MULTI-STREAM DIALECT CLASSIFICATION USING SVM-GMM HYBRID CLASSIFIERS

Rahul Chitturi, John. H.L. Hansen[1]
Center for Robust Speech Systems
Erik Jonson School of Engineering and Computer Science
University of Texas at Dallas, Texas, USA
rahul.ch@student.utdallas.edu, john.hansen@utdallas.edu

ABSTRACT

In this paper, we investigate two important issues that influence dialect classification: (i) exploring dialect dependent features, and (ii) an effective way of combining spectral, excitation, and vocal tract information to improve dialect classification. The motivation is that dialect dependent features such as formants, LSP (Line Spectral Pairs) and MEPZ (MFCCs + energy + pitch) span a wider range of speech production traits and are therefore better suited than traditional MFCCs for characterizing dialects. After establishing the proposed algorithm, we compare individual performances of each feature on a corpus of three dialects of Spanish. Next, we present a method for combining these features using GMM-SVM hybrid classifiers. The final combined system achieves a 30% relative improvement in dialect classification accuracy, confirming that the proposed advances significantly outperform conventional methods for dialect classification.

Index Terms— LSP, MEPZ, GMM, SVM, Dialect Classification

1. INTRODUCTION

Automatic Dialect Classification has recently emerged to be of substantial interest in the speech processing community [3, 11, 12]. Dialect classification systems can be used to improve the performance of Speech Recognition engines by employing dialect dependent acoustic and language models [1, 3]. Traditional speech recognition systems are not robust to variations due to speaker dialect. Speaker adaptation is a solution for this, but in real-time situations such as transcribing live broadcast news [19], speaker adaptation is not feasible. Therefore dialect classification is one solution where dialect dependent acoustic models can be trained to improve automatic speech recognition (ASR) performance. Many researchers have found that using dialect dependent speech recognition models would improve ASR performance. Diakoloukas et.al, [1] developed dialect dependent ASR adaptation, while Huggins et.al, [2] formulated automatic speaker classification based on dialect. Recently, Gray and Hansen [3] used dialect classification techniques for Rich Indexing of Historical speech databases and providing dialect information for Spoken Document Retrieval Systems.

The current state-of-the-art Dialect Identification systems are based on traditional spectral features such as MFCCs, MVDR, PMVDR, etc [12]. However dialect information has time domain characteristics contained in prosody which are not captured with spectral features. Therefore in this paper we explore the use of a range of features that can capture both the acoustic and prosodic/excitation characteristics as well. We observe that the performance of individual features over all dialects may not be good, but they are well-suited or biased for a particular dialect. For example formant features are well suited for identifying the Puerto Rican dialect of Spanish, while the MFCCs are good for the Peruvian Spanish dialect. However their performance over general Spanish dialects is lower.

In this study, we propose a classifier to exploit the positive characteristics of these individual features by employing a SVM-GMM hybrid system, which sets the weights to the baseline GMM classifiers (posterior probabilities) that are trained on these features separately. SVM-GMM classifiers have been explored for the problem of speaker recognition [9]. Recently [10], this approach has also been used for Language Identification. In this paper, we verify the efficiency of this approach for dialect classification with new feature domains. The individual performances of the baseline GMM classifiers are below 69.3%, whereas we achieve 85% accuracy with our approach. A 30% relative improvement in dialect classification accuracy confirms that the proposed advances significantly outperform conventional methods for dialect classification.

[1] This work was funded by RADC under contract A4104, and by University of Texas at Dallas under project EMMITT

978-1-4244-1746-9/07/$25.00 ©2007 IEEE

This paper is organized as follows. In Sec. 2, we describe the database that is used for algorithm development and evaluation. The baseline system which serves as the starting point for our proposed advances is described in Sec 3. Next, we explore dialect dependent features in Sec. 4. Sec. 5 shows how the features are combined using GMM-SVM hybrid classifiers. Results are shown in Sec. 6, with a summary and conclusions in Sec. 7.

2. DATABASE DESCRIPTION

The corpus used in our study is a Latin American Spanish dialect speech database with 3 different dialects from Cuba, Peru and Puerto Rico (PR), which is described in [8]. The spontaneous speech portion was recorded in an interview style. The interviewer gave sample topics such as "describe your family", and the subject would respond. The interviewer would give some hints during the collection in order to keep the subject talking. The subject used a head-mounted microphone, which also captured the speech from the interviewer at a much lower amplitude since the interviewer sat across from the subject and far away from the microphone. The speech from both the interviewer and the subject were recorded on the same channel. Table 1 summarizes the data used for our system development and evaluations.

Data	Training Data			Testing Data		
	Cuba	Peru	PR	Cuba	Peru	PR
Speakers	29	29	26	13	13	12
Minutes	52	53	36	21	23	17

Table 1: Spanish Dialect Database Description (Cuba, Peru, PR)

3. BASELINE SYSTEM

The Spanish speech database from Sec. 2 has no transcriptions and therefore, it is difficult to build a supervised generative model such as an HMM. GMMs are being used as one of the traditional unsupervised classifiers for dialect ID. The baseline system for dialect classification is a GMM based unsupervised dialect ID system originally proposed by Huang and Hansen [11]. In addition to their baseline system, we also test the performance of our method with the current state-of-the art methods (briefly explained in Sec 3.2 and 3.3). After considering dialect differences, sensitive features in Sec.4, and combining individual detectors in a machine learning/ SVM (Support Vector Machine) framework in Sec.5, we compare our results with the current techniques in Sec.6. Throughout this paper, we use 600 mixtures for all GMM models that are employed in this study.

3.1 GMM classifier

The GMM classifier is a popular method for text-independent speaker recognition and dialect classification. We use this as our baseline system. Fig. 1 shows the block diagram of the baseline GMM training system. The silence removal module sets aside silence in the audio files that are used for training and testing. A parallel gender ID system is used to select dialect sets for each gender. Next, gender dependent GMM models are trained for each dialect. While testing, the incoming audio is classified as a particular dialect based on the maximum posterior probability measure over all the Gaussian Mixture Models under test.

3.2 Mixture Selection: MS-GMM

In this method, the GMMs are trained in a manner similar to that described for the baseline. However by choosing the top representatives of the GMM mixtures, the MS-GMM classification scoring produces results which outperform the baseline system. In [11], the best performance is obtained when the top 75% of the GMM mixtures were chosen from a potential pool of 600 mixtures. This represents the MS-GMM (mixture selection based GMM) dialect classification system.

3.3 Frame Selection: FS-GMM

In this method, the most confusing speech frames across the dialects are removed in the training data and new GMMs are trained with the remaining speech data. The threshold which gives the maximum performance is set as the operational threshold. The FS-GMM approach for Dialect ID is discussed in more detail in [11]

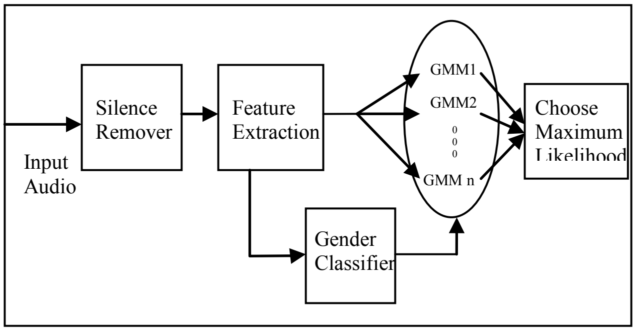

Figure 1: Baseline GMM based dialect classification

4. DIALECT BASED FEATURES

In this section, we describe the four features that are used within the integrated system. Later in Sec. 6.1, we consider the performance of all features that are described here. It can be observed that the individual features are biased for some particular dialects. In Sec. 5, we describe how individual feature based dialect detectors are combined into the overall final system.

Linguists and speech scientists have conducted extensive research in identifying dialect differences, including languages such as Spanish. Features such as formant trajectory have been explored which were found to be useful for accent and dialect classification [3, 17, 20]. Here, we consider a combination of acoustic information for the dialect classification with application to dialects of Spanish. We believe that our features are capable of identifying the dialect differences [5] such as phone deletions – (Peruvian: /D/), (PR:/s/), (Cuban:/D/) and replacements: (Peruvian: /x/ → /h/, /s/ → /h/), (PR: /d/ →/r/,/r/→/l/), Cuban(/h/ →/s/) [note: in this notation, → means that /x/ replaces /h/ in the first example].

4.1 MFCC

Many researchers have used spectral based features such as MFCC [11][12] for the purpose of dialect classification. In our experiments we use traditional 26-dimensional feature vector consisting of MFCC, ΔMFCC, Energy and ΔEnergy.

4.2 Formants

Gray and Hansen [3] showed that formant trajectory was very helpful in identifying the dialect. A great deal of research has considered formant structure and its influence on Accent/ Dialect. Yan and Vasegi [13] observed extensive variations in the formant structure of the phonemes across different dialects [13]. We employ the first four formants as the formant feature. These formant locations are estimated computing the roots of the LPC denominator (e.g.$\frac{G}{A(Z)}$) polynomial.

4.3 Line Spectral Pairs (LSP)

Line Spectral pairs are one of the important features for speech that are related to linear predictive analysis. LSPs are used extensively for various speech related applications like speech coding, speech enhancement [15], speaker recognition [14], etc. Liu et.al, studied LSP derived parameters in a VQ based text-dependent speaker verification system and concluded better performance using LSP frequencies over the cepstral coefficients [16].

Assuming we have obtained the p order LPC inverse filter $A_p\,(Z)$, LSPs are defined as the zeroes of the two polynomials constrained by $A_p\,(Z) \,\pm\, Z^{-(p+1)}A_p\,(Z^{-1})$. These zeroes are on the unit circle and their position and difference representatives are related to the formant frequencies locations. To overcome the ordering property of the LSPs, these are centered by subtracting the long term mean of each feature. Here, 20th order linear prediction is considered during LSP feature extraction.

4.4 MEPZ

All the features that are described in Sec.4.1-Sec.4.3 are all spectral in nature. As prosody is one of the important features for dialect classification which include the time domain information, we combine additional dimensions like energy, energy slope, pitch and zero crossing rate. This intermediate feature is a four dimensional feature which is then combined with MFCC and termed as MEPZ (MFCC + Energy + Pitch + ZCR). Features such as MEPZ were shown to be helpful in detecting speaker emotion. Since this is a feature that captures speaker level information, it motivated us to use this feature for dialect classification.

5. FEATURE COMBINATION

The motivation for combining features for dialect classification is that speech production employs an integrated combination of articulatory, excitation, prosody, linguistic and grammatical traits, and a combination is expected to be dialect dependent. In order to take advantage of the discriminating ability of the four feature dimensions: MFCCs, formants, LSPs and MEPZ we develop a strategy to combine individual GMM classifiers. Here, we compare two different techniques that can be employed to combine classifiers: (i) SVM-GMM hybrids and (ii) Bayesian GMM Hybrids

5.1 SVM-GMM Hybrid Classifier

Support Vector Machines (SVMs) are being used extensively in speech processing because they assume less on the feature distribution than HMMs/GMMs. Many researchers are exploring SVM hybrid classifiers for speech recognition [21], speech segmentation, speaker verification [9], etc. Here, we apply this SVM-GMM hybrid classifier for dialect classification.

In the decision space, we consider the performance of the individual feature detectors for dialect ID. If any individual detector gives far superior performance, the weight for this output is set high, and all others are set low (e.g., as an example, formant structure is shown to be very effective for Puerto Rican dialect of Spanish, and therefore weighted appropriately high). Next, we employ the SVM feature combination strategy to optimize the weights of the features using a greedy search approach and the hybrid system.

SVMs in their simplest form are hyper-plane linear classifiers, which maximize the margin between the in-class versus out-of-class minimizing the structural and the empirical risk [21]. In real world problems, the data is not linearly separable. In Fig 3, we show how the Cuban dialect mixes with the Peruvian and the Puerto Rican dialects, and there is no linear plane in this case which could separate these three dialects. In this case, the kernel functions can be used to transform the training data into a higher dimensional space. In our case, we use the Radial Basis Kernel Function (RBF). RBF kernel and the SVM function are defined as follows,

$$\boldsymbol{Kernel(x,y) = \exp\{-\gamma|x-y|^{2}\}} \tag{1}$$

$$\boldsymbol{SVM(x) = \sum_{i=1}^{n} \alpha_i \gamma_i Kernel(x,x_i) + b} \tag{2}$$

where γ is the kernel parameter that sets the extent of non-linearity of the decision surface and α corresponds to the weight of every sample point in the feature space – this is non-zero for support vectors and these support vectors decide the classification accuracy. We use the train data to obtain the best parameter settings for (α, γ, b). The SVMs return the distance from the hyper-plane while testing.

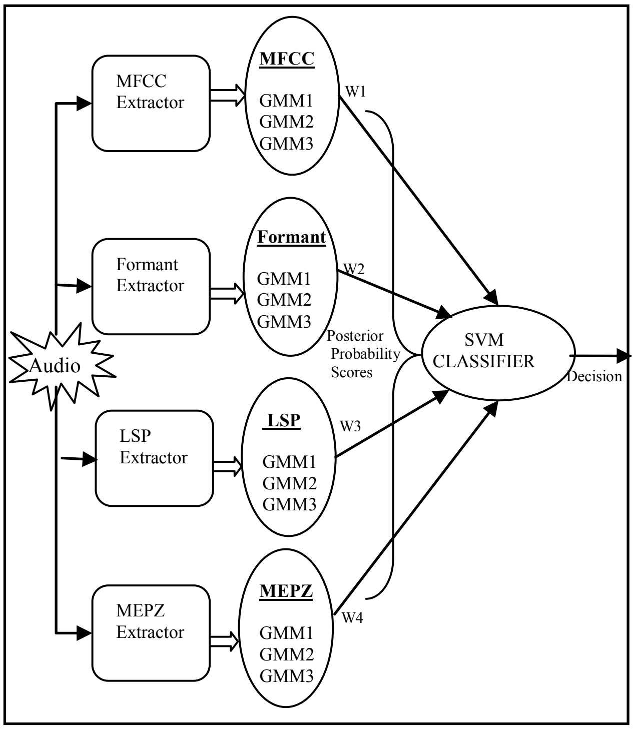

Figure 2: GMM-SVM Hybrid Classifier

In the proposed GMM-SVM hybrid classification, first the individual GMM classifiers are constructed using each feature separately, with 600 mixtures as illustrated in Fig 1. The probability estimates that are produced by the individual GMM classifiers are then set with optimal weights and submitted to the SVM classifier adding the weighted log probabilities (See Fig.2.).

During the training phase all GMM classifiers are trained and tested with the same train data in order to obtain the probability estimates for all dialect classes. Using the probability estimates produced by GMMs, we set weights in a greedy manner for each individual classifier and then train the combining SVMs. The best weight set for the individual GMM classifiers is selected for the GMM-SVM system to obtain the best dialect classification performance.

In the test phase, the posterior probabilities are calculated for each of the GMM classifiers and the weighted log probabilities are then tested with the SVM classifier.

5.2 Bayesian-GMM Hybrid Classifier

The setup for the SVM-GMM hybrid that was described in Sec. 5.1 is similar to the Bayesian-GMM hybrid setup except that, instead of employing the SVM classifier, the weighted posterior probability that is the maximum over all the dialects is chosen as the classified dialect. Therefore the weighted Bayesian classification is formulated as follows.

$$\boldsymbol{p(d|F1,..,F4) = \frac{p(d)p(F1,..,F4|d)}{p(F1,..,F4)}} \tag{3}$$

where F1,..,F4 are the four features and d is a dialect.

$$\begin{aligned}\boldsymbol{p(d,F1,F2,F3,F4)} &= \\ &\boldsymbol{p(d)p(F1,..,F4|d)} \\ &\boldsymbol{p(d)p(F1/d)p(F2,..,F4|d)} \\ &\boldsymbol{p(d)p(F1|d)p(F2|d,F1)p(F3,F4|d)}\end{aligned} \tag{4}$$

Assuming the conditional independence assumptions between the features, $p(Fi|d,Fj) = p(Fi|c)$

Therefore,

$$\boldsymbol{p(d,F1,..,F4) = p(d)p(F1|d)p(F2|d)..p(F4|d)} \tag{5}$$

$$\boldsymbol{p(d|F1,..,F4) \propto \; p(d)\prod_{i=1}^{4} p(Fi|d)} \tag{6}$$

Since we select the dialect which has the highest posterior probability score

$$\boldsymbol{d_{best} = argmax_{d_i} \prod_{i=1}^{4} p(Fi|d)} \tag{7}$$

In the weighted Bayesian case, we assign weights (w_i) to features, in which case,

$$\boldsymbol{d_{best} = argmax_{d_i} \prod_{i=1}^{4} p(Fi|d)^{w_i}} \tag{8}$$

6. EXPERIMENTAL RESULTS

All the experiments we describe in this section were conducted on the database as described in Sec. 2. The performance of each of these features is shown in section 6.1. In section 6.2 we describe our experimental setup that we employ to combine these GMM classifiers effectively to attain the best accuracy.

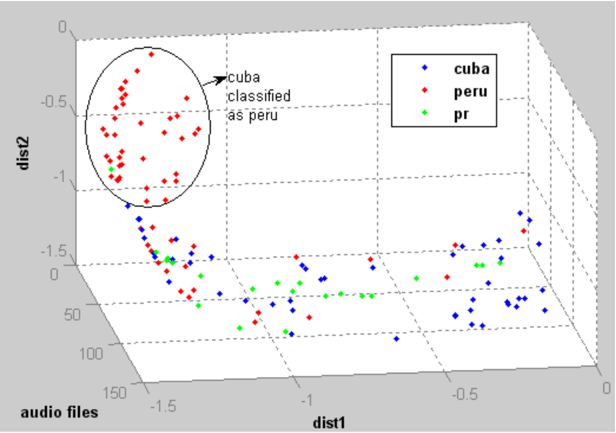

Figure 3: Cuban dialect compared to Peruvian and the Puerto Rico (PR) dialects. Dist1: represents the output score difference between Cuba and Peru GMMs; Dist2: represents the output score difference between Cuba and Puerto Rico GMMs.

6.1 Feature Comparison

Tables [2]-[5] show the individual performance of each of these features that was employed in our system. All the features are extracted using 20 ms frames with a 50% overlap between the windows. The accuracies that are best compared to other features are marked in gray in each table. Conclusions from individual feature classifiers are as follows: (i) for MFCCs, Peru is well detected, and Cuba has high confusion, (ii) for LSPs reasonable performance is achieved for all three dialects, (iii) for formants, Puerto Rico has outstanding performance (and is therefore the primary feature which should be used for PR dialect detection), and (iv) for MEPZ, okay performance exists for Peru, but confusion exists for Cuban dialect.

Train ⇨ **Test⇩**	**Cuba**	**Peru**	**Puerto Rico**
Cuba	**37.3%**	46.6%	16.6%
Peru	7.2%	**90.5%**	2.1%
Puerto Rico	12.6%	19.4%	**67.9%**
Overall Accuracy:	**65.23%**		

Table 2: Performance of MFCCs

Train ⇨ **Test⇩**	**Cuba**	**Peru**	**Puerto Rico**
Cuba	**67.4%**	32.5%	0
Peru	29.9%	**56.1**	10.9%
Puerto Rico	14..5%	0.0%	**85.4%**
Overall Accuracy:	**69.3%**		

Table 3: Performance of LSPs

Train ⇨ **Test⇩**	**Cuba**	**Peru**	**Puerto Rico**
Cuba	**69.0%**	30.9%	0.0%
Peru	60.5%	**39.4%**	0.0%
Puerto Rico	0.0%	0.0%	**100%**
Overall Accuracy:	**69.46%**		

Table 4: Formants Performance

Train ⇨ **Test⇩**	**Cuba**	**Peru**	**Puerto Rico**
Cuba	**54.7%**	40.4%	4..7%
Peru	20.4%	**79.5%**	0.0%
Puerto Rico	14.5	24.2	**61.1%**
Overall Accuracy:	**65.1%**		

Table 5: MEPZ Performance

6.2 SVM-GMM vs. Bayesian-GMM

As described in the Sec. 5, for a given train audio the posterior probabilities of all the classifiers are submitted to the SVM classifier, setting the appropriate weights. The optimal weights of these classifiers are set using a greedy strategy. Tables 6-7 show the best performance of the proposed method. The SVM-GMM and Bayesian-GMM systems are both superior to any individual feature from Tables 2-5.

Train ⇨ **Test⇩**	**Cuba**	**Peru**	**Puerto Rico**
Cuba	**72.2%**	27.8%	0%
Peru	16.91%	**83.09**	0%
Puerto Rico	0%	0%	**100%**
Overall Accuracy:	**85.09%**		

Table 6: Performance of SVM-GMM classifier

Train ⇨ **Test⇩**	**Cuba**	**Peru**	**Puerto Rico**
Cuba	**68.38%**	31.62%	0%
Peru	26.09%	**73.91%**	0%
Puerto Rico	0%	0%	**100%**
Overall Accuracy:	**82.56%**		

Table 7: Performance of Bayesian-GMM classifier

6.3 SVM-GMM vs. Traditional Dialect ID

The proposed system is compared with the current-state-of-the-art techniques that are used for dialect classification. Table 8 compares the results of our system with other techniques.

Dialect⇨ Method ⇩	Cuba	Peru	Puerto Rico	Overall
Baseline	37.3%	90.5%	67.9%	65.23%
MS-GMM	40.4%	91.2%	66.0%	65.87%
FS-GMM	41.2%	91.5%	67.9%	66.86%
SVM-GMM	**72.2%**	**83.09%**	**100%**	**85.09%**
Bayesian-GMM	68.38%	73.91%	100%	82.56%

Table 8: Proposed Method compared to traditional Dialect Classification Techniques

7. SUMMARY AND CONCLUSIONS

In this paper, we have investigated two important issues that influence dialect classification: (i) dialect dependent features, and (ii) effectively combining multiple feature sets to improve dialect classification. The motivation was based on the observation that dialect dependent features such as formants, LSP (Line Spectral Pairs) and MEPZ (MFCCs + energy + pitch) span a wider range of speech production traits and would therefore be better suited than traditional MFCCs for characterizing dialects. The proposed algorithm incorporates four features, with the output of the GMM feature detectors fused within an SVM system. We also considered a Bayesian-GMM scheme to incorporate the individual GMM classifiers. Evaluation on a corpus of Spanish dialects from Cuba, Peru and Puerto Rico showed reasonable performance for individual feature based detectors (overall dialect classification results ranging from 65.23-69.3%). When combining these within the GMM-SVM hybrid classifier, overall performance increased to 85.09% (a relative 30% improvement in dialect classification accuracy). This performance also outperformed the Bayesian-GMM scheme which achieved an overall dialect ID rate of 82.56%. The study therefore has demonstrated the importance of combining alternative feature domains, which conceptually would have a different set of errors that are partially corrected by combining classifier outputs for dialect classification.

8. REFERENCES

[1] *Diakoloukas, V.; Digalakis, V.; Neumeyer, L.; Kaja, J.;* ‟Development of dialect-specific speech recognizers using adaptation methods”, in *Proc. ICASSP-97.*

[2] *Huggins, A.W.F.; Patel, Y.;* ‟The use of shibboleth words for automatically classifying speakers by dialect”, in *Proc. ICSLP 96.*

[3] *Gray, S.; Hansen, J.H.L.;*” An integrated approach to the detection and classification of accents/dialects for a spoken document retrieval system”; IEEE ASRU 2005.

[4] Polikar R., Ensemble based systems in decision making, IEEE Circuits and Systems Mag., vol. 6, no. 3, pp. 21-45 , 2006.

[5] *Moreno.A. and Mariño J.B.*, “Spanish Dialects: Phonetic Transcription”, *in Proc. ICSLP 1998,*

[6] *Yanguas .L. R, O'Leary G. C., and Zissman M. A.,* “Incorporating LinguisticKnowledge into Automatic Dialect Identification of Spanish”, in *Proc.ICSLP*1998

[7] *Janin A., Ellis D., and Morgan N.,* “Multi-stream speech recognition: Ready for prime time,” *Proc Eurospeech-1999.*

[8] *Zissman, M.A.; Gleason, T.P.; Rekart, D.M.; Losiewicz, B.L.* Automatic dialect identification of extemporaneous conversational, Latin American Spanish speech- ICASSP-96.

[9] *Hou.F.; Wang.B.;* “Text-independent speaker recognition using probabilistic SVM with GMM adjustment”, Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003 International Conference on; 26-29 Oct. 2003 Page(s):305 - 308

[10] *Castaldo.F. et.al,*”Acoustic Language Identification Using Fast Discriminative Training”, in *Proc. ICSLP 2007*, August 27-31, 2007; Antwerp, Belgium

[11] *R. Huang, Hansen J.L.H.*; “Gaussian Mixture Selection and Data Selection for Unsupervised Spanish Dialect Classification”, in *Proc .ICSLP* 2006, Pittsburg, USA, November, 2006

[12] *Hansen J.H.L., Yapanel.U, Huang.R., Ikeno.A.* “Dialect Analysis and Modeling for Automatic Classification,” *in Proc. ICSLP-2004*, Jeju Island, South Korea, Oct. 2004.

[13] *Yan.Q; Vaseghi, S.;* Analysis, modelling and synthesis of formants of British, American and Australian accents; in *Proc .ICASSP* 2003, Volume 1, 6-10 April 2003 Page(s):I-712 - I-715

[14] *Campbell, J.P,* Speaker recognition: a tutorial; Proceedings of the IEEE, Vol.85, Iss.9, Sep 1997 Pages:1437-1462

[15] *Zhang. X.; Hansen, J.H.L.; Rehar, K.A.;* Speech enhancement based on a combined multi-channel array with constrained iterative and auditory masked processing; In Proc. ICASSP 2004.,Vol.1,Iss.,17-21May2004 Pages: I- 229-32 vol.1

[16] *Chi-Shi Liu, Min-Tau Lin,Wern-JunWang and Hsiao-ChuanWang;* Study of Line Spectrum Pair frequencies for speaker recognition. *ICASSP*, 1990.

[17] *Angkititrakul, P.; Hansen, J.H.L;* Advances in phone-based modeling for automatic accent classification. IEEE TASLP Vol.14, Iss.2, March 2006

[19] *Hansen, J.H.L.; Huang, R.; Zhou, B.; Seadle, M.; Deller, J.R.; Gurijala, A.R.; Kurimo, M.; Angkititrakul, P.;* SpeechFind: Advances in Spoken Document Retrieval for a National Gallery of the Spoken Word; Speech and Audio Processing, IEEE Transactions on, Vol.13, Iss.5, Sept. 2005Pages: 712- 730

[20] *Arslan L. M., Hansen, J.H.L.* "A study of Temporal Features and Frequency Characteristics in American English Foreign Accent," JASA, vol. 102(1), pp. 28-40, July, 1997.

[21] *Ganapathiraju. A.,* “Support Vector Machines for Speech Recognition,” *Ph.D. Dissertation*, Department of Electrical and Computer Engineering, Mississippi State University, January 2002.