PHONETIC LANDMARK DETECTION FOR AUTOMATIC LANGUAGE IDENTIFICATION

David Harwath and Mark Hasegawa-Johnson

University of Illinois at Urbana-Champaign
Department of Electrical and Computer Engineering
Urbana, IL 61801

ABSTRACT

This paper presents a method of augmenting shifted-delta cepstral coefficients (SDCCs) with the classification outputs of an array of support vector machines (SVMs) trained to detect a set of manner and place features on telephone speech. The SVM array allows for broad phoneme classification, and when this information is concatenated with SDCCs to form a hybrid feature vector for each acoustic frame, a set of Gaussian mixture models (GMMs) may be trained to perform automatic language identification (LID). The NTIMIT telephone band speech corpus was used to train the SVM-based distinctive feature recognizers, while the NIST callfriend telephone corpus was used for training and testing the rest of the system.

***Index Terms*—** Support Vector Machines, Gaussian Mixture Models, Distinctive Features, Language Identification

1. INTRODUCTION

The task of language identification (LID) is to determine what language a speaker is speaking given a recorded utterance. What makes this problem interesting is the number of levels at which languages differ, giving rise to a large number of algorithmic approaches to LID. Navratil distinguishes between four basic classes of LID systems: acoustic, phonotactic, prosodic, and LVCSR. Of these, acoustic and phonotactic are the most common approaches. Acoustic systems rely solely on features extracted from the input waveform, while phonotactic systems typically perform n-gram modeling to capitalize on linguistic differences between languages such as phoneme inventories and the frequencies of various phoneme sequences [1].

In this paper, we attempt to improve the classification accuracy of an acoustic baseline LID system with phonotactic information in the form of distinctive features. Distinctive features are a linguistic abstraction representing the building blocks of phonemes and are always binary, taking on a (+) or (-) value [2]. What makes distinctive features useful for our purposes is the fact that they have canonical acoustic properties, allowing systems to be trained to automatically detect the presence of a feature in a particular frame of speech. For any given speech frame, a vector of distinctive feature values can be estimated by means of SVMs and treated as a discriminant feature vector for a second level recognizer. This technique has been shown to perform well in phone recognizers [3]. For our purposes, the discriminant vector is concatenated with a separate vector of SDCCs (often used alone as the baseline in many state-of-the-art systems) and fed into a set of GMMs. The log-likelihood ratio outputs of each language-specific model versus a universal background model (UBM) are then used as a means of classifying an unknown input utterance.

2. FEATURES

2.1. DISTINCTIVE FEATURES

Linguistic distinctive features can be thought of as the building blocks of phones in the sense that any phone can be defined by its distinctive features. Distinctive features are always binary valued, taking on a (+) value when the feature is present, and a (-) value when it is not. They can be divided into two categories, manner and place. In our system, we are concerned with one manner feature and ten place features [2].

The manner feature we work with is the [sonorant] feature. Sonorance is essentially a measure of resonance in the vocal tract. Examples of [+sonorant] phones would be vowels, glides, and nasals.

Unlike manner features, place features are articulatory bound, meaning that they depend upon the location of the physical articulators in the vocal tract. We utilize the following features and describe their articulation:

Alveolar fricative: Sounds made by air flowing around the tounge when the tounge tip is pressed against the alveolar ridge. Example: /s/

Alveolar nasal: Sounds made by opening the velopharyngeal port when the tounge tip is pressed against the alveolar ridge. Example: /n/

Alveolar stop: Sounds made by the release of a closure in the vocal tract when the tounge tip is pressed against the alveolar ridge. Example: /d/

Interdental fricative: Sounds made by air flowing around

This work was supported by the National Science Foundation.

the tounge when the tounge body is between the teeth. Example: /th/

Labial nasal: Sounds made by opening the velopharyngeal port when the lips are pressed together. Example: /m/

Labial stop: Sounds made by the release of a closure in the vocal tract when the lips are pressed together. Example: /b/

Labiodental fricative: Sounds made by air flowing around the top teeth when they are pressed against the lips. Example: /f/

Postalveolar: Sounds made by pressing the tounge body against the palate behind the alveolar ridge. Example: /ch/

Retroflex: Sounds made by creating a resonant cavity in the mouth below the tounge. Example: /r/

Velar Stop: Sounds made by the closure and release of the glottis. Example: /k/

2.2. CEPSTRAL FEATURES

Mel-frequency cepstral coefficients (MFCCs) are a standard feature set in many speech recognition systems. They are computed by taking the Discrete Cosine Transform of the log magnitude spectrum of a signal, scaled logarithmically along the frequency axis into ”Mel” banks, which attempt to more closely match the human perception of pitch.

MFCCs are often used in phoneme recognizers, and have also been used in distinctive feature landmark detection [3] with high accuracy. In our system, we utilize MFCC features for the SVM distinctive feature detectors.

In single level GMM-based LID systems, Shifted Delta Cepstral Coefficients (SDCCs) have proved to be far more useful than MFCCs, as shown by [4]. SDCCs are computed by performing short-term differencing of MFCCs across many frames. In this way, SDCCs give a detailed representation of how MFCCs change over a period of time.

If we denote the MFCC vector at time t by $m(t)$, then the corresponding SDCCs, $s(t)$, are given by concatenating each $\Delta s(t)$ for $i = 0, 1, ...(k-1)$:

$$\Delta s(t) = m(t + iP + d) - m(t + iP - d) \tag{1}$$

Where d is the shift corresponding to the delta computation delay, k is the number of consecutive block-deltas computed, and P is the time shift between each block. The shifted delta cepstrum is also defined by a parameter N, equal to the number of MFCCs used in the computation. For our system, we use the configuration N = 7, d = 1, P = 3, and k = 7.

3. SYSTEM BUILDING BLOCKS

3.1. SUPPORT VECTOR MACHINES

Given a set of l observations, each consisting of a binary label $y_i \epsilon \{-1, 1\}$ and a feature vector x_i, we wish to find a suitable mapping function $f(x_i, \alpha)$ from x onto y with adjustable model parameters α.

A hyperplane can be defined in the feature space to separate the +1 and -1 distributions. We wish to maximize the ”margin” between the hyperplane and each distribution, where the margin is defined to be $\frac{2}{||w||}$. In this case, w is a vector normal to the hyperplane, and $\frac{|b|}{||w||}$ is the perpendicular distance from the hyperplane to the origin.

The classification decision for a feature vector is then defined to be:

$$y_i(x_i w + b) - 1 \geq 0, \forall i \tag{2}$$

We define the empirical error as

$$R(\alpha) = \frac{1}{2N} \sum_{i=1}^{l} |y_i - f(x_i, \alpha)| \tag{3}$$

In order to minimize the error, Burges shows that we can find an optimal Lagrangian of the form:

$$L = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i x_j \tag{4}$$

A detailed derivation can be found in [5].

For our problem, support vector machines offer a convenient means of classifying a frame of speech as either belonging to or not belonging to a particular distinctive feature. We trained a set of 11 distinctive feature recognizing SVMs for 1 manner feature, sonorant, and 10 place features: alveolar fricative, alveolar stop, alveolar nasal, interdental fricative, labial nasal, labial stop, labiodental fricative, postalveolar, retroflex, and velar stop.

3.2. GAUSSIAN MIXTURE MODELS

Gaussian mixture models (GMMs) enjoy widespread use in many speech recognition related applications. They provide a convenient means of modeling complex probability distributions by representing the probability density function of a random variable with a sum of weighted Gaussians. We give here a brief outline of the equations used to form our models; a more detailed explanation can be found in [6].

For the purposes of LID, we have a set of models of the form $g(x, \Theta^a)$. Each language is separately modeled by its parameters Θ^a, which correspond to the means, variances, and mixture weights of each Gaussian in the mixture. The general form of the PDF of each model is then given by:

$$g(x, \Theta^a) = \sum_{i=1}^{n} w_i \mathcal{N}(x_i; \mu_i, \sigma_i^2) \tag{5}$$

Bilmes shows that the parameters μ, σ^2, and w can be iteratively re-estimated using the Expectation-Maximization (EM) algorithm:

$$w_j = \frac{1}{N}\sum_{i=1}^{N}(p(j|x_i,\Theta^g) \tag{6}$$

$$\mu_j = \frac{\sum_{i=1}^{N} x_i p(j|x_i,\Theta^g)}{\sum_{i=1}^{N} p(j|x_i,\Theta^g)} \tag{7}$$

$$\Sigma_j = \frac{\sum_{i}^{N} p(j|x_i,\Theta^g)(x_i-\mu_j)(x_i-\mu_j)^T}{\sum_{i=1}^{N} p(j|x_i,\Theta^g)} \tag{8}$$

Where w_j, μ_j, and Σ_j in Eq. 8 are the newly estimated weights, means, and variances of the jth mixture component, Θ^g are the old parameters, x is the observation vector and N is the number of observations.

4. CORPUS

The SVM array was trained on the NTIMIT speech corpus. NTIMIT was created by re-recording the TIMIT corpus across telephone channels, effectively lowpass filtering TIMIT down to the telephone band. TIMIT was originally developed by Texas Instruments and MIT in order to provide a phoneme-rich corpus for training English language phoneme recognizers [7].

Training data for the GMM came from the CallFriend corpus provided by the Linguistic Data Consortium [8]. It contains three sets of 20 telephone conversations for each of 15 languages and dialects: Arabic (Egyptian), Mandarin (Mainland China), Mandarin (Taiwanese), English (American non-Southern), English (Southern), Farsi, French (Canadian), German, Hindi, Japanese, Korean, Spanish (Caribbean), Spanish (Non-Caribbean), Tamil, and Vietnamese. Each conversation lasts between 5 and 30 minutes and takes place between two speakers, one per stereo channel. Each conversation was split into two mono channels, and then silences were removed from each recording, resulting in 120 utterances per language. The corpus was then split into training and testing sets - the first 2.5 minutes of 80 utterances per language were used as training data, while the first 30 seconds of the remaining 40 utterances per language were used for testing.

5. EXPERIMENTS

Our system was comprised of two levels, with the first being an array of 11 radial basis function SVMs and the second a set of 15 language-specific GMMs, adapted from a Universal Background Model (UBM). At each timestep, we perform two feature extraction steps: a length 56 SDC vector, and a length 39 MFCC vector. Each of the 11 distinctive feature detection SVMs produces a classification output from the MFCC vector, resulting in a length 11 discriminitive feature vector where each element in the vector is the probability estimate for one SVM. This vector is then concatenated with the SDC vector to produce a length 67 hybrid vector. The hybrid vector is fed into the set of GMMs, each of which contains 1024 mixture components. The resulting output of the GMM set is a vector of log likelihoods for each language, and the test input is labelled with whichever language produced the greatest score. A UBM was first trained using data from all 15 languages, and then language-specific GMMs were obtained from the UBM by Maximum-a-Posteriori (MAP) adaptation, each using only training data from one language. This approach has previously been shown to work in speaker identification systems [9]. A system block diagram can be seen in figure 1.

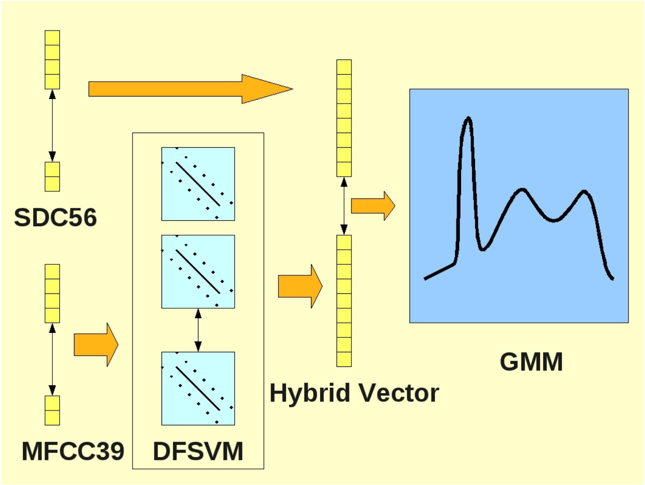

Fig. 1. System block diagram

6. RESULTS

The results of our experiment are given in table 1. Our hybrid system outperformed the baseline system with an increase in accuracy of 8.5% absolute and a decrease in the equal error rate of 5.5% absolute. The error rates can be seen in figure 2.

	Accuracy (%)	EER (%)
GMM	43.33%	24.90%
DFSVM-GMM	51.83%	19.40%

Table 1. Classification accuracies and equal error rates of baseline and GMM-SVM hybrid systems

7. DISCUSSION AND CONCLUSION

In this paper, we have shown that phonetic information in the form of linguistic distinctive features can be used to provide

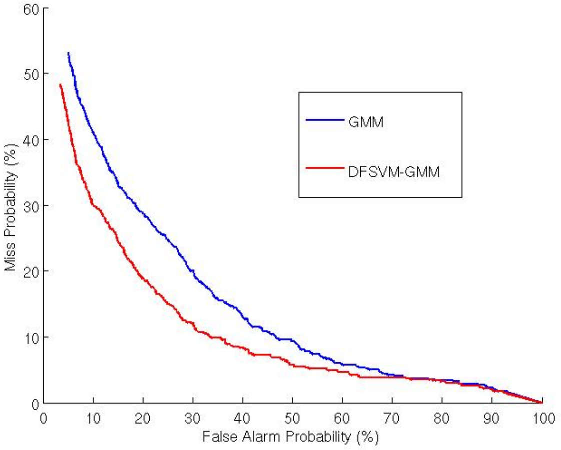

Fig. 2. Error rates for baseline and GMM-SVM hybrid system

an accuracy boost to an HMM-based LID system. In our future work, we propose to include additional distinctive features, and possibly expand the system to include prosodic features as well. Lastly, all of the distinctive feature SVMs were trained using the NTIMIT corpus of English speech, but it is possible that better accuracy would result if data from multiple languages was pooled together to train the SVMs.

8. REFERENCES

[1] Jiri Navratil, *Multilingual Speech Processing*, chapter 8: Automatic Language Identification, Academic Press, 2006.

[2] Ken Stevens, *Acoustic Phonetics*, MIT Press, Cambridge, MA, 1999.

[3] Sarah Borys and Mark Hasegawa-Johnson, “Distinctive Feature Based SVM Discriminant Features for Improvements to Phone Recognition on Telephone Band Speech,” *Interspeech*, 2005.

[4] Mary A. Kohler Richard J. Greene Douglas A. Reynolds Pedro A. Torres-Carrasquillo, Elliot Singer and J.R. Deller Jr., “Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features,” *Proc. ICSLP*, September 2002.

[5] Chris Burges, “A Tutorial on Support Vector Machines for Pattern Recognition,” *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[6] Jeff A. Bilmes, “A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models,” April 1998.

[7] Linguistic Data Consortium, “NTIMIT speech corpus,” .

[8] Linguistic Data Consortium, “CallFriend orpus,” .

[9] Thomas F. Quatieri Douglas A. Reynolds and Robert B. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models,” *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.