Int J Speech Technol (2014) 17:99–105
DOI 10.1007/s10772-013-9209-1

GMM based language identification system using robust features

Sadanandam Manchala · V. Kamakshi Prasad · V. Janaki

Received: 19 April 2013 / Accepted: 7 September 2013 / Published online: 3 October 2013

Abstract In this work, we have proposed new feature vectors for spoken language identification (LID) system. The Mel frequency cepstral coefficients (MFCC) and formant frequencies derived using short-time window speech signal. Formant frequencies are extracted from linear prediction (LP) analysis of speech signal. Using these two kind of features of speech signal, new feature vectors are derived using cluster based computation. A GMM based classifier has been designed using these new feature vectors. The language specific apriori knowledge is applied on the recognition output. The experiments are carried out on OGI database and LID recognition performance is improved.

Keywords Language identification · LID · MFCC · Formants · LPC · LID for OGI and Indic Languages and new feature set

1 Introduction

Automatic language identification (LID) is the task of classifying an unknown utterance of speech from a list of languages. Text independent LIDs do not require the underlying message to identify the language. As these LIDs require only raw signals, such LID systems are popular. These systems are very useful in several applications like call routing systems, language translation, spoken document retrieval and front-end processing in multilingual systems etc. (Schultz and Waibel 2001; Waibel et al. 2000; Chelba et al. 2008). It is also a topic of great interest in the areas of intelligence and security for information distillation.

In practice, text independent spoken language recognition is far more challenging than text-based language recognition because there is no guarantee that a machine is able to transcribe speech to text without errors. We know that humans recognize languages through a perceptual or psychoacoustic process that is inherent in the auditory system. Therefore, the type of perceptual cues that human listeners use is always the source of inspiration for automatic spoken language recognition (Zhao et al. 2008).

These systems do not require labeled and segmented speech. There are several cues to identify the spoken language like phonemes, prosody, phonotactics, syntax and structure etc. (Zissman 1995). Among those cues, one of the important cues is acoustic phonetic for the LID task. While the term acoustic refers to physical sound patterns, the term phonotactic refers to the constraints that determine permissible syllable structures in a language. We can consider acoustic features as the proxy of phonetic repertoire and call it acoustic-phonetic features. On the other hand, we see phonotactic features as the manifestation of the phonotactic constraints in a language (NIST Language Recognition Evaluations 2007; Martin and Garofolo 2007).

The acoustic LID approach aims at capturing the essential differences among languages by modeling the distribution of spectral vectors directly. These systems use acoustic features like Mel frequency cepstral coefficients, shift delta cepstral coefficients, perceptual linear production

S. Manchala (✉)
Kakatiya University, Warangal, Andhra Pradesh, India
e-mail: sadanb4u@yahoo.co.in

V. Kamakshi Prasad
Jawaharlal Nehru Technological University, Hyderabad, Andhra Pradesh, India
e-mail: kamakshiprasad@yahoo.com

V. Janaki
Vagdevi Engineering College, Warangal, Andhra Pradesh, India
e-mail: janakicse@yahoo.com

(PLP) features, Formants etc., which are extracted directly from speech signals.

In the past, LID systems were developed with different type of features using statistical methods like vector quantization, hidden Markov models and Gaussian mixture models. The pattern classifier approaches used to implement LID using GMM and experiments were carried using 100 dimensional LPC derived feature vectors in Cimarusti and Eves (1982). A GMM is used to approximate the acoustic-phonetic distribution of a language. It is generally believed that each Gaussian density in a GMM captures some broad phonetic classes (Torres-Carrasquillo et al. 2002). However, GMM is not intended to model the contextual or dynamic information of speech. The GMM based LID systems perform classification using information from a single observation (Torres-Carrasquillo et al. 2002).

Sound frequencies are different in different languages and this difference is characterized by acoustic features like Mel frequency cepstral coefficients and delta cepstral coefficients. Using these features, LID was implemented using Gaussian mixture classifier (Zissman 1995, 1996).

Nagarajan and Murthy (2004) developed LID based on syllable like units recognition using HMMs. The basic requirement for building syllable-like unit recognizers for all the languages to be identified is an efficient segmentation algorithm. Earlier Kamakshi Prasad et al. (2004) proposed an algorithm, which segmented the speech signal into syllable-like units using minimum phase group delay.

There are several systems which were implemented using vector quantization (VQ), discrete HMM and Gaussian mixture model (GMM) using acoustic features like MFCCs and shifted delta cepstral (SDC) (Torres Carrasquillo et al. 2002; Nakagawa and Suzuki 1993). Muthusamy et al. (1994) discussed the segment approaches to LID with acoustic, phonotactic and prosodic features to develop LID and carried out experiments by combining the spectral feature vectors and pitch. Nagarajan and Murthy (2002) proposed VQ based LID using several statistical methods using MFCCs and used usefulness parameter to improve LID performance.

The communication among humans is established by Speech, wherein the information to be conveyed is embedded in the sequence of sound units produced. These basic sound units are normally referred to as phonemes. The sequences of phonemes used for communication are governed by the rules of the language. The acoustic characteristics of the phonemes are closely related to the manner in which they are produced. The phonemes that are produced depend on the type of excitation and the shape of the vocal tract system. As the basic sound units are characterized by a set of formant frequencies which correspond to the resonances of the vocal tract system, formants are one of the major acoustical cues for the identification of language from speech. Formant frequencies have rarely been used as acoustic features for language recognition, in spite of their phonetic significance.

State-of-the-art formant estimators locate candidate peaks of the spectra from short-time analysis of speech and perform temporal tracking. Traditional formant frequency estimation methods are based on spectral analysis and peak picking techniques. The characteristics of phonemes are generally manifested in spectral properties of speech signal and formants are best choice to represent the acoustic features of basic sound units which are useful for LIDs (Yegnanarayana 1978). Formants are extracted using linear prediction coefficients (LPC) and these formants have the information about different sounds (Bruce et al. 2002; Bruce and Mustafa 2006).

In this paper, we have proposed new features for a text independent LID system. Mel frequency cepstral coefficients (MFCCs) and formant frequencies are extracted from short-time processing of speech and these two kinds of features are concatenated to form feature vector. Formant features are extracted using linear prediction coefficients of speech signal. In this work, GMMs are created one for each language using new feature vectors of each language in training phase of LID. In the testing phase, MFCCs and formants are extracted from unknown utterance and these combined features are transformed into the new feature vectors. These new feature vectors of unknown utterance are evaluated against GMM of each listed languages. Usefulness of derived feature vectors is computed as the weightage of feature vectors. A language is hypothesized based on maximum usefulness value of sequence of new feature vectors of the unknown utterance against each GMM model. The steps followed in the implementation of LID system is explained in the following sections.

2 Feature extraction

The performance of any automatic LID system depends on several parameters and among them the selection of feature vectors is very important. For text Independent LID, feature vectors are extracted from speech signal without considering the knowledge of speech. From the existing systems (Nagarajan and Murthy 2002), it is observed that if the frequency of phonemes is different in different languages, the frequency of feature vectors is also different in different languages, as there is a correspondence between these two entities.

2.1 Derivation of new features

In this work, m-dimensional MFCC features and n-dimensional formant frequencies are extracted from speech signal and concatenated them to form $(m+n)$-dimensional feature

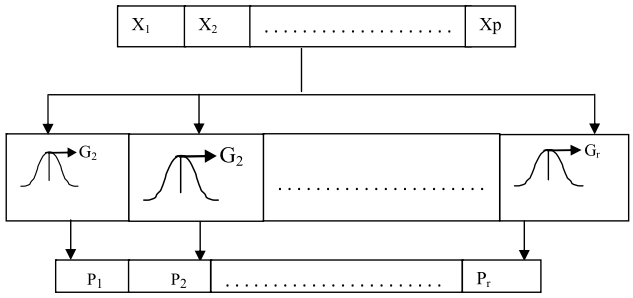

Fig. 1 Transformation of $(m+n)$-dimensional concatenated features into r-dimensional new feature vectors

vectors. These feature vectors are grouped into r-clusters using clustering algorithm. One Gaussian is designed for each cluster. These combined features are passed through all Gaussians and calculated probability using probability density function of the respective Gaussians. The probabilities of Gaussians are treated as coefficients of new feature vectors. Each feature vector ($m+n$-dimensional) is transformed into a r-dimensional new feature vector as described in Fig. 1. This proposed method for the derivation of new features is used in both training and testing phases of LID system. In training phase, new features are extracted from huge corpus of language specific speech for each language. In testing phase, new features are extracted from unknown utterance of speech.

3 Motivation of this work

The human speech apparatus is capable of producing a wide range of sounds. Speech sounds as concrete acoustic events are referred to as phones, whereas speech sounds as entities in a linguistic system are termed as phonemes (Bruce et al. 2002). The number of phonemes used in a language ranges from about 15 to 50, with the majority having around 30 phonemes each. Phonetic repertoires differ from language to language although languages may share some common phonemes. These differences between phonetic repertoires imply that each language has its unique set of phonemes thus acoustic-phonetic feature distributions (Kirchhoff 2006).

The performance of any LID system depends on the type of feature vectors and the classifier used. If the feature vectors do not represent underlying phonetic content of the speech, the system will perform poorly irrespective of the classifier used. The selection of features is very important for LID to get the good recognition performance. The fundamental cue to recognize the spoken language is the frequency of occurrence of basic sound units is different in different languages. In short-term speech processing, it is very likely that most of the cues of basic sound units are covered in a short-term window. Hence there is a close resemblance between basic sound units and derived feature vectors. This has motivated us to explore new features. In earlier systems, phonemes are described with the acoustic features. The acoustic features are represented well with MFCC. But the state of art LID systems gave the poor results for tonal languages with only MFCC features. This has motivated us to form the features by combining MFCCs and formants.

In VQ based LID systems, for each feature vector only one code book index is considered, discarding second best and third best indices etc. But these cues are also very important in making comprehensive decisions. Hence it is proposed to use k-best alternatives in decision making process instead of a single code book index.

As the probability of a feature vector in a language is greater than that of some languages and lesser than that of some other languages, significance of feature vector cannot be estimated in isolation. In such case, weightage is given to the feature vectors based on the log likelihood ratio of the feature vector for identification of language. To estimate the significance of feature vector among the languages, compute the usefulness of feature vector between a pair of languages and a language which gives maximum usefulness is allowed for further comparisons with other languages (Nagarajan and Murthy 2004). We propose to evaluate our new features using this usefulness criterion.

4 Design of text-independent LID using proposed model

In the proposed model, GMM based LID is implemented using new feature vectors of speech signal and the usefulness of new feature vectors. The LID system involves two phases namely training and testing phases. Each spoken language is represented by one GMM. If there are M languages, correspondingly there are M GMMs in the recognition system. For training each of M GMMs, language specific speech corpus is used unlike in training of each of r-Gaussians, the speech corpus of all languages is combined, r-clusters are formed and one Gaussian for each cluster is designed as describe in Sect. 2.

4.1 Training phase

The training phase involves two steps. In the first step, consider a huge speech corpus consisting of speech of 25 minutes duration for each language. m-Dimensional MFCC feature vectors and n-dimensional formant frequencies are extracted from each of the listed languages by applying overlapped short time windows. These $(m+n)$-dimensional feature vectors are converted into r-dimensional new feature vectors as discussed in Sect. 2.1.

The second step of training phase involves training of GMMs one for each language using Baum-Welch re-estimation algorithm (Nagarajan and Murthy 2002). The k-dimensional new feature vectors of speech corpus are used

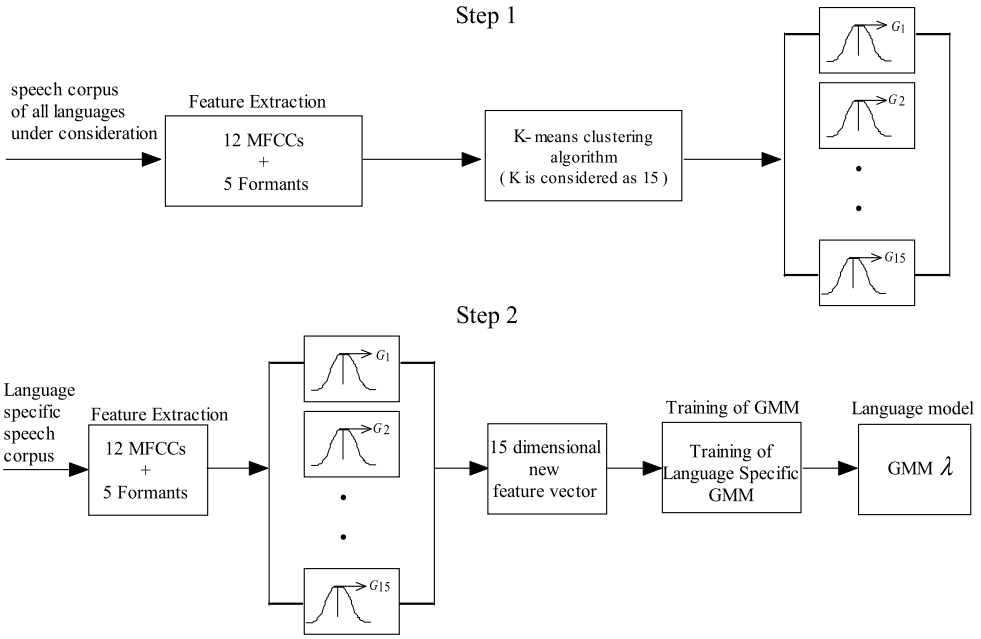

Fig. 2 Training phase of LID

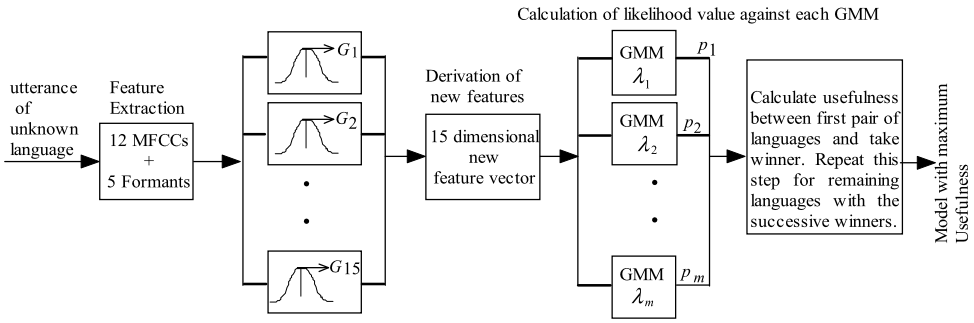

Fig. 3 Testing phase of LID

to train GMMs. The flow diagram of training phase of the language identification system is illustrated in Fig. 2.

4.2 Testing phase

In the first step, the new feature vectors of unknown utterance of speech are derived using the procedure followed in training phase as in Fig. 2. These new feature vectors of unknown utterance of speech are used as observation sequence of GMM.

In the second step, the new feature vectors of speech utterance of unknown language is evaluated against each of M GMMs, where M is the number of languages under consideration using forward-backward algorithm as in Fig. 3.

The significance of feature vectors among the languages is obtained by computing the usefulness of feature vector between a pair of languages and a language which gives maximum usefulness is allowed for further comparisons with other languages (Nagarajan and Murthy 2002). Usefulness of spectral vector is defined (Nagarajan and Murthy 2002) as

$$U(V_k, \lambda_i)/\lambda_j = P(V_k/\lambda_i) \log \frac{P(V_k/\lambda_j)}{P(V_k/\lambda_i)} \tag{1}$$

where V is sequence of feature vectors, λ_i, λ_j are the languages which are considered in training and $P(V_k/\lambda_i)$ is the likelihood of feature vector V_k in λ_i Language.

In the third step, the usefulness of all feature vectors for a pair of languages is calculated using (1). The Language which gives maximum usefulness of all feature vectors will be allowed to further comparisons with other language one at a time. This process is repeated for all languages under consideration. If the considered languages are M, then the number of comparisons is $(M-1)$ only. In this process, after the comparison of all languages, the language which gives maximum usefulness in the last comparison is identified as the recognized language.

Table 1 LID performance in % using usefulness for OGI database with MFCCs

Language	Performance (%) for 8 mixtures			Performance (%) for 16 mixtures			Performance (%) for 32 mixtures		
	1 sec	2 sec	3 sec	1 sec	2 sec	3 sec	1 sec	2 sec	3 sec
English	78	79	80	82	82	82	83	84	83
French	72	72	72	70	71	71	70	72	72
Farsi	38	42	45	42	44	44	43	48	50
German	42	42	44	43	43	44	45	49	51
Mandarin	0	0	12	2	4	18	0	14	21
Spanish	14	32	38	33	42	44	42	45	50
Japanese	55	52	58	57	62	62	66	72	73
Korean	63	58	59	59	62	62	63	64	64
Tamil	82	80	82	82	82	83	84	85	88
Vietnam	45	59	50	51	51	50	51	52	53
Average	48.9	51.6	54	52.1	54.3	56	54.7	58.5	60.5

Table 2 LID performance in % using usefulness for OGI database with MFCCs and formants

Language	Performance (%) for 8 mixtures			Performance (%) for 16 mixtures			Performance (%) for 32 mixtures		
	1 sec	2 sec	3 sec	1 sec	2 sec	3 sec	1 sec	2 sec	3 sec
English	80	80	81	82	83	83	82	84	85
French	72	70	72	71	72	72	72	74	76
Farsi	45	49	50	47	47	48	49	50	51
German	50	51	51	52	52	54	54	55	55
Mandarin	20	22	22	21	26	24	26	26	27
Spanish	44	47	49	50	52	52	53	55	59
Japanese	65	62	66	67	68	66	69	70	70
Korean	52	57	57	59	60	61	66	68	73
Tamil	90	90	89	92	91	94	94	93	95
Vietnam	46	50	52	51	52	55	54	58	62
Average	56.4	57.8	58.9	59.2	60.3	60.9	61.9	63.3	65.3

5 Experimental setup

The experiments are carried out using MATLAB 9.0 on Windows 7 platform. The OGI database has been used for this study (OGI Multi Language Telephone Speech 2004). MFCC and formant features are extracted from each short-term window and the successive windows are overlapped. In our experiment, 12 MFCCs ($m = 12$) and five formant frequencies ($n = 5$) are considered. We have evaluated the 17 ($m + n$)-dimensional concatenated feature vectors using 15 Gaussians ($r = 15$) to derive new feature vectors. Gaussian mixture models with varying number of mixtures 8, 16 and 32 are implemented using new features. Testing is performed for different utterances of 1 s, 2 s and 3 s duration using the proposed method.

6 Results

The performance of language identification for OGI database for different duration of test utterances and varying number of mixtures of GMM using usefulness value is calculated and results are obtained using different feature vectors of short-term windowed speech signal. The performance of language identification using only MFCC feature vectors is depicted in Table 1. The performance of LID is also measured with the features of MFCCs and formants and the results are furnished in Table 2. New features are obtained from concatenated features of MFCCs and formants. Formants are extracted using LP spectrum for training and testing. The performance of the LID system is evaluated using new features and results are shown in Table 3. The performance is measured in terms of percentage of correct identification of test samples from the given test samples.

Table 3 LID performance in % using usefulness for OGI database with new feature vectors

Language	Performance (%) for 8 mixtures			Performance (%) for 16 mixtures			Performance (%) for 32 mixtures		
	1 sec	2 sec	3 sec	1 sec	2 sec	3 sec	1 sec	2 sec	3 sec
English	100	100	100	100	100	100	100	100	100
French	99	99	100	99	100	100	100	100	100
Farsi	85	89	93	85	90	95	97	97	99
German	94	95	98	95	95	99	95	97	99
Mandarin	82	84	86	88	90	92	90	94	96
Spanish	86	89	89	89	90	92	92	95	99
Japanese	90	90	93	90	94	94	94	96	99
Korean	81	86	86	91	93	95	95	96	98
Tamil	99	99	100	100	100	100	100	100	100
Vietnam	85	88	90	100	100	100	100	100	100
Average	90.1	91.9	93.5	93.7	95.2	96.7	96.3	97.5	98.8

Table 4 GMM based LID performance using new features and Usefulness for OGI database

	8 mixtures	16 mixtures	32 mixtures
IR	91.5	95.2	98.8
FRR	9.2	4.9	1.8
FAR	13.5	8.6	3.4

Table 5 The average time required in milliseconds for testing LID system with different feature vectors

Feature vectors	Duration of utterance of unknown language		
	1 sec	2 sec	3 sec
MFCCs+formants	62.879	74.896	77.677
New features of MFCCs+formants	56.696	62.596	77.183

The performance of LID system is also referred in terms of Identification Rate (IR), False Acceptance Rate (FAR) and False Rejection Rate (FRR). IR is the percentage of test utterances that in certain languages and classified as "true" for those languages. FAR is the percentage of test utterances that are not in certain languages but classified them as "true" for those languages. FRR is the percentage of test utterances that in certain languages but classified as "false" for those languages.

The performance of LID in terms of IR, FAR, FRR for different duration of test utterances and varying number of Mixtures of GMM using likelihood value and usefulness is depicted in Table 4.

It is observed that the average performance of LID task of OGI languages of the duration of 1 s, 2 s, 3 s utterance for 32 mixtures is increased to comparatively to 16 mixtures and 8 mixtures as specified in Tables 3 and 4.

The computational time analysis is also performed for the experiments carried out in this work using core i3 processor with 2 GB RAM. The comparison of the time taken for testing of unknown utterance of speech with the different durations of test speech utterance is demonstrated in Table 5.

7 Conclusions

In this paper, a new GMM based approach has been proposed for text independent language recognition using new feature vectors derived from MFCC feature vectors and formants. Formants are extracted using LP spectrum of speech signal. LID system is developed using Gaussian mixture model with different mixtures. Formant and MFCC feature vectors represent the acoustic features of speech signals so that LID performance is improved. A significant improvement in the recognition performance was found with usefulness criterion combined with the new feature vectors. The procedures adopted in this paper are general in nature and hence could be extended to the implementation of any GMM or HMM based pattern recognition tasks. The average recognition performance of this text independent LID system is achieved more for 32 mixtures for OGI languages is 98.8 %.

References

Bruce, I. C., & Mustafa, K. (2006). Robust formant tracking for continuous speech with speaker variability. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *14*(2), 435–444.

Bruce, I. C., Karkhanis, N. V., Young, E. D., & Sachs, M. B. (2002). Robust formant tracking in noise. In *ICASSP*.

Chelba, C., Hazen, T., & Saraclar, M. (2008). Retrieval and browsing of spoken content. *IEEE Signal Processing Magazine*, *25*(3), 39–49.

Cimarusti, D., & Eves, R. B. (1982). Development of an automatic identification system for spoken languages, phase I. In *Proc. IEEE int. conf. acoust., speech, and signal processing* (pp. 1661–1663).

Kamakshi Prasad, V., Nagarajan, T., & Murthy, H. A. (2004). Automatic segmentation of continuous speech using minimum phase group delay functions. *Speech Communication*, *42*, 429–446.

Kirchhoff, K. (2006). Language characteristics. In T. Schultz & K. Kirchhoff (Eds.), *Multilingual speechprocessing*. Amsterdam: Elsevier.

Martin, A. F., & Garofolo, J. S. (2007). NIST speech processing evaluations: LVCSR, speaker recognition, language recognition. In *Proc. IEEE workshop on signal processing applications for public security and forensics* (pp. 1–7).

Muthusamy, Y. K., Barnard, E., & Cole, R. A. (1994). Automatic language identification: a review/tutorial. *IEEE Signal Processing Magazine*, *Oct. 1994*

Nagarajan, T., & Murthy, H. A. (2002). Language identification using spectral vector distribution across the languages. In *Proceedings of int. conf. natural language processing*.

Nagarajan, T., & Murthy, H. A. (2004). Language identification using parallel syllable-like unit recognition. In *Proc. IEEE int. conf. acoust. speech, and signal processing*.

Nakagawa, S., & Suzuki, H. (1993). A new speech recognition method based on VQ-distortion measure and HMM. In *Proc. int. conf. ASSP* (pp. 673–679).

NIST language recognition evaluations (2007). http://nist.gov/itl/iad/mig/lre.cfm.

OGI multi language telephone speech. www.cslu.ogi.edu/corpora/mlts/, January 2004.

Schultz, T., & Waibel, A. (2001). Language independent and language adaptive. *Speech Communication*, *35*(1–2), 31–51.

Torres Carrasquillo, P. A., Reynolds, D. A., & Deller, J. R. (2002). Language identification using Gaussian mixture model tokenization. In *Proc. IEEE int. conf. acoust., speech, and signal processing* (Vol. 1, pp. 757–760).

Torres-Carrasquillo, P., Singer, E., Kohler, M., Greene, R., Reynolds, D., & Deller, J. Jr. (2002). Approaches to language identification using Gaussian mixture models and shifted delta cepstral features. In *Proc. ICSLP* (pp. 89–92).

Waibel, A., Geutner, P., Tomokiyo, L. M., Schultz, T., & Woszczyna, M. (2000). Multilinguality in speech and spoken language systems. *Proceedings of the IEEE*, *88*(8), 1181–1190.

Yegnanarayana, B. (1978). Formant extraction from linear prediction phase spectrum. *The Journal of the Acoustical Society of America*, *63*, 1638–1640.

Zhao, J., Shu, H., Zhang, L., Wang, X., Gong, Q., & Li, P. (2008). Cortical competition during language discrimination. *NeuroImage*, *43*, 624–633.

Zissman, M. A. (1995). Overview of current techniques for automatic language identification of speech. In *Proceedings of the IEEE automatic speech recognition workshop* (pp. 60–62).

Zissman, M. A. (1995). Automatic language identification of telephone speech. *The Lincoln Laboratory Journal*, *8*(2), 115–144.

Zissman, M. A. (1996). Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on Speech and Audio Processing*, *SAP-4*(1), 31–44.