing accent knowledge would provide valuable information in the development of more effective speech systems such as speech recognition, speaker identification, audio stream tagging in spoken document retrieval, channel monitoring, or voice conversion. Accent knowledge could be used for selection of alternative pronunciations in a lexicon, engage adaptation for acoustic modeling, or provide information for biasing a language model in large vocabulary speech recognition. In this paper, we propose a text-independent automatic accent classification system using phone-based models. Algorithm formulation begins with a series of experiments focused on capturing the spectral evolution information as potential accent sensitive cues. Alternative subspace representations using principal component analysis and linear discriminant analysis with projected trajectories are considered. Finally, an experimental study is performed to compare the spectral trajectory model framework to a traditional hidden Markov model recognition framework using an accent sensitive word corpus. System evaluation is performed using a corpus representing five English speaker groups with native American English, and English spoken with Mandarin Chinese, French, Thai, and Turkish accents for both male and female speakers.**

***Index Terms*—Automatic accent classification, dialect modeling, open accent classification, phoneme recognition, spectral trajectory modeling, speech recognition.**

I. Introduction

ACCENT classification, or as it is sometimes referred to as accent identification, is an emerging topic of interest in the speech recognition community since accent is one of the most important factors next to gender that influences speech recognition performance [15], [18], [29]. Accent knowledge could be used for selection of alternative pronunciations in a lexicon, engage adaptation for acoustic modeling, or provide information for biasing a language model in large vocabulary speech recognition. Accent knowledge can be useful in speaker profiling for call classification (e.g., route incoming Spanish accented calls to bilingual Spanish operators), as well as for data mining and spoken document retrieval [33].

Manuscript received December 29, 2003; revised June 24, 2004. This work was supported by the U.S. Air Force Research Laboratory, Rome, NY, under Contract F30602-01-1-0511. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Geoffrey Zweig.

P. Angkititrakul is with the Center for Robust Speech Systems, Eric Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX 75083-0688 USA (e-mail: angkitit@utdallas.edu).

J. H. L. Hansen was with the Robust Speech Processing Group, Center for Spoken Language Research, University of Colorado Boulder, Boulder, CO 80302 USA and is now with the Center for Robust Speech Systems, Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX 75083–0688 USA (e-mail: john.hansen@utdallas.edu).

Digital Object Identifier 10.1109/TSA.2005.851980

Foreign accent (which researchers simply refer to as "accent") can be defined as the patterns of pronunciation features which characterize an individual's speech as belonging to a particular language group. Accent, as well as nonnative grammatical utterances, is often due to a negative transfer from the first language (L1) of a speaker to a second language (L2) [22]. A number of studies have been directed at analysis and classification of movement from different L1 to L2 accent groups [1], [5], [19], [21]. The few accent classification techniques that have been proposed are based on the underlying assumption that a nonnative speaker will systematically substitute sounds (phones) from his/her native language (L1) for those foreign sounds (phones) they are less familiar. Given this reasonable assumption on phone substitution in speech production, accent classification represents a significant challenge since nonnative speech is so diverse. Successful accent classification must be able to address the tremendous intra- and inter-speaker variation in the production of spoken language.

Accent Classification assumes that speakers are intentionally speaking the same language in a natural manner. The level of accent exhibited in second language pronunciation will depend on a number of speaker related factors such as: 1) the age at which a speaker learns the second language; 2) the nationality of the speaker's language instructor; and 3) the amount of interactive contact the speaker has with native talkers or radio/TV/text exposure of the second language. For our study, we maintain a reasonable balance in age distribution across each accent. Our approach to addressing the three speaker factors effecting accent is to collect a reasonable number of speakers for each accent to reduce inter-speaker variability and thereby focus on accent sensitive traits. We point out that a number of studies on accent typically focus on a small number of speakers (i.e., typically 1–5 per accent) [4], [6], [7], [31]. In our study, the goal was to obtain at least 10 speakers (either a male set, and/or female set) per accent. We did not consider specific differences between speakers for a given accent regarding age of learning L2, instructor nationality, and exposure/contact with L2 materials such newspaper, books, TV, radio, etc. It would be difficult to quantify the exposure to English as an L2 language for these speakers, and therefore we assume that a sample size of at least ten offers a reasonable balance. We also did not consider dialect issues among the same native language, since we assumed that speakers share similar accent traits in speech production. A recent study of native Japanese speakers acquiring English (L2) in the mid-United States and in south eastern United States strongly suggested that speakers acquire similar accent characteristics that are less dialect dependent for American English [26].

Over the past decade, several research studies have investigated accent sensitive acoustic speech characteristics and robust accent classification/detection schemes. Many features have been proposed to use as accent sensitive traits at both high and low levels of acoustic/linguistic knowledge. Grover [14] verified that French, English and German speakers differ in the slope of their intonation contours. Flege [6] performed experiments to test the detect-ability of native listeners for French accented speakers, where detection rates varied between 63% and 95%. Kumpf [20] applied linear discriminant analysis (LDA) for the discrimination of three accents in Australian English. In that system, a single feature vector was extracted from each phoneme segment which combines acoustic (12 MFCC + log energy), prosodic (phoneme segment duration, pitch, and pitch slope), and contextual information (categorical features describing the phonetic left and right context of the segment). Schults, *et al.* [31] employed a nonverbal cue identification framework, using phone strings produced by different context-independent phone-recognizers instead of traditional short-term acoustic vectors, to differentiate between native and nonnative speakers of English. Blackburn, *et al.* [5] employed a cascade of multi-layer perceptrons for speech segmentation and subsequent foreign speaker accent classification. Zissman [34] discriminated dialects of Latin American Spanish by combining phoneme recognition with dialect-dependent language modeling. Lincoln, *et al.* [21] investigated two unsupervised approaches, one acoustic and the second a phonotactic model approach, to automatically identify accent and reported results on the problem of discriminating American- and British English- accented speech. Berkling [4] showed that the position within the syllable is important for accent classification because the pronunciation patterns of accented speakers vary as a function of the phoneme's position and lexical stress. Ghesquiere, *et al.* [10] employed formant frequency location and duration as features to identify Flemish accents in read speech. That study showed that vocal tract length normalization (VTLN) based on eigenvoices followed by mean normalization produced an effective method to reduce organic speech characteristics.

In the past, the process of accent classification has been typically used in front-end analysis for an accent-specific automatic speech recognition system. Such a front-end would not be responsible for understanding the input speech sequence. It is therefore not necessary to decode the speech segment into a string of possible words as is common in speech recognition. One could argue that such information is also useful for discriminating accents, since the selection of words, grammar rules employed, and overall sentence structure can also convey accent information. In our study, we will consider these features as high-level accent knowledge cues. In this study, we describe our text-independent accent classification system using phoneme-based Hidden Markov Model (HMM) recognition as a baseline system, expanding from our previous work which was based on the likelihood scores produced by a text-dependent accent-dependent HMM recognizer [1]. Unfortunately, the conventional HMM assumes that the sequence of features vectors is produced by a piecewise stationary process. Typical HMM modeling assumes that adjacent frames are acoustically uncorrelated and that the state-dependent duration distributions are exponentially decreasing, which breaks down in a number of speech classification applications. Several segmental models have been proposed to reduce this independence assumption within the HMM, and attempts to explicitly capture the time variations of particular features within a segment have also been considered. Segment-based models offer a better framework for modeling the evolution of the spectral dynamics of speech, adding flexibility and power as seen in whole-segment classification, and in contrast to the traditional frame-by-frame based paradigm.

Ostendorf and Roukos [28] introduced the stochastic segment model (SSM) and used a fixed-length trajectory representation in each phone model for the task of traditional large vocabulary speech recognition. Stochastic trajectory models (STMs) were used for modeling phone-based speech units as clusters of trajectories in the parameter space. Gong *et al.* [13] focused on trajectories which consist of N sample points within a segment and are represented by a mean and covariance vector for each point in a large vocabulary continuous speech recognition (LVCSR). The trajectories are modeled by mixtures of state sequences of multivariate Gaussian density functions to explain inter-frame dependencies within a segment. Fukada, *et al.* [8] represented trajectories by polynomial-constrained trajectory models within a phone segment. Goldenthal [12] employed a nonparametric trajectory model which describes the dynamics of acoustic attributes over a segment. This method tracks the parameters that generate a synthetic segment model which is then compared against the speech segment to be classified. The resulting error is used to estimate the likelihood of the segment being classified as a particular phone. Gish and Ng [11] employed a linear least-square error estimation for the polynomial parametric trajectory, which exploits the time dependence of speech frames by representing the speech features of a speech segment as Gaussian mixtures with time-varying parameters. Most, if not all, of these recent studies have considered spectral trajectory modeling as a way to improve LVCSR, and not specifically for accent classification.

The proposed procedure for accent classification is as follows. An input utterance or word sequence is submitted, with text knowledge of the input.[1] The audio sequence is forced-aligned using an HMM reference phoneme sequence. The HMM phoneme sequence includes combinations of phonemes in English and each accent of interest. The final likelihood score for each accent and for each utterance is calculated as the linear combination of the individual normalized log likelihoods resulting from the corresponding acoustic models associated with each accent. The hypothesized accent is chosen from the maximum likelihood (ML) score at each decision time. The motivation for this approach is the notion that explicitly modeling trajectories of phone-transitions as the acoustic representation will be an important accent-sensitive trait.

This paper is organized as follow. Section II begins with a discussion of two different approaches to represent speech trajectories, and their trajectory scoring mechanisms. Our proposed subspace projection approach is discussed in Section III. After a brief description of our database in Section IV, we develop

[1]In our proposed method, we assume the text is known. In cases where this information is not available, we would perform an initial ASR pass to obtain an estimated text sequence.

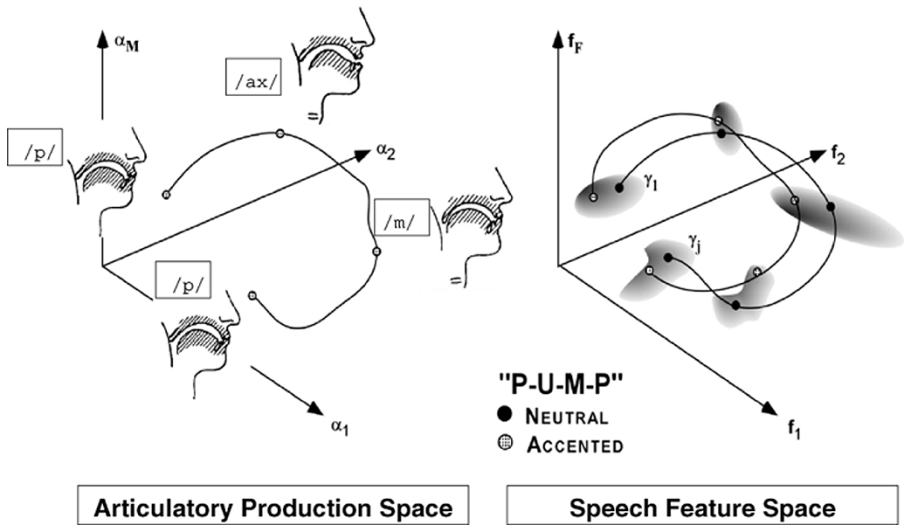

Fig. 1. Hypothetical speech production space and the corresponding feature space with movement based on native and nonnative speakers.

our baseline system and establish performance in Section V. Section VI presents our proposed trajectory-based system with subspace projection technique, followed by a series of accent classification experiments. Section VII closes with a discussion and conclusions.

II. Trajectory Models

In a parametric or cepstral space, a speech signal can be represented as a point which moves as the articulatory configuration changes. Fig. 1 shows a hypothetical speech production space and the corresponding feature space with movement based on native and nonnative speakers. A similar framework has been previously proposed, entitled Source Generator Framework for characterizing speech production under stress [16], [17]. Here, the sequence of points reflects movement in the speech production and feature spaces which can be called the *trajectory* of speech. The motivation from this observation is that the speech signal tends to follow certain paths corresponding to the underlying speech segments. Fig. 2 shows the vocal tract movement of the phoneme sequence /aa/–/r/ from the word "target" and the corresponding trajectories in the first two cepstral dimensional space for speakers with English, Chinese, Thai, and Turkish accents. We can see that there are certain trajectory similarities between Thai and Turkish accent speech, while the native English and Chinese accented speech are more similar. Clearly, a model that captures this spectral evolution could be useful for accent classification.

A. Stochastic Trajectory Model (STM)

An STM [13] represents the acoustic observations of a phoneme as clusters of trajectories in a parametric space. Each state is associated with a multivariate Gaussian density function, optimized at the state sequence level.

Let $\mathbf{X}$ be a sequence of N points: $\mathbf{X} = (\mathbf{x_0}, \mathbf{x_1}, \ldots, \mathbf{x_{N-1}})$, where each point is a D-dimensional vector in a speech production space. $\mathbf{X}$ is obtained by resampling a sequence of d frames along a linear time scale. The resampled N–frame vector $\mathbf{X}$ is

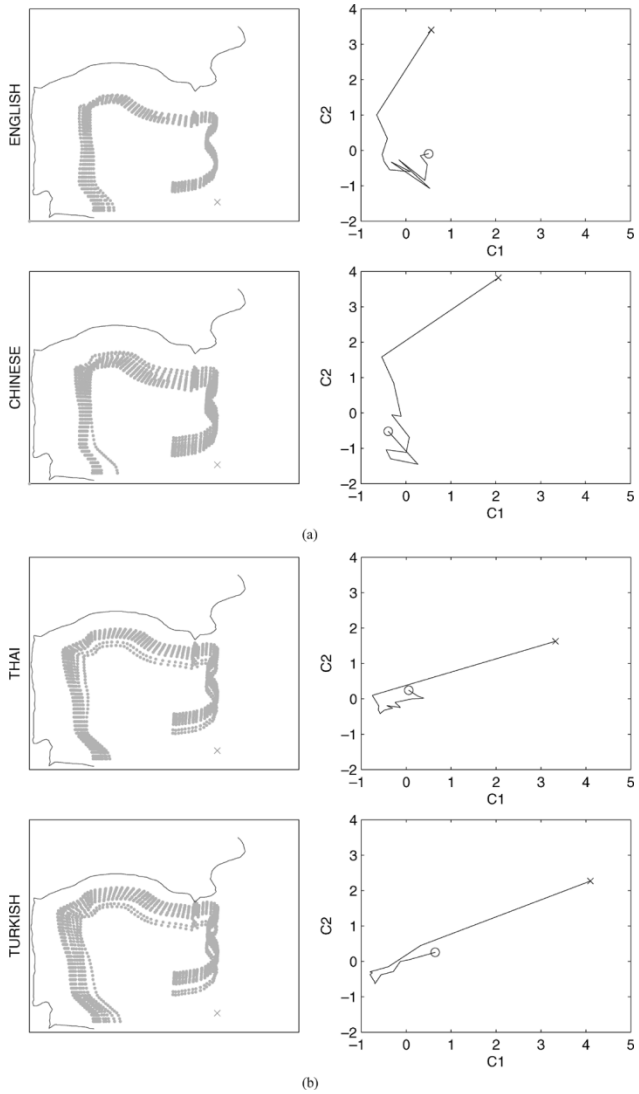

Fig. 2. Vocal tract movements of /aa/-/r/ from the word "target", and the corresponding trajectories in MFCC c1-c2 space ("o" start position of trajectory, "x" stop position of trajectory). (a) American English and Chinese. (b) Thai and Turkish.

considered to be the *underlying spectral trajectory* of the original d–frame vector $\mathbf{X}$. Here, we assume each speech segment is associated with a set of stochastic generators of trajectories, and a segment model may be viewed as a mixture of trajectory models. Here, the probability density function (*pdf*) of a segment $\mathbf{X}$, given a duration d and the segment symbol s is written as

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

where T_s is the set of all trajectory components associated with s. $Pr(t_k|s)$ is the probability of observing trajectory t_k given that we have segment class s, with the constraint that $\sum_{k\in T_s} Pr(t_k|s)=1, \forall s$. $p(\mathbf{X}|t_k,d,s)$ is the *pdf* of the vector sequence $\mathbf{X}$, given that we know the component trajectory t_k, duration d and symbol s. The distribution assigned to each of the N samples points on a trajectory is characterized by a multivariate Gaussian distribution with a mean vector $\mathbf{m}^s_{k,i}$, and covariance matrix $\mathbf{\Sigma}^s_{k,i}$. With the assumption of frame independent trajectories, the *pdf* is modeled as

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

Our training algorithm performs maximum likelihood estimation (MLE) of the parameters of the Gaussian distribution ($\mathbf{m}^s_{k,i}$, $\mathbf{\Sigma}^s_{k,i}$) which are equal to the sample mean and sample covariance matrix, respectively. The training algorithm consists of finding the ML estimate of the parameters, based on the application of the Linde–Buzo–Gray (LBG) algorithm in trajectory classification.

B. Parametric Trajectory Model (PTM)

An alternative to the STM is the PTM. The PTM treats each speech unit to be modeled by a collection of curves in the feature space, where the features typically are cepstral based. The class of trajectories we consider are low order polynomials such as a quadratic polynomial, based on previous PTM studies for LVCSR [11], [24].

For the parametric trajectory, we model each speech segment feature dimension as

$$c(n)=\mu(n)+e(n),\ \text{for}\ n=1,\ldots,N \tag{3}$$

where $c(n)$ are the observed cepstral features of the speech segment with a length of N frames, $\mu(n)$ is the mean feature vector representing the dynamics of the features in the segment, and $e(n)$ is the residual error term which is assumed to have a Gaussian distribution whose errors are assumed to be independent from frame to frame. For the mean of the feature vector, which models the trajectory, we consider this to be a quadratic function of time.

Given a speech segment with a duration of N frames, we obtain a resampled sequence of d frames according to a linear time scale, where each frame is represented by a D dimensional feature vector. The speech segment can be modeled as

$$C=ZB+E \tag{4}$$

where Z is an $N\times R$ design matrix which is determined by the nature of the trajectory, and for our case $R=3$ for quadratic trajectories. B is an $R\times D$ trajectory parameter matrix we wish to model, E is an $N\times D$ residual error matrix, and C is the feature vector matrix. We can express (4) in matrix notation as

$$\begin{pmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,D}\\ c_{2,1} & c_{2,2} & \cdots & c_{2,D}\\ \vdots & \vdots & & \vdots\\ c_{N,1} & c_{N,2} & \cdots & c_{N,D}\end{pmatrix}=\begin{pmatrix} 1 & 0 & 0\\ 1 & \frac{1}{N-1} & (\frac{1}{N-1})^2\\ \vdots & \vdots & \vdots\\ 1 & \frac{N-1}{N-1} & (\frac{N-1}{N-1})^2\end{pmatrix}$$
$$\times\begin{pmatrix} \beta_{1,1} & \beta_{1,2} & \cdots & \beta_{1,D}\\ \beta_{2,1} & \beta_{2,2} & \cdots & \beta_{2,D}\\ \beta_{3,1} & \beta_{3,2} & \cdots & \beta_{3,D}\end{pmatrix}+\begin{pmatrix} e_{1,1} & e_{1,2} & \cdots & e_{1,D}\\ e_{2,1} & e_{2,2} & \cdots & e_{2,D}\\ \vdots & \vdots & & \vdots\\ e_{N,1} & e_{N,2} & \cdots & e_{N,D}\end{pmatrix}. \tag{5}$$

By minimizing the least-square objection function, given K training tokens to train an M mixture component segmental trajectory model, the ML solution can be found using the expectation maximization algorithm (EM) as discussed below.

Let the likelihood of a sequence of speech features in segment k being generated by the model mean B_m, and model covariance Σ_m, be dependent on the segment via the estimate of the trajectory parameter $\hat{B}_k$, the estimate of the covariance matrix $\hat{\Sigma}_k$, and the number of frames in segment k, N_k, as

$$\begin{aligned} &L(\hat{B}_k,\hat{\Sigma}_k|B_m,\Sigma_m)\\ &=l(k|m)\\ &=(2\pi)^{-DN_k/2}|\Sigma_m|^{-N_k/2}\cdot\exp\left(-\frac{N_k}{2}tr[\Sigma_m^{-1}\hat{\Sigma}_k]\right)\\ &\quad\cdot\exp\Big(-\frac{1}{2}tr[Z_k(\hat{B}_k-B_m)\Sigma_m^{-1}(\hat{B}_k-B_m)^TZ_k^T]\Big). \end{aligned} \tag{6}$$

After each likelihood score $l(k|m)$ is calculated from the K training tokens, the posterior probability of the m^{th}-mixture, given segment k, can be computed as

$$p(m|k)=\frac{l(k|m)p(m)}{\sum_{j=1}^{M} l(k|j)p(j)} \tag{7}$$

where the following hold.

1) The prior probability for mixture component m is computed as

$$p(m)=\frac{1}{K}\sum_{k=1}^{K}p(m|k). \tag{8}$$

2) The trajectory parameters for mixture component m is computed as

$$B_m=\left[\sum_{k=1}^{K}p(m|k)Z_k^TZ_k\right]^{-1}\left[\sum_{k=1}^{K}p(m|k)Z_k^TZ_k\hat{B}_k\right] \tag{9}$$

$$\Sigma_m=\frac{\sum_{k=1}^{K}p(m|k)(C_k-Z_kB_m)^T(C_k-Z_kB_m)}{\sum_{k=1}^{K}p(m|k)N_k}. \tag{10}$$

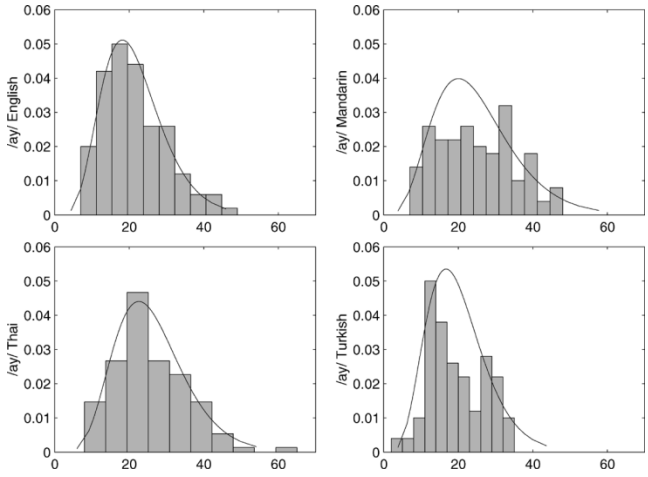

Fig. 3. Typical duration histograms (in frames, where each frame is 20-ms duration with 50% overlap) and their Γ-function modeling of phoneme /ay/ for four accents.

The updated parameters are then used to calculate the likelihood $l(k|m)$ for the next EM iteration.

C. Likelihood Score and Duration Distribution

At the classification stage, the likelihood of an unknown speech segment $\mathbf{X}$ given segment class s with T_s trajectories can be expressed as

$$p(\mathbf{X}, s) = p(\mathbf{X}|d, s)^{\alpha} \cdot Pr(d|s)^{\beta} \tag{11}$$

while the likelihood of an unknown test segment k coming from model m with M mixtures can be expressed as

$$p(k|d,s) = L(\hat{B}_k, \hat{\Sigma}_k) = \sum_{m=1}^{M} p(m) \cdot L(\hat{B}_k, \hat{\Sigma}_k | B_m, \Sigma m) \tag{12}$$

$$p(k,s) = p(k|d,s)^{\alpha} \cdot Pr(d|s)^{\beta}. \tag{13}$$

Here, $Pr(d|s)$ is the duration probability that segment $\mathbf{X}$ (from the STM model) or k (from the PTM model) of class s has frame duration d. This probability distribution can be modeled as a Γ-function obtained from a histogram of the training data, as in [13]. We also note that α and β are the control weights which are determined experimentally. Fig. 3 shows typical example duration histograms and Γ-function modeling of the phoneme /ay/ for four American-English accents.

III. Subspace Classification

The motivation for employing a subspace based classification technique is that a range of speech production factors are expected to be accent sensitive, and we wish to focus our effort on traits that are more accent sensitive and less intra-speaker dependent. Therefore, we wish to develop a systematic process by which to determine the optimal subspace where the projected trajectories are most discriminant. To address this, we will consider two approaches based on PCA and LDA.

A. Principal Component Analysis (PCA)

PCA or Karhunen-Loève expansion is a well established unsupervised technique for dimensionality reduction and data analysis. The principal directions resulting from this analysis are given by the eigenvectors of the data covariance matrix. The aim is to find a set of M orthogonal vectors in the data space that reflect the largest percentage of the data variance. Projecting the original D-dimensional data onto the M-dimensional subspace spanned by these vectors performs a dimensionality reduction, since $M < D$. The time-constrained PCA method [30] has been shown to be effective for trajectory projection when frame ordering is considered. Given data which consists of a set of K tokens, each containing N observations of dimensionality D, we form the data sequence $\vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_D$ with $\vec{\mathbf{x}}_k = \vec{\mathbf{x}}_k^{\{1\}}, \ldots, \vec{\mathbf{x}}_k^{\{N\}}$, where $\vec{\mathbf{x}}_k^{\{*\}}$ is a D-dimensional vector from token k consisting of N observations. We wish to evaluate the temporal evolution of these vectors in each sequence that is of interest. In order to preserve the temporal sequence information, we represent a scalable frame ordering as a time constrained transformation. Let $\boldsymbol{\tau}$ be a set of time constraints, $\boldsymbol{\tau} = 1, \ldots, N$. Our subspace definition using this time-constrained PCA framework[2] can be described by solving the sample covariance matrix at each time rescaling block τ, as follows:

$$C^{\{\tau\}} = \frac{1}{K} \sum_{k=1}^{K} (\mathbf{x_k}^{\{\tau\}} - \mu^{\{\tau\}})(\mathbf{x_k}^{\{\tau\}} - \mu^{\{\tau\}})^T,$$

$$\mu^{\{\tau\}} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{x_k}^{\{\tau\}}. \tag{14}$$

If we assume that the eigenpairs of $C^{\{\tau\}}$ are $(c_1^{\{\tau\}}, \lambda_1^{\{\tau\}})$, $\ldots, (c_N^{\{\tau\}}, \lambda_N^{\{\tau\}})$ and the eigenvalues are rank ordered $\lambda_1 \geq \cdots \geq \lambda_N$, then the transformation matrix $A^{\{\tau\}}$ will be $[c_1^{\{\tau\}}, \ldots, c_M^{\{\tau\}}]$ (i.e., the eigenvectors with the largest M eigenvalues, for each time constraint). The eigenvector with the highest eigenvalue represents the dimension in the eigenspace in which the variance of vectors is maximum in a correlation sense. We can therefore determine a threshold by which an eigen-dimension must have to be included in the reduced space. This represents the PCA approach to dimensionality reduction.

B. Linear Discriminant Analysis (LDA)

The main objective of discriminant projections is to minimize the variation of the projected features within a particular class, while maximizing the distance between the projected means of different classes. Fisher's LDA is a popular method for classification based on multidimensional predictor variables. LDA aims at improving discrimination between classes in a vector space, by finding a low-dimensional projection of the raw data such that the resulting space maximizes the ratio of between-class variation and within-class variation.

[2]We point out that the notation $\{\tau\}$ represents the τ^{th} time representation of the data covariance matrix, and the context of $\{\bullet\}$ represents a time location and not a power exponent.

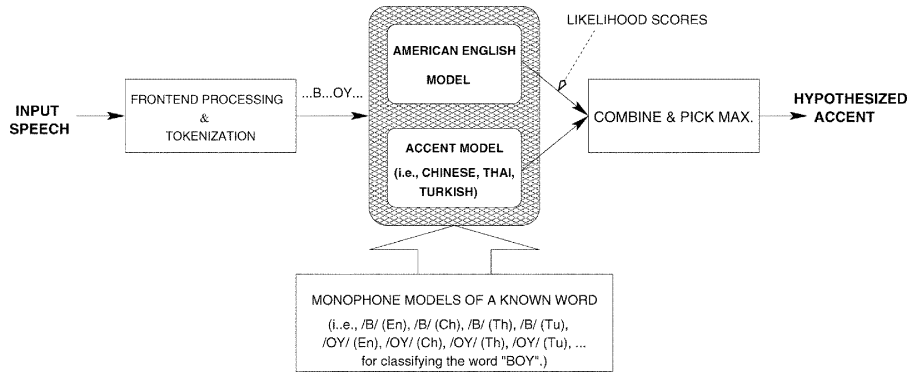

Fig. 4. Framework for the accent classification algorithm.

Let $\vec{\mathbf{x}}_k^{\{\tau\}}$ be a D-dimensional vector with time-constraint τ from token k as described in the previous section, and $U^{\{\tau\}}$ a $D \times M$ transform matrix. The M-dimensional transform vector is then expressed as $U^{\{\tau\}T}\vec{\mathbf{x}}_k^{\{\tau\}}$. The transformation is defined according to the usual criterion which maximizes $tr(\mathbf{W}^{\{\tau\}-1}\mathbf{B}^{\{\tau\}})$, where $\mathbf{W}^{\{\tau\}}$ and $\mathbf{B}^{\{\tau\}}$ are the within and between class covariance matrices, at time constraint τ, defined as

$$\mathbf{B}^{\{\tau\}} = \frac{1}{N}\sum_{k=1}^{K} n_k(\mu_k^{\{\tau\}} - \mu^{\{\tau\}})(\mu_k^{\{\tau\}} - \mu^{\{\tau\}})^T \quad (15)$$

$$\mathbf{W}^{\{\tau\}} = \frac{1}{N}\sum_{k=1}^{K} n_k \sum_{n=1}^{n_k}(x_{kn}^{\{\tau\}} - \mu_k^{\{\tau\}})(x_{kn}^{\{\tau\}} - \mu_k^{\{\tau\}})^T \quad (16)$$

where N denotes the total number of training tokens, K the number of classes (i.e., accents), and n_k the number of training tokens of the k^{th} class. The mean $\mu_k^{\{\tau\}}$ class and the overall mean $\mu^{\{\tau\}}$ are given by

$$\mu_k^{\{\tau\}} = \frac{1}{n_k}\sum_{n=1}^{n_k} x_{kn}^{\{\tau\}} \quad (17)$$

$$\mu^{\{\tau\}} = \frac{1}{N}\sum_{k=1}^{K} n_k \mu_k^{\{\tau\}} \quad (18)$$

where $x_{kn}^{\{\tau\}}$ is the n^{th} training token from the k^{th} class. It should be noted that there are at most $K-1$ nonzero generalized eigenvectors, so the upper bound on M is $K-1$. In addition, we require at least $D+K$ samples to guarantee that $\mathbf{W}^{\{\tau\}}$ does not become singular. If $W^{\{\tau\}}$ is a nonsingular matrix, then the column vector of the projection matrix, $U^{\{\tau\}}$ are the eigenvectors of $\mathbf{W}^{\{\tau\}-1}\mathbf{B}^{\{\tau\}}$. This completes our formulation of the LDA subspace accent partitioning process. Next, we turn to database and classifier formulation.

IV. CU-Accent Database

The CU-Accent corpus [36] was organized and collected at CSLR for algorithm formulation in acoustic-phonetic studies in automatic accent classification and speaker identification. The corpus consists of 179 speakers (72 male, 107 female) that belong to a variety of accent groups, such as neutral American English, Chinese, French, Thai, Turkish, and others. The majority of the speakers are from the Boulder, CO community who have moved there within a 1–3-year period. Each speaker was asked to pronounce isolated words, a few short sentences in English and their native languages, and one minute of spontaneous speech focused on a specific topic. The speakers were motivated to make multiple calls with at least one day separation between each call if possible (more than 50% of the entire database contain speakers with multiple session calls). The corpus was recorded in 2001 using an ATT DSP32 based Gradient/DeskLab216 unit, connected via an ISDN telephone line and SUN UNIX environment. The original corpus was digitized at 8000 Hz and stored in 16-bit linear PCM format without headers in the speech file (i.e.,raw audio format). This corpus is similar in structure to the earlier corpus that we collected in 1996 [1], but was extended in a number of ways including format, multiple calls for session-to-session variability, and adding a section with one-minute of spontaneous speech.

V. Baseline Classification System

A baseline Markov model based classifier is formulated in order to investigate trajectory modeling and subspace feature analysis. A phone-based accent classifier is established that assumes knowledge of the word sequence, but the system is capable of handling any word sequences for an accumulated accent score. Here, we assume the speaker gender to be known *a priori*. In general, front-end gender discrimination using Gaussian mixture models (GMM) is extremely effective [33], [37].

A flow diagram of the proposed accent classification system is shown in Fig. 4. Acoustic feature extraction is performed on sampled data on a frame-by-frame basis, and pre-emphasized by a first-order FIR filter, $1 - 0.97z^{-1}$. Hanning window analysis is applied to frames that are 20 ms in duration with a 10-ms overlap between successive frames. Next, 12 mel frequency cepstral coefficients (MFCCs) and normalized log-frame energy are computed per frame. The final vector $\vec{\mathbf{x}}_i$ includes these features and their first-order time derivatives, resulting in a 26-dimensional vector per frame.

During training, each speech utterance is converted into a stream of feature vectors, and then tokenized into a set of phone sequences using automatic forced-alignment. By relying on automatic alignment, our system is portable and practical. The set of continuous-density HMM-based phone models are trained with each accented English speech set. Each continuous-density HMM uses five states. The emission probabilities are modeled by Gaussian mixture models with two mixtures per state. The transition probabilities are not trained, but instead are given fixed *a priori* values.

The classification stage is based on the accumulation of normalized log-likelihood scores produced by accent-dependent phoneme models, that require no high-level language modeling information.

During the classification phase, each test speech utterance is also converted into feature vectors and then tokenized into a sequence of phone features. The Bayes's decision rule is used to determine the accent class based on the accumulated accent likelihood scores computed from each accent phone-based HMM in the phoneme sequence. The estimated log-likelihood score per frame from a Viterbi decoder is used to minimize the effect of different phoneme durations in the scores.

A. Bayes Classification and Log-Likelihood Scoring

Our reference classification method employs a ML classifier (i.e., a well-known Bayes classifier) with equal *a priori* class probabilities. The Bayes classifier uses the Bayes decision rule to determine the class for the present data. The objective in Bayes decision theory is to minimize the probability of decision error. Let A_i represent the accents(classes) and $\mathbf{X}$ a sequence of feature vectors for one phoneme. The decision rule can be stated as

$$\text{Decide accent } A_i \text{ if } P(A_i|\mathbf{X}) > P(A_k|\mathbf{X}), \text{ for all } k \neq i,$$
$$\text{when the } a\,posteriori \text{ probability } P(A|\mathbf{X})$$
$$\text{can be calculated using Bayes rule.}$$

The classifier assigns feature vector $\mathbf{X}$ to accent A^* such that: $A^* = \arg\max_{A_i} P(A_i)P(\mathbf{X}|A_i)$, when $P(\mathbf{X}|A_i)$ is the log-likelihood score per frame estimated from a Viterbi decoder. Having established our classifier and scoring strategies, we now turn to experimental evaluations.

VI. Experiments

In our evaluation, we will consider how effective accent classification is for 1) pairwise native/nonnative; 2) four-way accent conditions; 3) human performance; and 4) trajectory feature based classification.

A. Baseline System

In order to establish the performance and characteristics of our baseline classification system, we conducted some exploratory experiments. It is important to note that for LVCSR, it is not uncommon to have several thousand training and test utterances from many speakers. The accent database used here represent one of the largest available, but there are still some limitations. Due to the relatively small vocabulary size, we used 36 context-independent phone models for each accent and ignored silence between speech sections. All training and testing speech materials were obtained from isolated words and short sentences, except otherwise noted (phonemes that were not considered include /ah, ao, el, ix, ow, z, zh/). Ten speakers were chosen for each class(i.e., 80 speakers employed total): native American-English male, native American-English female, Chinese male, Chinese female, Thai male, Turkish male, Turkish female, and French female. The experiments were conducted in a round-robin fashion, five-fold cross validation. For each gender-dependent experiment, acoustic models were trained from eight speakers and tested on the remaining two speakers. On average, each speaker has approximately 80-s worth of speech (note that each word is between 620–1150 ms in duration.) The classification results reported are the average of all five rounds, except otherwise specified.

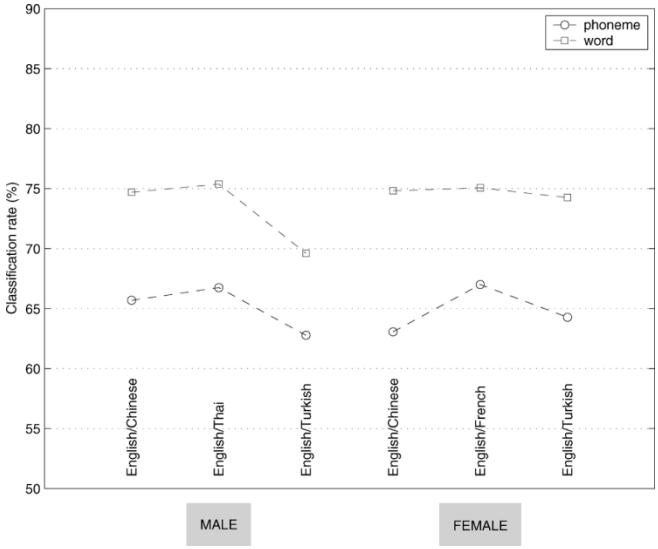

Fig. 5. Pairwise classification accuracy rates.

1) Pairwise Classification: In this evaluation, the classifier makes a binary decision between two hypothesis accents consisting of native American-English and a second foreign accent (i.e., neutral versus Chinese, neutral versus Thai, and neutral versus Turkish). Pairwise classification can be considered as a specific case of accent detection where the competing accent class is known *a priori*. Fig. 5 shows pairwise classification experimental performance using single phoneme testing, and single word testing. From this figure, the baseline performance shows consistent accent discrimination for both genders, and a consistent increase in discrimination when we move from a single phoneme test to an isolated word test for all pairs (on the average, each word contains 4–5 phonemes). The increase for males from individual phoneme based accent classification to word based classification is 8.16%, and 12.39% for females. Female performance in general was slightly higher when compared to males for the same accent. For completeness, we

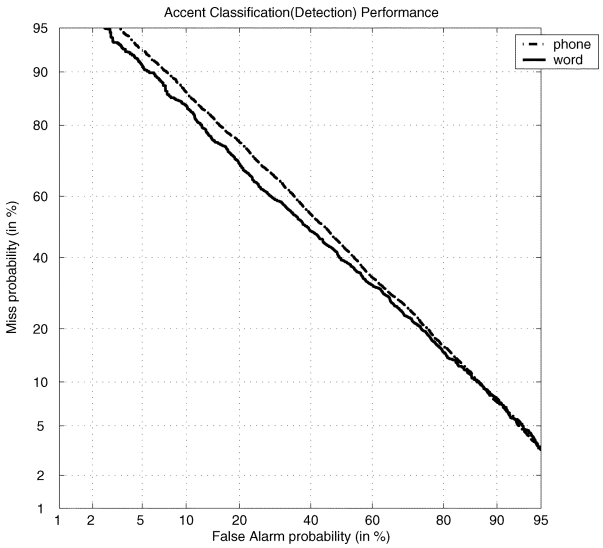

Fig. 6. Accent detection performance for classifying Chinese accent against American English accent (Male speaker group).

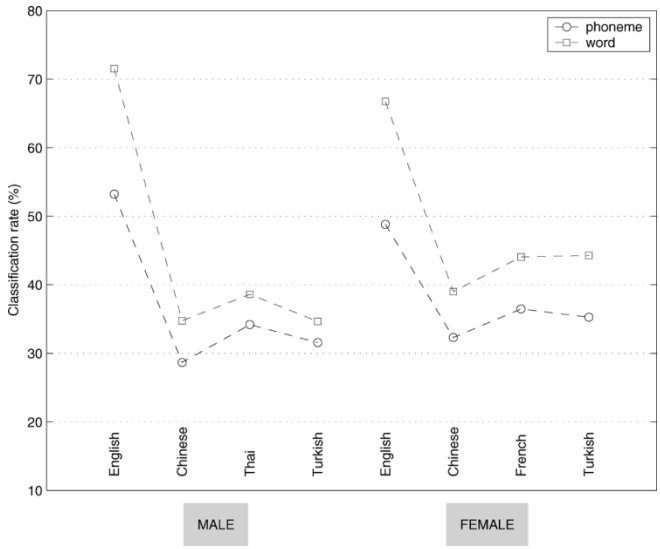

Fig. 7. Four-way accent classification performance.

also illustrate in Fig. 6, Detection error tradeoff (DET) [25][3] plots for a sample set of accents from Fig. 5. We see that the resulting curves are approximately straight lines, corresponding to the assumed normality of the likelihood distributions. If the performance is reasonably good, we limit the curves to the lower left quadrant. Clearly, word accent performance always outperforms single phoneme level performance.

2) Four-Way Classification: In this experiment, on a per gender basis, accent classification is performed over a pool of four accents. Fig. 7 shows four-way classification performance using single phonemes and single words for each accent. While the figure shows similar rates for native American-English and Turkish speakers, there is a significant difference in the overall average classification performance across accents between males and females (44.86% for male, and 51.60% for female; note that 25% is chance). Unlike the pairwise classification task, four-way classification allows more choices of hypothesized accents in cases where we are not certain of a binary decision scenario. In this experiment, results point to inferior performance for the male group where confusion was high between Chinese and Thai accents versus the confusion between French and Chinese accents in the female group. We hypothesize that these results occur because Chinese and Thai speakers share more similar articulatory movements when they speak English than French speakers. In this domain, we also wanted to explore the potential for accent classification in the limit as extensive test data is accumulated. Fig. 8 show examples of accumulated probability scores of one Thai male and one French female using 10-seconds of open test spontaneous speech. The plot illustrates the accumulation of the classifier score difference generated by a speaker's accent models against the competitive accent models. The simulated score space showed that the distance between Thai and Mandarin Chinese accents has the least separation compared to English and Turkish accents as the number of accumulated test phonemes increases, while the distance between French accent and Mandarin Chinese accent is well separated compared to English and Turkish accents. These results suggest what the accent pairwise performance would be in the limit as more speech is accumulated (i.e., we should expect confusion between Thai and Chinese accented speakers, and less confusion for other pairwise sets).

For our baseline system, the average classification accuracy rate is 64.90% at the phone level, and increases to 75.18% at the word level for pairwise classification. For a pool of four accents, the averaged classification accuracy rate is 37.57% at the phone level, and 46.72% at the word level. These results suggest a baseline system in which to formulate more effective accent sensitive features and classifiers. In theory, there is no limit to the number of speech segments that could be used in a sequence for discriminating between accents. In practice, as the number of speech segments increases, the discriminant performance of the classifier will also increase. However, this might not be true for those accents which have a strong degree of similarity in their linguistic structure. An evaluation such as that shown in the results in Fig. 8 will establish how discriminating a pair of accents will be in the limit, and if it may be more profitable to group two or more accents into a single class.

3) Human Perceptual Comparison: This section briefly describes a human perception study which was conducted on two native American Speakers, and six 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 from five speakers from each accent. This listener evaluation was performed in three phases: human accent classification based on: 1) one-word; 2) two-word; and 3) three-word sequences. On the average, each isolated word contained four phonemes from our vocabulary set(i.e., on average, human performance was based on phoneme strings of length 4, 8, and 12). Each listener was able to listen to each test token multiple times

[3]The DET curve form of presentation is relevant to any detection task where a tradeoff of error types, false alarm probability, and miss probability, is involved. The DET curve has distinct advantages over the standard ROC type curve for presenting performance results in speaker and language recognition evaluations.

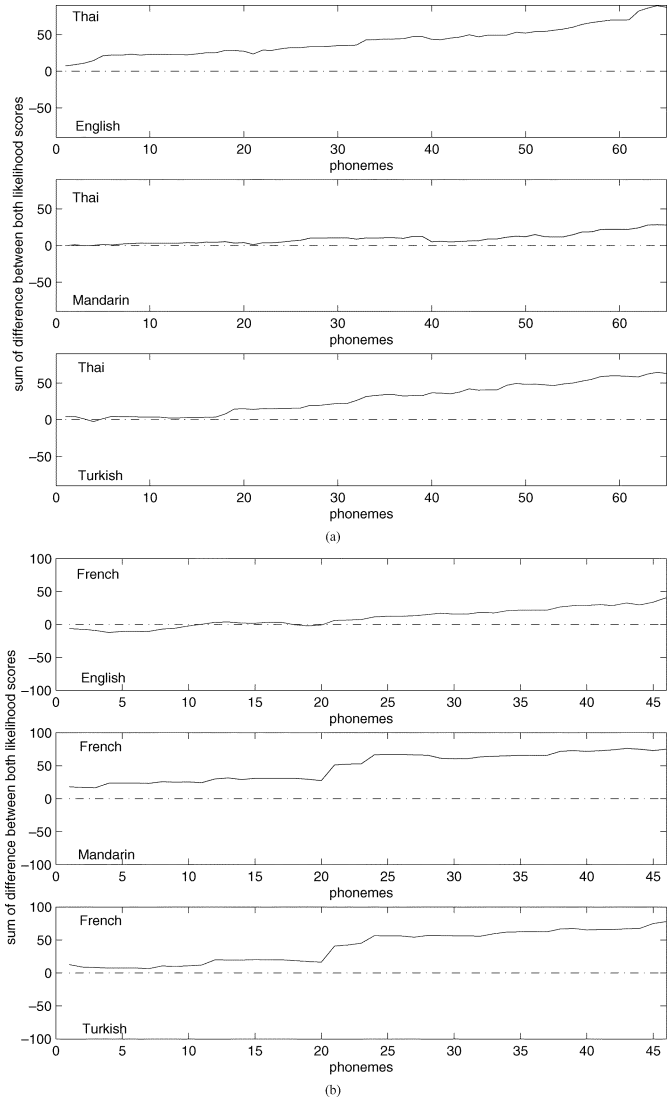

Fig. 8. Accumulated summation of different likelihood scores between two accents with up to 10-s worth of speech. (a) Thai. (b) French.

before making their decision. No listener reported any history of hearing loss, and each listener was able to adjust the listening volume to a comfortable level. The text word list was kept the same for human listener tests across all accents. Fig. 9 shows the resulting human classification performance. We see that human classification performance is higher for Chinese and Thai accents(i.e., 87%–90% with 2–3 words) compare to Turkish accent(i.e., 63%–74% with 2–3 words), which is consistent with our automatic baseline classifier performance. A general observation was that with Thai and Chinese, human accent classification performance typically improved when we increased the human test material size from one to three words. The results for Turkish was unusual since results decreased with more data (two words) before returning back to single word performance at the three-word test duration size. We believe that these results point to similar and consistent traits which listeners detect and track when making their subjective decision as to accent content. Again, performing such a study, we feel offers insight as to what type of performance we might expect for automatic classification.

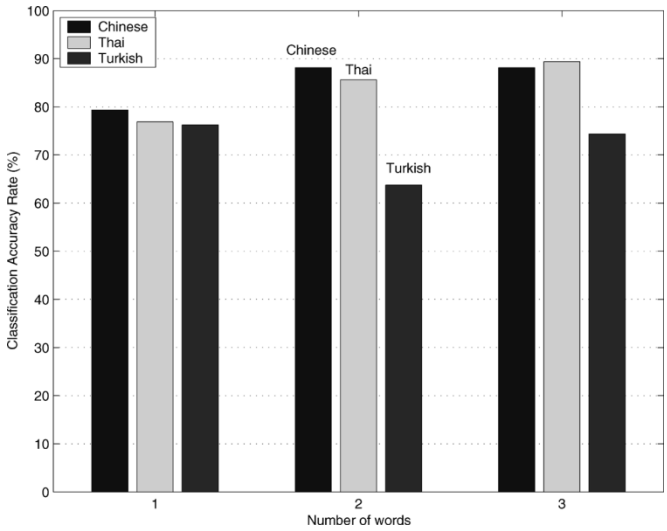

Fig. 9. Human perceptual rates.

B. Trajectory-Model Based Classification

To evaluate the trajectory-based accent classification system, we performed our classification experiments on vowels, semi-vowels, and diphthongs. The task includes the following 18 phonemes: /iy, ih, eh, ae, aa, er, axr, ax, uw, uh, ay, oy, aw, ey, w, l, r, y/ which are considered perceptually important and represent information-rich portions of the speech signal.[4] From our experiments, by employing acoustic models, these phonemes play an important role for discriminating accent. Due to the small diversity of phoneme sequences in our database and avoiding sparseness during the training stage, we trained STM and PTM feature based classifiers with two trajectory mixtures in each case for each gender. For a fair comparison, the CD-HMM was also trained with two mixtures per state. We conjecture that two Gaussian mixtures for each HMM also represents the global accent traits for a small database. Speech was parameterized using 13-dimensional MFCCs (energy included). Figs. 10 and 11 show the pairwise classification performance employing three acoustic models using a decision with single versus phoneme sequence string (i.e., 1, 5, 11, and 17 phonemes used per decision). Increasing the numbers of phonemes was obtained from random concatenation of different isolated words. From Figs. 10 and 11, STM showed the best performance among the three acoustic model types, except for the Turkish/English male speaker group, which consistently showed a lower level of performance versus all other accents. For the best cases, STM could classify the English/Chinese pair and English/French pair for the female speaker with up to 90% performance after 17 phonemes. On the average, PTM and HMM showed comparable discrimination performance. As the likelihood accumulated scores of phoneme sequences are combined in obtaining the decision, all classifiers showed improved performance and in general converged in their levels of classification.

Next, a further set of experiments using a subspace representation or dimensionality reduction was conducted. Here, the

[4]Note that for some speech tasks such as speaker recognition, low energy consonants such as stops and fricatives are often discarded.

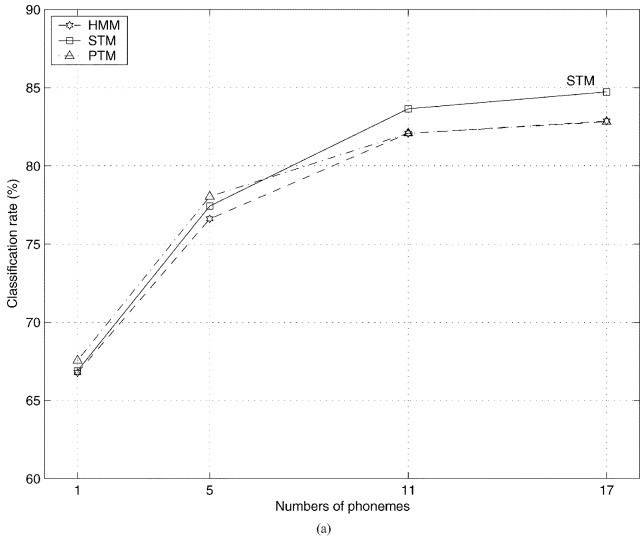

Fig. 10. Accent classification accuracy rates versus number of phonemes for STM: Stochastic trajectory model, HMM: Hidden Markov model, and PTM: Parametric trajectory model. (a) Male: Chinese versus American-English. (b) Male: Thai versus American-English. (c) Male: Turkish versus American-English.

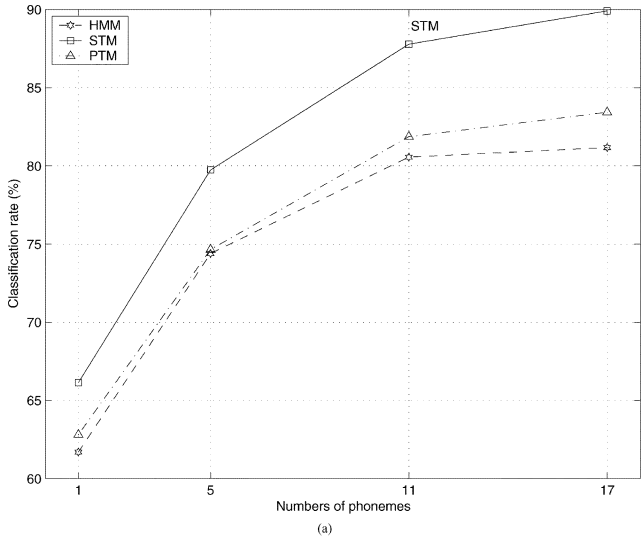

Fig. 11. Accent classification accuracy rates versus number of phonemes, for STM: Stochastic trajectory model, HMM: Hidden Markov model, and PTM: Parametric trajectory model. (a) Female:Chinese versus American-English. (b) Female:French versus American-English. (c) Female:Turkish versus American-English.

feature space was originally 13 and we explored performance with dimensionality reduction to three dimensions. Table I shows the accent classification accuracy rates among four accents at the phoneme level. Again, STM(13) and PTM(13)

TABLE I
ACCENT CLASSIFICATION ACCURACY RATES WITH SUBSPACE PROCESSING USING PHONEME-LEVEL CLASSIFICATION

classification accuracy rate (%)									
	Male: Four-way				Female: Four-way				OVERALL AVG.
	EN	CH	TH	TU	EN	CH	FR	TU	
STM(13)	50.22	30.43	36.29	35.48	49.08	34.03	41.92	38.55	39.50
STM–PCA(3)	41.01	26.15	32.02	31.68	43.84	25.39	38.91	38.47	34.68
STM–PCA(13)	42.25	30.49	**38.67**	31.33	47.55	36.58	40.46	40.41	38.47
STM–LDA(3)	50.96	30.63	36.12	34.35	52.37	**35.68**	39.81	40.33	40.03
PTM(13)	49.23	**30.67**	31.10	**40.55**	43.53	27.20	**44.69**	**49.03**	39.50
PTM–PCA(3)	50.27	24.98	31.89	25.40	48.54	24.65	43.20	34.68	35.45
PTM–PCA(13)	35.65	25.21	37.09	37.68	37.64	29.52	40.70	46.54	36.25
PTM–LDA(3)	**53.32**	29.27	36.07	26.64	**58.93**	34.21	36.65	40.05	40.64

represents the 13-dimensional trajectory based models from the original feature space discussed above, while PCA(3), PCA(13), and LDA(3) represents the subspaces with dimensionality reduction. For subspace representation, we added delta-parameters to the original features, resulting in 26-dimensional feature space, and then applied projection analysis into new subspaces. The particular transformation matrix was trained as discussed in Section III and applied to each individual phoneme. PCA(3) and PCA(13) represents the first three and 13 principal components respectively. For LDA(3), we transformed 26 feature space into the first three linear discriminant feature subspace, since the total number of accent classes for each phoneme was four. Considering both transformation techniques, LDA outperformed PCA in both 13- and 3-dimensional spaces [i.e., PCA(3) and PCA(13)]. While dimensionality transformation using PCA processing showed a small degradation in performance for most tasks, LDA preserved the discrimination performance and showed small improvements for female speakers ($+2.81\%$ for STM, and $+3.28\%$ for PTM).

C. Open-Accent Detection Performance

Our last experiment was to study how well the classifier can detect open accented speech from a limited set of trained accent models or prototype models, using a subspace representation. The experimental setup was the same as in the previous section. However, in this case accent likelihood score was computed from the collection of all prototype accent models, where STM models are used. Here, we focus on accent detection, and not accent classification. Therefore, we use an English STM model for native, and three STM models (e.g., MALE: Chinese, Turkish, Thai) grouped together for the accent decision. If any of the three accent STMs are selected, the speech is labeled as accented, otherwise the American-English model is selected and the speech is native. Our testing here will consist of closed accent-set conditions (e.g., MALE: Chinese, Turkish, Thai), as well as open accent-set conditions (i.e., accents the models have not seen). The maximum score is used as the accent match score for the input speech. The classifier decides the input as being accented speech if the accent match score is higher than the neutral score computed from the American-English accent model. The first part of experiment considered accented speech for which we have their accent models in the collection, referred to as closed accent-set detection. Next, detection was performed on four unseen (open) accents (MALE: Dutch, French, German, and Spanish, FEMALE: Arabic, Thai, German, and Spanish), using two speakers for each group. Our underlying assumption was that the collection of closed trained accent groups could represent the shift of coarticulation from neutral speech to nonneutral speech, especially when the first languages of the speaker and the prototype accents in the collection share the same branches in the world language tree.

TABLE II
OPEN ACCENT DETECTION PERFORMANCE WITH SUBSPACE REPRESENTATION, AT PHONEME LEVEL

Accent Detection Performance (%)					
	Male			Female	
	LDA(3)	PCA(3)		LDA(3)	PCA(3)
Closed accent group					
Chinese	80.52	76.60	Chinese	78.42	77.78
Thai	82.45	78.03	French	83.63	79.24
Turkish	75.77	76.82	Turkish	81.78	80.87
Open accent group					
Dutch	85.39	74.99	Arabic	63.87	69.70
German	81.82	69.01	German	79.29	80.10
Spanish	83.83	75.92	Spanish	73.83	66.59

From Table II, the average detection performance of the closed accent-set was 80.43% for LDA, and 78.23% for PCA (note: all speakers and tokens are open in the test). For the open accent-set, the average detection performance was 78.75% for LDA, and 73.95% for PCA. While LDA performed slightly better than PCA for the closed accent-set group, LDA outperformed PCA measurably on the open accent-set group. We believe that LDA improves the discrimination between classes in the subspace, and that resulting space would therefore be better able to discriminate between native and other nonnative speech classes as well. The results in Table II are quite impressive, since for males, the unseen open-set accents achieve similar performance to the closed accent group. There is a slight reduction, however for females. The Arabic accent case showed the lowest detection rate among all accent classes. The results for the Arabic accent suggest that languages which are close to each other in the world language tree might achieve similar accent detection performance if neighbors are included in the closed accent-set base models. The use of Chinese, Turkish, and Thai for males, and Chinese, Turkish, and French for females, show only slight differences in detection performance when

seeking out open accent classes. On the average, there was only a 1.68% reduction in accent detection performance using STM-LDA modeling versus a 4.28% reduction with STM-PCA modeling.

VII. Conclusion

In this study, we have considered an approach to automatically classify accent from speech. The proposed classifiers were based on STM and PTM trajectory models, and were intended to capture the speech production based spectral evolution caused by coarticulation. Several experiments were conducted to study performance of a baseline phone-based HMM classifier and the new classifier schemes. Evaluations were also compared with human accent classification performance in an effect to understand ground-truth in a speaker's variable level of accent information. Experimental evaluations illustrated that the likelihood scores produced by an accent-specific phoneme classifier can be used to discriminate the accents of speech, and accent sensitive traits can be captured by such trajectory models. In general, trajectory-based classifiers showed better discriminating performance as likelihood scores are accumulated with more duration based test data. Using a set of concatenated isolated words, the best accent classification performance can approach 90% with an STM framework for a pairwise classification task (50% is chance). Alternative subspace feature partitioning based on PCA and LDA analysis was also studied within unsupervised and supervised frameworks. No significant consistent improvement was found after PCA projecting to a reduced dimensionality, but classification performance was preserved and slightly improved using LDA. Finally, it was shown that using an LDA based subspace accent detection framework could detect a set of unseen accents by up to 78.73% (50% is chance), at the phoneme level, from a limited set of prototype accents.

Automatically identifying an accent from only acoustic data is a challenging problem. We must recognize that individual speaker variability can influence accent model development. For a given set of speakers and/or accents, it may not be possible to achieve perfect accent classification performance because speakers may not consistently portray accent sensitive structure, or may have acquired sufficient L2 language skills so they do not convey their accent trait (i.e., the speaker sounds like a Chinese accented speaker part of the time, or has mild versus heavy accent). Compared with high quality speech, telephone speech has reduced bandwidth and is more noisy, thereby increasing the difficulty in reliable accent discrimination. Also, the task or goal of accent classification must consider the following problems: 1) the segmentation process; 2) suitable populations of speakers for statistical estimation of the segmental durations, together with all model parameters; and 3) features must be effective for discriminating accents. We recognize that while this work has contributed to improve accent classifier algorithm development, other features could also be considered (i.e., prosodic-based pitch structure and energy contour, specific formant location shifts, etc.) [2]. We feel that future studies might consider incorporating the contributions here with additional speech production features.

References

[1] L. M. Arslan and J. H. L. Hansen, "Language accent classification in American English," *Speech Commun.*, vol. 18, pp. 353–367, 1996.

[2] ——, "A study of temporal features and frequency characteristics in American English foreign accent," *J. Acoust. Soc. Amer.*, vol. 102, no. 1, pp. 28–40, Jul. 1997.

[3] A. Batliner *et al.*, "Boiling down prosody for the classification of boundaries and accents in German and English," in *Proc. EUROSPEECH*, Aalborg, Denmark, 2001.

[4] K. Berkling, "Scope, syllable core and periphery evaluation: Automatic syllabification and foreign accent identification," *Speech Commun.*, vol. 35, pp. 125–138, 2002.

[5] C. S. Blackburn, J. P. Vonwiller, and R. W. King, "Automatic accent classification using artificial neural networks," in *Proc. EUROSPEECH*, 1993, pp. 1241–1244.

[6] J. E. Flege, "The selection of French accent by American listeners," *J. Acoust. Soc. Amer.*, vol. 76, pp. 692–707, 1984.

[7] ——, "Factor affecting degree of perceived foreign accent in English sentences," *J. Acoust. Soc. Amer.*, vol. 84, pp. 70–77, 1988.

[8] T. Fukada, Y. Sagisaka, and K. K. Paliwal, "Model parameter estimation for mixture density polynomial segment models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1997, pp. 1403–1406.

[9] S. Furui, "Speaker independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 1, pp. 52–59, Jan. 1986.

[10] P. J. Ghesquiere and D. V. Compernolle, "Flemish accent identification based on formant and duration features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, 2002, pp. 749–752.

[11] H. Gish and K. Ng, "Parametric trajectory models for speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, vol. 1, 1996, pp. 466–469.

[12] W. Goldenthal, "Statistical Trajectory Models for Phonetic Recognition," Ph.D. dissertation, Dept. Aeronaut. Astronaut., Mass. Inst. Technol., Cambridge, 1994.

[13] Y. Gong, "Stochastic trajectory modeling and sentence searching for continuous speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 1, pp. 33–44, Jan. 1997.

[14] C. Grover, D. G. Jamieson, and M. B. Dobrovolsky, "Intonation in English, French, and German: Perception and production," *Lang. and Speech*, vol. 30, no. 3, pp. 277–295, 1987.

[15] V. Gupta and P. Mermelstein, "Effect of speaker accent on the performance of a speaker-independent, isolated word recognizer," *J. Acoust. Soc. Amer.*, vol. 71, pp. 1581–1587, 1982.

[16] J. H. L. Hansen, "Adaptive source generator compensation and enhancement for speech recognition in noisy stressful environments," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Apr. 1993, pp. 95–98.

[17] ——, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Commun.*, vol. 20, pp. 151–173, Nov. 1996.

[18] J. H. L. Hansen and L. M. Arslan, "Foreign accent classification using source generator based prosodic features," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, 1995, pp. 836–839.

[19] K. Kumpf and R. W. King, "Automatic accent classification of foreign accented Australian English speech," in *Proc. Int. Conf. Spoken Language Processing*, 1996, pp. 1740–1743.

[20] ——, "Foreign speaker accent classification using phoneme-dependent accent discrimination models and comparison with human perception benchmarks," in *Proc. EUROSPEECH*, vol. 4, 1997.

[21] M. Lincoln, S. Cox, and R. Ringlind, "A comparison of two unsupervised approaches to accent identification," in *Proc. Int. Conf. Spoken Language Processing*, vol. 2, 1998, pp. 109–112.

[22] R. C. Major, *Foreign Accent*. Mahwah, NJ: Lawrence Erlbaum, 2001.

[23] N. Malayath, H. Hermansky, and A. Kain, "Toward decomposing the sources of variability in speech," in *Proc. EUROSPEECH*, vol. 1, 1997, pp. 497–500.

[24] S. Man-Hung, R. Iyer, H. Gish, and C. Quillen, "Parametric trajectory mixtures for LVCSR," in *Proc. Int. Conf. Spoken Language Processing*, 1998.

[25] A. Martin, G. Doddington, T. Kamn, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Sep. 1997, pp. 1895–1898.

[26] J. McGory, E. Frieda, S. Nissen, and R. A. Fox, "Acquisition of dialectal differences in English by native Japanese speakers," *J. Acoust. Soc. Amer.*, vol. 109, no. 5, 2001.
[27] L. Neumeyer, H. Franco, M. Weintraub, and P. Price, "Automatic text-independent pronunciation scoring of foreign language student speech," in *Proc. Int. Conf. Spoken Language Processing*, 1996, pp. 1457–1460.
[28] M. Ostendorf and S. Roukos, "A stochastic segment model for phoneme-based continuous speech recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 12, pp. 1857–1869, Dec. 1989.
[29] L. R. Rabiner and J. G. Wilpon, "Speaker-independent isolated word recognition for a moderate size (54 word) vocabulary," in *Proc. IEEE Trans. Conf. Acoustics, Speech, Signal Processing*, vol. 27, 1977, pp. 583–587.
[30] K. Reinhard and M. Niranjan, "Parametric subspace modeling of speech transitions," *Speech Commun.*, vol. 27, no. 1, pp. 19–42, 1999.
[31] T. Schultz, Q. Jin, K. Laskowski, A. Tribble, and A. Waibel, "Speaker, accent, and language identification using multilingual phone strings," *HLT*, 2002.
[32] J. Vonwiller, C. Blackburn, and R. King, "Automatic accent classification using artificial neural networks," in *Proc. EUROSPEECH*, vol. 2, 1993, pp. 1241–1244.
[33] B. Zhou and J. H. L. Hansen, "SpeechFind: An experimental on-line spoken document retrieval system for historical audio archives," in *Proc. Int. Conf. Spoken Language Processing*, 2002, pp. 1969–1972.
[34] M. A. Zissman, T. P. Gleason, D. M. Rekart, and B. L. Losiewicz, "Automatic dialect identification of extemporaneous conversational, Latin American Spanish speech," in *Proc. IEEE Conf. Acoustics, Speech, Signal Processing*, vol. 2, 1996, pp. 777–780.
[35] M. A. Zissman and K. M. Berkling, "Automatic language identification," *Speech Commun.*, vol. 35, pp. 115–124, 2001.
[36] University of Colorado. (2005). Tech. Rep. [Online]. Available: http://crss.utdallas.edu/accent/
[37] J.H.L. Hansen, R. Huang, B. Zhou, M. Seadle, J.R. Deller, Jr., A.R. Gurijala, and P. Angkititrakul, "SpeechFind: Advances in spoken document retrieval for a national gallery of the spoken word," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 5, pp. 712–730, Sep. 2005.

Pongtep Angkititrakul (S'04–M'05) was born in Khonkaen, Thailand. He received the B.S.E.E. degree from Chulalongkorn University, Bangkok, Thailand, in 1996 and the M.S. degree in electrical engineering from the University of Colorado at Boulder in 1999, where he is currently pursuing the Ph.D. degree in electrical engineering.

He was a Research Assistant with the Center for Spoken Language Research (CSLR), University of Colorado, from 2000 to 2004. He was with Eliza Corporation, Beverly, MA, from 2005 to 2006. In February 2006, he joined The Center for Robust Speech Systems, University of Texas at Dallas, Richardson, TX.

John H. L. Hansen (S'81–M'82–SM'93) was born in Plainfield, NJ. He received the B.S.E.E. degree (with highest honors) from Rutgers University, New Brunswick, NJ, in 1982 and the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1983 and 1988, respectively.

He joined the University of Texas at Dallas (UTD), Richardson, in Fall 2005, where he is Professor and Department Chairman of Electrical Engineering, and holds the Endowed Chair in Telecommunications Engineering in the Erik Jonsson School of Engineering and Computer Science. He also holds a joint appointment in the School of Brain and Behavioral Sciences (Speech and Hearing). At UTD, he established the Center for Robust Speech Systems (CRSS), which is part of the Human Language Technology Research Institute (HLTRI). Previously, he served as Department Chairman and Professor with the Department of Speech, Language, and Hearing Sciences (SLHS), and Professor with the Department of Electrical and Computer Engineering, University of Colorado, Boulder, from 1998 to 2005, where he co-founded the Center for Spoken Language Research. In 1988, he established the Robust Speech Processing Laboratory (RSPL) and continues to direct research activities in CRSS at UTD. He was a Faculty Member with the Departments of Electrical and Biomedical Engineering of Duke University, Durham, NC, 11 years before joining the University of Colorado in 1999. He has served as a Technical Consultant to industry and the U.S. Government, including ATT Bell Labs, IBM, Sparta, Signalscape, BAE Systems, ASEC, VeriVoice, Audience, HRL, and DoD in the areas of voice communications, wireless telephony, robust speech recognition, and forensic speech/speaker analysis. He has also served as a Technical Advisor to the U.S. Delegate for NATO (IST/TG-01: Research Study Group on Speech Processing, 1996 to 1999). He has supervised 33 Ph.D./M.S./M.A. thesis candidates, and is the author/co-author of 209 journal and conference papers in the field of speech processing and communications, co-author of the textbook *Discrete-Time Processing of Speech Signals* (New York: IEEE Press, 2000), and co-editor of *DSP for In-Vehicle and Mobile Systems* (Norwell, MA: Kluwer, 2004). His research interests span the areas of digital speech processing, analysis and modeling of speech and speaker traits, speech enhancement and feature estimation in noise, robust speech recognition with current emphasis on robust recognition and training methods for spoken document retrieval in noise, accent, stress/emotion, and Lombard effect, and speech feature enhancement in hands-free environments for human-computer interaction.

Dr. Hansen received the Whitaker Foundation Biomedical Research Award, the National Science Foundation Research Initiation Award, and was named a Lilly Foundation Teaching Fellow for "Contributions to the Advancement of Engineering Education." He was an Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1992–1998), Associate Editor for the IEEE SIGNAL PROCESSING LETTERS (1998–2000), and a Member of the Editorial Board for the *IEEE Signal Processing Magazine* (2001–2003). He also served as Guest Editor of the Special Issue on Robust Speech Recognition for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING . He was an invited Tutorial Speaker for the IEEE ICASSP-95, the 1995 ESCA-NATO Speech Under Stress Research Workshop (Lisbon, Portugal), and the 2004 IMI-COE Symposium (Nagoya, Japan). He served as Chairman for the IEEE Communication and Signal Processing Society of North Carolina (1992–1994), Advisor for the Duke University IEEE Student Branch (1990–1997), Tutorials Chair for IEEE ICASSP-96. He organized and served as the General Chair for ICSLP-2002: International Conference on Spoken Language Processing. He is presently serving as an IEEE Signal Processing Society Distinguished Lecturer (2005–2006) and is a member of the IEEE Signal Processing Society Speech Technical Committee (2006–2008).