LANGUAGE IDENTIFICATION SYSTEM USING MFCC AND SDC FEATURE

Mr. Jangam Shrinivas Suresh
M. Tech Student
Dept. of Computer science and Engineering.
Rajarambapu Institute of Technology, Sakhrale, India
shrinivas.jangam@gmail.com

Prof. S. A.Thorat
HOP
Dept. of Computer science and Engineering.
Rajarambapu Institute of Technology, Sakhrale, India
sandip.thorat@ritindia.edu

***Abstract*:**

Speech recognition is technology which recognizes the spoken words and phrases and converts them to a machine-readable format. Speech recognition gives information about the spoken word, speaker, and language. According to this information, speech recognition has classes as text recognition, speaker recognition, and language identification. This system is to find out specific language from speech samples. The speech signal is basically intended to carry the information about the linguistic message, it also contains the language-specific information. In this regard, this work undertakes the study and implementation of Language Identification System using GMM classifiers. This system is based on Mel-frequency Cepstral Coefficients and Shifted Delta Cepstral feature extraction techniques. MFCC gives the information about human vocal tract shape and SDC gives the information about phonemes. In this language identification system combination of MFCC and SDC feature is used for better results and Gaussian Mixture Model is used as a classifier to increase the accuracy of identifying a language. This system works with 17 languages as Eastern Arabic, Bengali, German, Hindi, Hungarian, Japanese, Kannada, Malayalam, Kashmiri, Portuguese, Urdu, Russian, Spanish, Marathi, Tamil, Panjabi, and English.

Keywords: Language Identification, MFCC, SDC, GMM.

I. INTRODUCTION

Speech recognition is a process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words. Speech recognition is technology which recognizes the spoken words and phrases and converts them to a machine-readable format. Speech recognition gives information about a spoken word, speaker, and language. According to this information, speech recognition has classes as Text recognition, Speaker recognition, and language identification. Speech recognition system has input as speech sample and extracts required information from a speech signal.In our everyday life, there are many forms of communication, for example: body language, a language of the text, pictorial language and speech, etc. But between these forms speech is always considered to be the most powerful form because of its rich dimension of dimensions. Except for river text (words), the rich dimensions also apply to the gender, attitude, emotion, health, and identity of the speaker. This information is very important for effective communication. In terms of signal processing, speech can be characterized with respect to the signal transmission information. The waveform can be one of the representations of speech, and this type of signal is most useful in practical applications. By retrieving the signal, we can get three basic types of information: spoken text, language, and speaker identification. Speech signal with three recognition systems Speech (text) recognition, speaker recognition and language recognition [14].

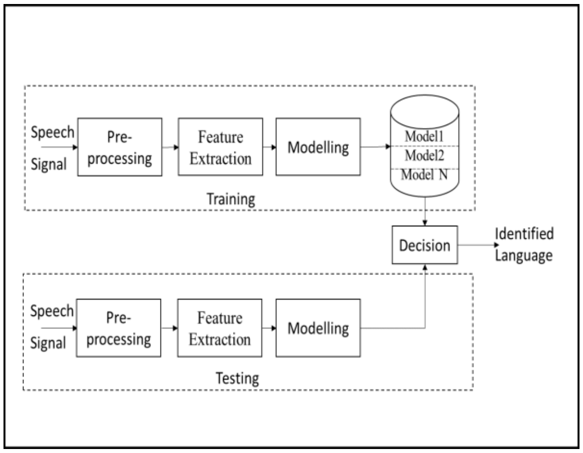

Fig 1.1 Language Identification System

Language recognition is a concept of identifying language from speech samples. To such type of recognition system has to work on feature extraction operation. Speech signals contain parametric information that is features of a speech signal. Language identification depends on phonetics and phonemes features. Each language is having unique phonemes according to that we can identify a particular language. In language recognition system, feature extraction operation computes MFCC that is Mel Frequency Cepstral Coefficients and SDC that is Shifted Delta Cepstral. Combining these MFCC and SDC system got accurate phoneme for particular language and according to that recognition is done [4].

II. LITERATURE REVIEW

A. Feature Extraction

Speech sample carries a special type of information that can be express speaker identity and language identity. This included language and speaker-related information such as phonemes, vocal tract shape, excitation source, behavior features, temporal information and acceleration information. Temporal information and acceleration information can be used for language recognition. Feature extraction is a technique to extract these type of information from speech sample and representing this into parametric information. This parametric information knows as features then used in language identification.

Manish Gupta, ShambhuBharati*et al.* [1] gives the systematic approach to identify language from speech samples using Linear Predictive Coding and Mel-Frequency Cepstral Coefficient features. They increase Performance of language identification by combining MFCC and LPC features. In this system reason behind using MFCC and LPC features were these features contain information about vocal tract shape and information about language which can be used in identification. They developed three systems according to MFCC, LPC individually and a combination of features MFCC and LPC. They compared these three systems and conclude that combination of MFCC and LPC features gives better results.

S. Jothilakshmi, V. Ramlingam*et al.* [2] developed a language identification system based on MFCC, Delta and acceleration, SDC. They developed three types of system 1st using MFCC 2nd using MFCC, delta MFCC and acceleration 3rd using SDC. Delta MFCC and acceleration features give temporal information of language speech. They compared system performance between MFCC, MFCC with delta and acceleration and SDC features for 9 Indian languages. This system declared MFCC with delta and acceleration features has better results than other features.

Shashidhar G. Koolagudi*et al.* [3] developed a system for identifying language from speech samples using Mel Frequency Cepstral Coefficient. MFCC features contain information of vocal tract system. Theydeveloped this system to compare system accuracy using different number of MFCC features. The accuracy of this system using 19 MFCC coefficients is 88.4% and for 29 MFCC coefficients 87.6 %. Hence in this work, they declared as language identification system based on MFCC features gives better result on more number of features extracted by speech.

UtpalBhattacharjee*et al* [4] developed a language identification system using MFCC and SDC as features. They developed based line system using MFCC feature and made an addition to this system by extracting SDC feature. Later they prepared system based on SDC and combination of MFCC with SDC features. They compared system using MFCC features, SDC features and combination of MFCC and SDC features. This work has shown that combination of MFCC and SDC features gives a better result.

Bo Yin, Fang Chen *et al.* [6] proposed a system of language identification to study different features extraction techniques for SDC, MFCC and PLP features. They developed language identification system using SDC feature which gives information about phoneme of languages. PLP features are mostly used in emotion and gender recognition by speech. They also compared system using MFCC and PLP features with different coefficient numbers. After successfully tested the system it declared that less number of coefficient of feature gives a better result.

Jue HOU, Yi LIU *et al.*[7] gives a systematic approach to identify language from speech samples using Cepstral and prosodic features. They developed a system using combination of SDC and pitch counter features. The accuracy of this system using SDC feature was 75 % and using Pitch features 71.6%.

AbhijeetSangwan*et al.* [8] implemented a system for language analysis based on knowledge of speech. The developed method automatically extracts key production traits or "hot-spots" which have significant language discriminative capability. In this system, the speech utterances were parsed into consonant and vowel clusters. The production traits for each cluster is represented by the corresponding temporal evolution of speech articulatory states.

Abhijeet Kumar, H. Hemani*et al.* [12] implemented language identification system to train with acoustic features. MFCC, SDC features were used in this system as feature extraction technique. They have done work on various preprocessing techniques for noise removal, speech activity detection, speaker normalization, channel normalization. These preprocessing technique improved accuracy in feature extraction and decreased the time taken by train operation. They compared MFCC and SDC features in language identification system.

ChithraMadhu, Anu George *et al.* [13] gives an approach for identifying language implemented on language dependent phonotactic features and prosodic information of particular language. They developed two types of language identification system as phonotactic system with 72 % accuracy and prosodic system with 68% accuracy.

B. Classification

Manish Gupta, ShambhuBharati*et al.* [1] gives a systematic approach to identify language from speech samples using Linear Predictive Coding and Mel-Frequency Cepstral Coefficient features. They developed a system for identifying language based on Support Vector Machine and Random Forest classification technique. This system worked well with 92.60% on 6 different languages.

S. Jothilakshmi, V. Ramlingam*et al.* [2] developed a language identification system which worked in two levels of identification. In first level identification system identifies a family of language and in the second level of identification system identifies particular language from identified family. This system developed to study three types of the classifier as Artificial Neural Network, Hidden Markov Model, and Gaussian Mixture Model. It is shown that LID system using GMM classifier worked better than other classifiers with 80.56 % accuracy.

Shashidhar G. Koolagudi*et al.* [3] developed a system for identifying language from speech samples. This system is developed for 15 Indian languages and used GMM (Gaussian Mixture Model) as a classifier. Using GMM classifier system was trained with parameters of Gaussian distribution and for K-means algorithm was used to make clusters. In all conditions, this system gives 100% accuracy for Assamese, Marathi, Nepali, Tamil, Urdu languages in identification. This system worked with 88% accuracy on 15 languages.

Ravi Kumar, Hari Krishna *et al.* [5] implemented a system of language identification using Cepstral features. They used MFCC as Cepstral feature and GMM-UBM as a classifier and developed this system to compare the accuracy of GGM based language identification to GMM_UBM based language identification. This system declared that GMM-UBM classifier is better than GMM classifier and improves the accuracy by 7% to 8% of identification. This system developed using speech samples from IITKGP-MLILSC.

AbhijeetSangwan*et al.* [8] implemented a system for language analysis based on knowledge of speech. The developed method automatically extracts key production traits or "hot-spots" which have significant language discriminative capability. In this system, the speech utterances were parsed into consonant and vowel clusters. The production traits for each cluster is represented by the corresponding temporal evolution of speech articulatory states. A selection of these production traits are strongly tied to the underlying language and exploited for identifying languages. This system carried out on 5 closely related languages spoken in India namely, Kannada, Tamil, Telugu, Malayalam, and Marathi. The LID accuracy of 65 % is achieved with this system.

Mendoza, Sergio, *et al.* [11] they developed a highly accurate automatic language identification system based on large vocabulary continuous speech recognition (LVCSR). Each test utterance is recognized in a number of languages, and the language ID decision is based on the probability of the output word sequence reported by each recognizer. Recognizers were implemented for this test in English, Japanese, and Spanish, using the Ricardo corpus of telephone monologues. When tested on the OGI corpus of digitally recorded telephone speech, they obtained error rates of 3% or lower on 2-way and 3-way closed-set classification of ten-second and one-minute speech segments.

Chakroun, Rania, Yassine Ben Ayed*et al.* [21] developed language identification system to study Support Vector Machine Classifier. This system was implemented in six languages as German, English, Arabic, Spanish, French, and Italian. This system was implemented using acoustic features classification. In support vector machine every sample of the speech signal is characterized by a unique vector of parameters, which does not correspond to the perceptive reality of the speech, which is continuous by nature. The system worked well with 89.72% accuracy.

Ossama Abdel Hamid, Hue Jiang *et al.* [24] developed a system for speech recognition. They applied Convolutional Neural Network to recognize speech within the framework of hybrid Neural Network – Hidden Markov Model. In this system, they used local filtering and max-pooling in frequency domain to normalize speaker variance to achieve higher multi-speaker speech recognition performance. They proposed CNN architecture to evaluate speaker independent speech recognition task using the standard TIMIT datasets

Table 2.1 Summary of prior works on language identification system.

Sr. No	Features	Classification	No. of Lang.	Remark %	Refer ence
1	MFCC,LPC	SVM, Random Forest Techniques	6	92.6	[1]
2	MFCC, Delta, Acceleration	ANN, HMM, GMM	9	80.56	[2]
3	MFCC	GMM	15	87.60	[3]
4	MFCC, SDC	GMM	4	93.60	[4]
5	MFCC	GMM, UBM, GMM-UBM	27	88.40	[5]
6	SDC, Cepstral	GMM-UBM	10	81.70	[6]
7	SDC, Pitch Counter	GMM, SVM	2	79.60	[7]
8	Phonological	HMM	5	79.00	[10]
9	MFCC, SDC	GMM	4	78.00	[12]
10	Phonotactic, Prosodic	ANN	7	72.00	[13]
11	MFCC	SVM	6	89.72	[21]
12	Rhythm, Vowel	GMM	7	88±5	[24]

III. Definition and Model

A. Feature Extraction

Speech sample carries special type of information that can be express speaker identity and language identity. This include language and speaker-related information such as phonemes, vocal tract shape, excitation source, behavior features, temporal information and acceleration information. Temporal information and acceleration information can be used for language recognition. Feature extraction is a technique to extract these type of information from speech sample and representing this into parametric information. This parametric information knows as features then used in language identification. This system works on MFCC and SDC feature extraction technique [12].

- **Mel Frequency Cepstral Coefficient**

Mel frequency Cepstral coefficient feature is mostly used in speaker recognition system. Using MFCC feature extraction technique can extract vocal tract shape, excitation source from speech sample. This information is necessary for speaker identification. By computing delta of MFCC feature can get temporal and acceleration information of speech signal and that will be used in language identification. In this system combination of MFCC and SDC features are used to identify language [8].

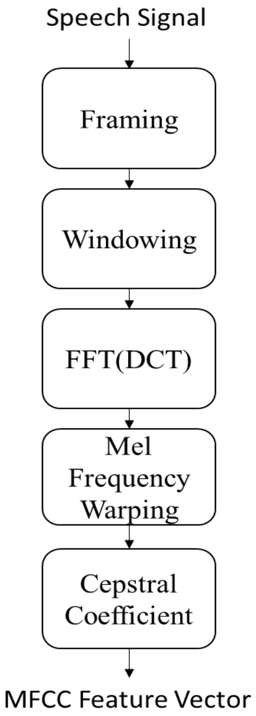

Fig. 3.1 Illustrates steps in MFCC feature extraction.

Fig. 3.1 is block diagram for MFCC feature extraction technique illustrates the steps to compute MFCC feature vector. The speech signal is having a large amount of data and that is difficult to analyze at one time. Therefore perform framing on the speech signal to minimize difficulties in analysis of speech signal. In framing speech signal is divided into small parts of 20 or 30 milliseconds known as frames. So short time analysis is possible by framing then its properties are fairly stationary. Framing is a process of cutting speech signal on some time duration so there is a possibility to lose some points from a signal and hence the loss in information. To overcome this problem windowing is applied. Thatis 2nd step in MFCC feature extraction technique. In this system hamming window is used to applying windowing function. Next step is applying Fast Fourier Transform on speech frames. FFT is an algorithm to compute discrete Fourier transform of speech signals.

Discrete Fourier transform is a process of converting time domain signal to frequency domain signal which helps analyze speech signal easily. Next step is to applying Mel frequency warping on signal to make the feature more closely what humans hear. Mel frequency warping converts the real frequency scale into human perceived frequency scale as per human auditory system. Mel frequency warping is done by applying filter banks on the speech signal. The formula for converting from frequency to Mel scale:

$$Mel\ (f) = 2595.log_{10}(1+f\ /\ 700)$$

The final step to compute MFCC feature vector is calculating log and applying inverse DCT on the speech signal.

- **Shifted Delta Cepstral Coefficient**

Language identification system improves the performance by combining MFCC and SDC feature. SDC feature gives the information about phonemes of languages. Every language is having a unique phoneme. By this information language identification system identifies language correctly. SDC feature vector is calculated by stacking delta Cepstral calculated across multiple speech frames [2]. In SDC feature extraction shifting delta, operation is applied to frame-basedacoustic feature vectors in order to create the newly combined featurevectors for each frame.To compute SDC feature vector we have to pass four parameters namely written as N-d-p-k. SDC feature is calculated from MFCC features so N is a number of coefficient in MFCC vector.The parameter d determine the spread over which deltas are computed, and the parameter P determines the gaps between successive delta computations and k defines the number of blocks [12]. For time t obtain:

$$\triangle\, c(t,i) = c(t+iP+d) - c(t+iP-d)$$

For i =0, 1, 2, 3,..., k-1

The stacked SDC coefficients are:

$$SDC(t) = [\triangle\, c(t,0)^t \triangle\, c(t,1)^t \ldots..\triangle\, c(t,k-1)^t]^t$$

B. *Classification Technique*

- **Gaussian Mixture Model**

Gaussian Mixture Model has mostly used classifier in language identification system to classify languages by using acoustic features of a speech signal. In this work, GMM is used to create language models for individual languages to train. In this EM algorithm is used to find out parameters and K-means algorithm is used to clustering the languages models. GMM is mostly used the parametric model of probability distribution.A GMM is a probabilistic model for density estimation using a mixture distribution and is defined as a weighted sum of multivariate Gaussian densities [2].The Gaussian mixture model is a weighted sum of Gaussian distributions which is able to model an arbitrary distribution of observations. The likelihood of a GMM model l for an observation x is given as below:

$$P(x|\boldsymbol{\lambda}) = \sum_{m=1}^{M} w_m p_m(x),$$

Where $w_m$is the weight of the mth Gaussian density$p_m(x)$:

$$p_m(x) = \frac{1}{(2\pi)^{\frac{D}{2}}|\sum_m|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(x-\mu_m)\sum_m^{-1}(x-\mu_m)\}$$

In above equation, μ_m and $\sum_m$are the mean vector and the covariance matrix of the mth Gaussian, respectively. Additionally in first equation, $\sum_m^M w_m$= 1 and w_m> 0.

IV. EXPERIMENTAL SETUP

A. *Training Data and Testing Data*

This system works with 17 languages as Eastern Arabic, Bengali, German, Hindi, Hungarian, Japanese, Kannada, Malayalam, Kashmiri, Portuguese, Urdu, Russian, Spanish, Marathi, Tamil, Panjabi, and English. We trained 300 speech samples of 17 different languages which are collected by CAIR, DRDO. The trained speech samples are of 40 to 60 sec all are wave files. It takes 20 to 25 minutes to train system with these 300 speech samples. Duration of testing speech sample is 30 to 40 seconds. The test set consist of 90 such speech samples to identify.

B. *Model Training*

MFCC and SDC feature vectors are computed for all 17 languages. 17 MFCC features and 54 SDC features are computed and combined together to create language model using GMM. Multivariate GMM consisting of 17, 54 mixture components were trained.

V. RESULT AND DISCUSSION

A. *Feature Extraction*

This system mainly works to combine MFCC and SDC features and trained with GMM. The system first computes MFCC feature vector. To compute MFCC feature speech samples are divided into N numbers of frames with 30 ms duration. Below fig. 5.1 shows an example of a single frame of the speech sample.

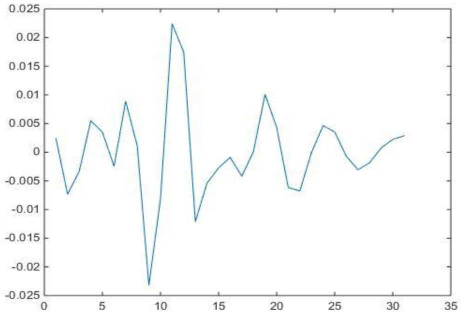

Fig. 5.1 Single frame of 30 ms.

We computed MFCC feature vector of 12, 15 and 17 no of the coefficient for each speech sample. The figure gives the graphical representation of MFCC feature vector of 12 coefficients.

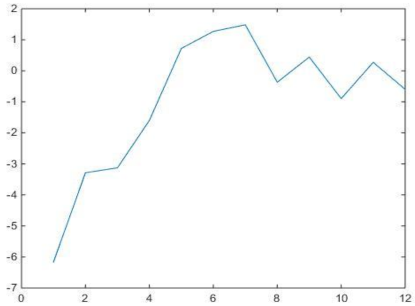

Fig. 5.2MFCC spectrum.

To compute SDC first need to calculate delta of MFCC features. For 12 coefficient of MFCC feature of single frame 12 delta coefficients are calculated. The figure shows a graphical representation of MFCC delta coefficients of a single frame.

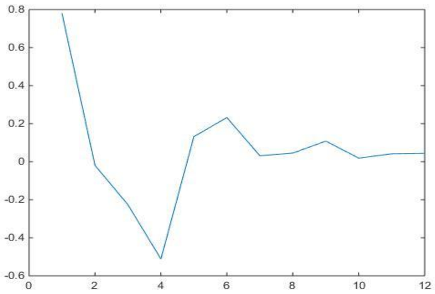

Fig. 5.3 MFCC delta spectrum

Then Computed SDC features by using MFCC and MFCC delta coefficients. In this system to compute SDC features, we have taken parameter (N, d, P, k) as (12, 3, 1, 3) for 12 MFCC coefficients and (17, 3, 1, 3) for 17 MFCC coefficients. The figure shows a graphical representation of SDC features of a single frame.

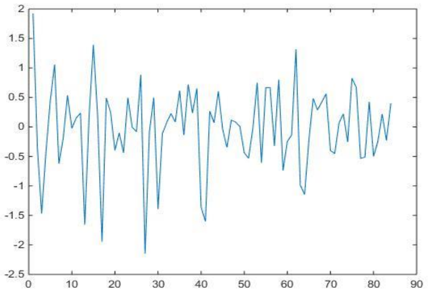

Fig. 5.4Shifted Delta Coefficients.

B. *Results of Overall System*

This system is tested with 3 different combinations of features. For 12 MFCC coefficient computed 24 MFCC and delta features and 36 SDC features. The testing accuracy of this combination is 76.47%.For 15 MFCC coefficient computed 30 MFCC and delta features and 45 SDC features. The testing accuracy of this combination is 88.23%.For 17 MFCC coefficient computed 34 MFCC and delta features and 51 SDC features. The testing accuracy of this combination is 94.12%.This system gives a better result for third combination that is 17 MFCC, 34 delta and 51 SDC coefficients than others.

VI. CONCLUSION

Language identification system is a system to identify language from speech samples. This LID system is based on GMM model and MFCC and SDC features. From this system, we observed that combining MFCC and SDC features gives better results to identify the language. With this system, we are getting 94.12% of accuracy in 17 languages. By this results, we observed that using more number of features to train, gives a better accuracy of identifying language but increase in computational work and time. Further work is to a testing system with more number of languages and decreasing computational work by applying various preprocessing methods. System accuracy of classifying languages will be improved by combining two classifiers.

REFERENCES

[1] M. Gupta, S. Bharati and S. Agarwal, "Implicit Language Identification System based on Random Forest and Support Vector Machine for Speech", IEEE conference on Power, Control & Embedded Systems (ICPCES), pp. 1-6, 2017.

[2]S. Jothilakshmi, V. Ramalingam, and S. Palanivel, "A hierarchical language identification system for Indian languages", Digital Signal Processing, vol. 22, no. 3, pp. 544-553, 2012.

[3]S. Koolagudi, D. Rastogi and K. Rao, "Identification of Language using Mel-Frequency Cepstral Coefficients (MFCC)", *Procedia Engineering*, vol. 38, pp. 3391-3398, 2012.

[4]K. Sarmah and U. Bhattacharjee, "GMM based Language Identification using MFCC and SDC Features", *International Journal of Computer Applications*, vol. 85, no. 5, pp. 36-42, 2014.

[5]V. Kumar, H. Vydana and A. Vuppala, "Significance of GMM-UBM based Modelling for Indian Language Identification", *Procedia Computer Science*, vol. 54, pp. 231-236, 2015.

[6]B. Yin and F. Chen, "Combining cepstral and prosodic features in language identification", *Pattern Recognition, IEEE*, vol. 4, pp. 254-257, 2006.

[7]H. Jue, Y. Liu and T. Fang, "Using cepstral and prosodic features for Chinese accent identification", *7th International Symposium on Chinese Spoken Language Processing (ISCSLP), IEEE*, pp. 177-181, 2010.

[8] SangwanAbhijeet, MahnooshMehrabani, and John HL Hansen, "Automatic language analysis and identification based on speech production knowledge." *IEEE International Conference*, 2010.

[9] Rizvi M, Akram B, and Sheikh MJ, "Language identification from raw speech", IEEE Students Conference, vol. 1, pp. 27--33, 2002.

[10] F. Pellegrino and J. Rouas, "An unsupervised approach to language identification", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, no. pp. 833-836, 1999.

[11] I. Yoshiko and G. Larry, "Automatic language identification using large vocabulary continuous speech recognition", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 785-788, 1996.

[12] A. Kumar and H. Hemani, "Effective Preprocessing of Speech and Acoustic Features Extraction for Spoken Language Identification", *Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), IEEE*, pp. 81-88, 2015.

[13] C. Madhu and A. Goerge, " Automatic Language Identification for Seven Indian Languages using Higher Level Features ", *Signal Processing, Informatics, Communication, and Energy Systems (SPICES), IEEE*, pp. 1-6, 2017.

[14] L. Feng, Speaker recognition. Lyngby, 2004.

[15] M. Savic, E. Acosta and S. K. Gupta, "An automatic language identification system," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 817-820 vol.2. 1991.

[16] F. Ernawan, N. Abu and N. Suryana, "Spectrum analysis of speech recognition via discrete Tchebichef transform", *International Society for Optics and Photonics*, vol. 8285, p. 82856L, 2011.

[17] H. Wang, C. Leung, T. Lee, B. Ma and H. Li, "Shifted-Delta MLP Features for Spoken Language Recognition", *IEEE Signal Processing Letters*, vol. 20, no. 1, pp. 15-18, 2013.

[18] M. Kockmann, L. Burget, "Application of speaker- and language identification state-of-the-art techniques for emotion recognition", *Speech Communication*, vol. 53, no. 9-10, pp. 1172-1185, 2011.

[19] X. Wang, Y. Wan, L. Yang, R. Zhou and Y. Yan, "Phonotactic language recognition using dynamic pronunciation and language branch discriminative information", *Speech Communication*, vol. 75, pp. 50-61, 2015.

[20] S. Sadjadi and J. Hansen, "Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification", *Speech Communication*, vol. 72, pp. 138-148, 2015.

[21]Y. Ben Ayed and R. Chakroun, "Automatic Language Identification in speech streams", *IEEE, International Multi-Conference on Systems, Signals and Devices (SSD)*, pp. 1-4, 2012.

[22]Gonzalez-Dominguez, Javier, "Frame-by-frame language identification in short utterances using deep neural networks." *Neural Networks* pp. 49-58, 2015.

[23]Lopez-Moreno I, Gonzalez-Dominguez J, Plchot O, Martinez D, Gonzalez-Rodriguez J, Moreno P. “Automatic language identification using deep neural networks.” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5337-5341, 2014.

[24] Rouas, Jean-Luc, "Rhythmic unit extraction and modeling for automatic language identification." *Speech Communication* pp.436-456, 2005

[25] Abdel-Hamid O, Mohamed AR, Jiang H, Penn G.," Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition.", *IEEE International Conference on In Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4277-4280, 2012.