International Journal of Electrical and Computer Engineering (IJECE)
Vol. 9, No. 4, August 2019, pp. 2443~2450
ISSN: 2088-8708, DOI: 10.11591/ijece.v9i4.pp2443-2450

High level speaker specific features as an efficiency enhancing parameters in speaker recognition system

Satyanand Singh
School of Electrical and Electronics Engineering, Fiji National University, Fiji Island

Article Info

Article history:

Received Dec 5, 2018
Revised Jan 21, 2019
Accepted Mar 11, 2019

Keywords:

Automatic speaker recognition (ASR)
Confidence measure (CM)
Deep neural networks (DNN)
Gaussian mixer model (GMM)
Mel-frequency cepstral coefficients (MFCC)

ABSTRACT

In this paper, I present high-level speaker specific feature extraction considering intonation, linguistics rhythm, linguistics stress, prosodic features directly from speech signals. I assume that the rhythm is related to language units such as syllables and appears as changes in measurable parameters such as fundamental frequency F_o, duration, and energy. In this work, the syllable type features are selected as the basic unit for expressing the prosodic features. The approximate segmentation of continuous speech to syllable units is achieved by automatically locating the vowel starting point. The knowledge of high-level speaker's specific speakers is used as a reference for extracting the prosodic features of the speech signal. High-level speaker-specific features extracted using this method may be useful in applications such as speaker recognition where explicit phoneme/syllable boundaries are not readily available. The efficiency of the particular characteristics of the specific features used for automatic speaker recognition was evaluated on TIMIT and HTIMIT corpora initially sampled in the TIMIT at 16 kHz to 8 kHz. In summary, the experiment, the basic discriminating system, and the HMM system are formed on TIMIT corpus with a set of 48 phonemes. Proposed ASR system shows 1.99%, 2.10%, 2.16% and 2.19 % of efficiency improvements compare to traditional ASR system for <10 ms, <20 ms, <30 ms and <40 ms of 16KHz TIMIT utterances.

Corresponding Author:

Satyanand Singh,
School of Electrical and Electronics Engineering,
Fiji National University, Fiji Island.
Email: satyanand.singh@fnu.ac.fj

1. INTRODUCTION

The language is mainly for passing information from one person to other person in day to day life [1]. It is transmitted through a series of legal sound units. This sequence must respect the constraints imposed by the language. Therefore, speech and language are compliments of each other and it can not be separated. Because each speaker has unique physiological characteristics of speech and speech generation style and speaker-specific features are also integrated into the speech signal. Therefore, the speech signal contains not only the expected message but also the language and speaker specific characteristics. In addition, the emotional state of the speaker is also transmitted through words [2, 3]. The speech message part is mainly expressed as a series of legal sound units, each corresponding to the manner and location of speech production by a particular sound unit. The language, emotions and speaker parts of the information contained in the speech signal are derived using several levels of functionality. Existing speaker, language, emotion, and speech recognition systems rely on features derived from the short-term spectral analysis. However, the spectral characteristics are affected by channel and noise characteristics. This has prompted researchers to explore the use of additional features that may provide additional evidence of a spectrum-based system.

Speech processing research aims to implement machines capable of performing automatic speech recognition, speech synthesis, speaker recognition, and many other speech processing tasks such as speech recognition by machine like a human [4, 5]. The researchers succeeded in developing speech systems operating in a restricted environment. Many of these systems rely solely on acoustic models formed using spectral characteristics. These acoustic models lack much higher-level information that humans use for the same task. The highest levels of information include prosody, context, and vocabulary knowledge.

It is understood that the introduction of the knowledge of prosody into automatic speaker recognition (ASR) system of the vocal systems will make them more intelligent and similar to humans [6]. Various researchers in the past have established the importance of prosodic features for speech processing applications [7]. Unfortunately, incorporation of prosody into the speech systems has to address several issues. One major issue is the automatic extraction and representation of prosody and its application in speaker recognition to enhance the efficiency of ASR. Our fundamental understanding of the processes in most of the speech perception modules in Figure 1 is rudimentary at best, but it is generally agreed that some physical correlate of each of the steps in the speech perception model occur within the human brain, and thus the entire model is useful for thinking about the processes that occur.

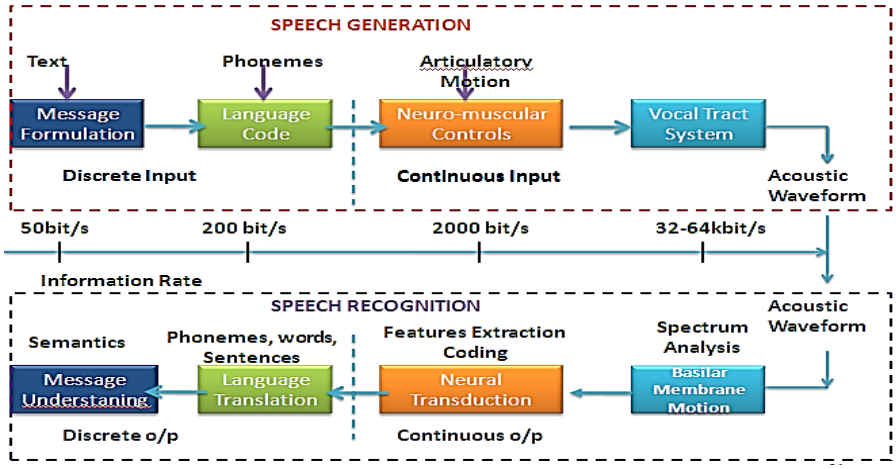

Figure 1. The Speech Generation Chain of a Normal Auditory System

2. PROSODY HIGH LEVEL SPEAKER SPECIFIC FEATURES IN SPEAKER RECOGNITION

Short-term cepstral features are often referred to as low levels reflects the speaker's voice rather than capturing high level speaker specific features, rhythm, and vocabulary information. Unfortunately, some prosodic features are very difficult to calculate, while others are difficult to deduce solely from acoustics (eg, the roundness of lips). As a result, more and more features are receiving increasing attention over the past decade.

Speech is transmitted through a series of legal sound units in the language. With the order of sound units, some built-in features give a natural voice. The pitch change provides identifiable melody attributes for speech. This controlled modulation of sound is called intonation. The unit of sound is shortened or lengthened according to some basic modes to give a certain rhythm to the voice.

There are few syllables or words may be more important than others, causing language pressure. The intonation, rhythm, and pressure of speech increase the intelligibility of speech information, allowing listeners to easily divide continuous speech into sentences and words [8]. It also conveys more vocabulary and non-verbal information such as vocal tones, loud tones, accents, and emotions. The characteristics that make us perceive these effects are collectively called prosody. Human Prosody is used to obtain information such as emotions, word/sentence boundaries, speaker characteristics and language characteristics, which are used for speaker identification. Each prompt is a complex perceptual entity mainly represented by three acoustic parameters: tone, energy, and duration.

2.1. Intonation as speaker specific features in ASR system

Pitch is the perceived property of sound and can be described as a perception of sound relative to "pitch" [9]. The physical correlation of pitch is the fundamental frequency (F_0) determined by the vibrational

rate of the vocal cords. The set of pitch changes during speech is defined as intonation [10]. The F_0 range depends on the length and mass of the vocal cord of a speaker. For men, it is usually between 80 and 200 Hz, and for women, between 180 and 400 Hz (approximately) for conversational speech. In this range, each speaker can produce an increase and a decrease of F_0. The direction in which F_0 changes (up or down) is determined by the speech pattern that makes up the word. Figure 2 shows the evolution F_0 of the man statement in "**College of Engineering Science & Technology"**. It can be observed that the value F_0 is discontinuous because of the periodic character of the speech during the vocal region (vowel, nasal, semitone, consonant, etc.) and the aperiodic nature during the silent region. The small disturbances in the F_0 process are mainly due to the involuntary aspects of the language.

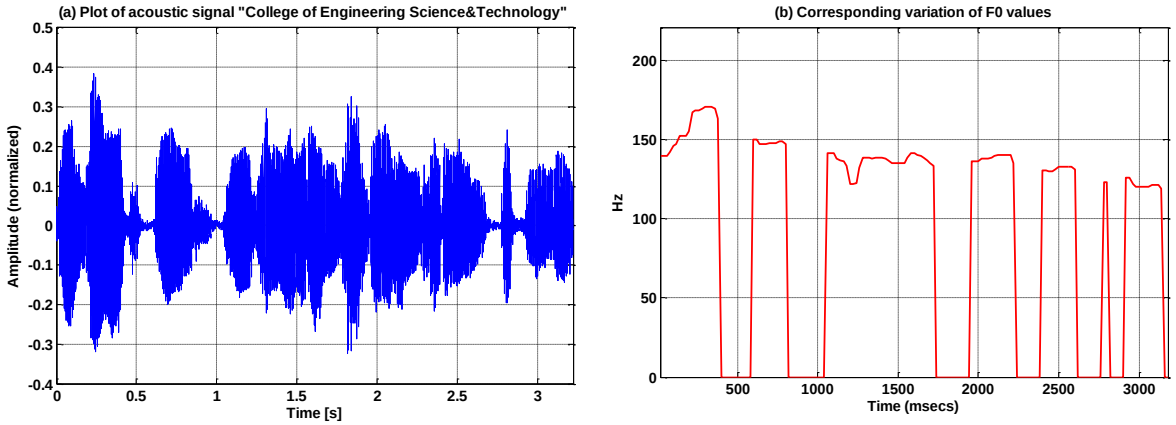

Figure 2. Variation of F_0 for the Utterance College of Engineering Science & Technology words stressed

2.2. Linguistics Stress as Speaker Specific Features in ASR System

In linguistics, stress is the ability to give relative importance to certain syllables or sentences of a word or to certain words of a sentence. There are pressures in many languages of the world. Stress is an attribute of the structural language of a word that indicates which syllable in a word is stronger in one sense than any other syllable. One of the important themes of the research on speech has always been the emphasis on the acoustic and perceptual characteristics of attributes: syllables are distinguished from unstressed syllables that surround them, or in a more controllable way, the emphasis on syllables differs from the unaccented implementation of the same syllable. The introduction of the knowledge of prosody into automation of the ASR systems will make them more intelligent and similar to humans [11].

2.3. Linguistics Rhythm as Speaker Specific Features in ASR System

The rhythm corresponds to the total duration of speech. Several experiments were conducted to study the rhythmic pattern of speech by replacing the original syllable with a meaningless syllable, preserving the original duration / duration and the original stress pattern. For example, "MAN in STREET" mimics "adDadaDa" in which capital letters are accented, assuming that the syllable is the basic unit of speech synchronization. This can be done in two ways, either to preserve the tone pattern of the original utterance, or to remain monotonous. This experiment deals with the temporal models associated with the perceived structure of the phonetic rhythms are no longer emphasizes the aspects that are not enhancing efficienicy of ASR [12]. Even in the absence of language, babies are able to recognize the familiar knowledge of rhythmic patterns. However, different modes that cause continuous changes can not be easily separated.

Histogram of all the pitch periods found in the speech signal college distributed acording to their fundamental frequency is shown in Figure 3. Where Alt Tx is a histogram of all the pitch periods, regular Tx is a histogram of all the regular pitch periods and Qx is a histogram of the closed quotient values distributed acording to their fundamental frequency F_o. The closed quotient is an estimate of ther pecentage time the vocal folds remained closed in each pitch period. Jitter is a measure of period-to-period fluctuations in duration from the mean taken over 5 pitch cycles.

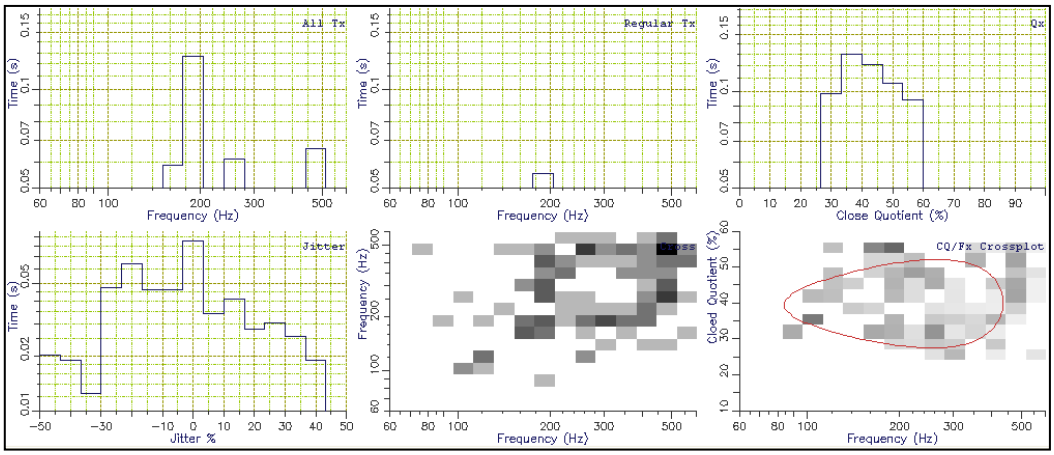

Figure 3. Histogram of all the pitch periods found in the recording, distributed acording to their fundamental frequency

3. PROBABILISTIC FORMULATION OF HIGH LEVEL SPEAKER SPECIFIC FEATURES

Identify the problem that can be expressed as the most likely speaker or language or emotional or sound B* of input speech from known speakers or languages or moods or sound units. Let $\{B_i\}, 1 \leq j \leq N$ is denoted the set of classes which is representing language, speaker and sount unit. The observation derived from the input of sample speech signal is denoted by O. The probabilistic formulation of high level speaker specific features can be formulated as follows:

$$B^* = arm^{max}_j P(B_i|O) \quad (1)$$

Where posterior probability of class B_i is represented as $P(B_i|O)$ for a considered speech signal utterance of a speker expressed in terms of O. To represent probabilistic formulation assuming observation O belonging to on of the N classes $\{B_i\}, 1 \leq j \leq N$. As per rule defined in (1) the main aim is to choose the objective of class B_i for posterior probability $P(B_i|O)$ must be maximum for a given O. Applying Bayes rule to obtain posteriori probability,

$$P(B_i|O) = \frac{P(O|B_i)P(B_i)}{P(O)} \quad (2)$$

Where likelihood probability is represented as $P(O|B_i)$ of O which is corresponding to the class B_i . The priori probability of the class B_i is represented as $P(B_i)$. Then the problem can be formulated as follows:

$$B^* = arm^{max}_j \frac{P(O|B_i)P(B_i)}{P(O)} \quad (3)$$

There is no reason to consider overlapping the class, $P(B_i)$ can be considered equal for all classes of different speaker groups. Here $P(O)$ belongs to all classes, the probabilistic problem can be simplified to reduce the computational complexity as follows:

$$B^* = arm^{max}_j \ P(O|B_i) \quad (4)$$

Thus speaker or linguistic or emotional or speech recognition tasks are regarded as estimates of posterior probabilities and can be reduced to likelihood probability estimates under specific assumptions.

3.1. Speaker specific feature aspect of individual speech signal

Speaker characteristics vary due to the difference in physiological characteristics of speech production organs and acquired or learned habits. Features of ASR are roughly divided into four groups of continuous, qualitative, spectral, and teager based energy operator features, and prosodic features are classified into categories of continuous speech features [13]. Rhythmic features are reliable indicators of emotion and are widely used for emotional recognition [14]. The arousal state of the speaker has been studied

to influence the overall energy, frequency, and duration of the voice pause [15]. Emotions like anger are characterized by a high speech rate, but feelings of sadness are related to whispered speed.

Gaussian Mixer Model (GMM) [16] and neural network [17] were successfully used for emotional recognition. SVM is widely used by researchers to classify emotions [18]. Deep neural networks (DNN) can be used to obtain higher level features from low-level acoustic features and then to other classifiers for emotion recognition. In [19], features of the segmentation level including Mel-frequency Cepstral Coefficients (MFCC), pitch-based features (pitch period and harmonic to noise ratio), and their delta values are extracted.

3.2 Fusing higher speaker specific feature into conventional ASR application

The prosodic model provides an additional knowledge source that the acoustic model cannot provide. This may help to overcome some of the miss identifications. Therefore, combining information from multiple sources of evidence, known as fusion technology, has been widely used in speakers, languages, emotions, and speech. Typically, many different feature sets are extracted from the speech signal, then a separate classifier is used for each feature set, then sub-scores or decisions are combined. This means that each speaker stores a plurality of speaker models in the database. It is generally believed that successful fusion systems should be combined into independent features. Possible low-level spectral characteristics, prosodic function, advanced function. The simplest fusion method is to combine classifier output scores with weighted sums. That is, a given subscore s_k is a fusion match of the index classifier k.

$$s = \sum_{n=1}^{N_c} (w_i, s_k) \ where\ n_c = number\ of\ classifier, w_n = nth\ classifier \tag{5}$$

Another way to combine features at the score level is to use confidence measure. In [16], the author confirmed that confidence based fusion complementary feature method of combining wavelet multiplication coefficients residual (WOCOR) and MFCC function for speaker recognition. This metric is derived from the likelihood score obtained from the two features. In order to campute the confidence measure (CM), the discrimination ability of each feature in a particular recognition test is first calculated is given as follows:

$$D_j = LLR_j / \left| logP\left(s_j / \lambda_{Ij,j}\right) \right| \tag{6}$$

where

$$\mathrm{LLR_j} = \mathrm{logP}\left(\mathrm{s_j}/\lambda_{\mathrm{c,j}}\right) - \mathrm{logP}\left(\mathrm{s_j}/\lambda_{\mathrm{u,j}}\right) \tag{7}$$

The log-likelihoods of the client model and background model are represented in (6) and eqn. (7) respectively. The computation of the discrimination ratio based on the value function of each trial is $\mathrm{DR} = \mathrm{D_1}/\mathrm{D_2}$. Next, confidance metric is camputed based on the DR value as follows:

$$CM = -log\left(\frac{1}{1 + e^{\alpha(DR-\beta)}}\right) \tag{8}$$

The values of α and β were determined by setting the development data to 0.75 and 2, respectively. Based on CM, score level fusion is done, which represented as follows:

$$\mathrm{LLR} = \mathrm{LLR_1} + \mathrm{LLR_2}.\mathrm{CM} \tag{9}$$

As the fusion score combines weighted $\mathrm{LLR_1}$ and $\mathrm{LLR_2}$, this CM based scoring fusion method yields better results in terms of fixed weight fusion are represented as follows:
2.5403, 2.4960, 2.3400, 1.6502, 1.5585, 1.4669, 1.3768, 1.2097, 1.1377, 1.0320, 1.0054, 1.0239, 1.2813, 1.4358, 2.7555, 2.9106, 2.9807, 2.8617, 1.0193, 1.0392, 2.9974, 2.8392, 1.0696, 2.9909, 2.6717, 1.0231, 2.9789, 1.0293, 2.9191, 1.9560, 1.5887, 1.1993, 1.3973, 1.0538, 2.9990, 1.0577, 1.6913, 1.0357.

4. LARGE MARGIN APPROACH FOR LEARNING ALIGNMENT IN ASR APPLICATION

A supervised learning algorithm for alignment receives a training set as input $\mathcal{T} = \{(\bar{\mathrm{x}}_1, \bar{\mathrm{p}}_1, \bar{\mathrm{s}}_1)\}, \ldots\ldots \{(\bar{\mathrm{x}}_{\mathrm{m}-1}, \bar{\mathrm{p}}_{\mathrm{m}-1}, \bar{\mathrm{s}}_{\mathrm{m}-1})\}, \{(\bar{\mathrm{x}}_{\mathrm{m}}, \bar{\mathrm{p}}_{\mathrm{m}}, \bar{\mathrm{s}}_{\mathrm{m}})\}$ which returns a alinmened function f.

To promote efficient algorithms, I restrict to a limited kind of alignment function. More specifically, it is assumed that there is a set of predefined basic aligned feature functions$\{\varphi_j\}_{j=1}^{n}$. Each base alignment speaker specific feature is a function of the form $\varphi_j: \mathcal{X}^*.\mathcal{P}^*.\mathbb{N}^* \rightarrow \mathbb{R}$. That is, each basic alignment speaker specific feature $\bar{x}$ and speaker specific phoneme sequence $\bar{p}$, together with the candidate timing sequence $\bar{s}$. $\varphi\{\bar{x}, \bar{p}, \bar{s}\}$ is denoting $\mathbb{R}^n$ vector whose jth elment is $\varphi_j\{\bar{x}, \bar{p}, \bar{s}\}$. In this paper I am using the the alignment function defined as follows:

$$f(\bar{x}, \bar{p}) = arg \max_{\bar{s}} W . \varphi\{\bar{x}, \bar{p}, \bar{s}\} \quad (10)$$

With the SVM algorithm for binary classification, the method of selecting the weight vector w is based on the concept of large margin separation. But in this case, timing is not just right or wrong. Therefore, my goal is not to separate the right timing from the wrong timing, but to try to sort the sequence by quality. In theory, my method can be described as a two-step process: first build a vector $\varphi\{\bar{x}, \bar{p}, \bar{s}'\}$ in vector space $\mathbb{R}^n$ in insident based approach $(\bar{x}_i, \bar{p}_i)$ in a training set $\mathcal{T}$ and each possible timing sequence $\bar{s}'$. Second I find a vector K that projects the vector to w and sorts the vectors $w \in \mathbb{R}^n$constructed in the first step above based on its quality. Ideally, for each instance $(\bar{x}_i, \bar{p}_i)$ and every suggestable timing to keep the following constraints:

$$W. \varphi(\bar{x}_i, \bar{p}_i, \bar{s}_i) - w. \varphi(\bar{x}_i, \bar{p}_i, \bar{s}') \geq \gamma(\bar{s}_i, \bar{s}') \quad (11)$$

The computer simulated vectors $w \in \mathbb{R}^n$constructed in the first step based on its quality is represented as follow:
0.0000, 0.0505, 0.2020, 0.6566, 0.7071, 0.7576, 0.8081, 0.9091, 0.9596, 1.0606, 1.1111, 1.2121, 1.3636, 1.4141, 1.7172, 1.7677, 1.8687, 1.9192, 2.2222, 2.2727, 2.5253, 2.5758, 2.7778, 2.9293, 2.9798, 3.0808, 3.2323, 3.3333, 3.4343, 3.5859, 3.7374, 3.9899, 4.2929, 4.4444, 4.5455, 4.6970, 4.7980, 4.9495.

Where $\gamma(\bar{s}_i, \bar{s}')$ is cost function assessing the quality of sequences. The constraint of the expression in (10) means that the margin of w with respect to possible timing sequence $\bar{s}'$ must be greater than the cost of the prediction $\bar{s}'$, not the true timing $\bar{s}_i$. Of course, if the w rank is different and the possible timing is calculated correctly, the margin requirement given by (11) can simply be satisfied by multiplying w by a large scalar. The SVM algorithm is subjected to the constraints given in (11) by minimizing $\frac{1}{2}\|w\|^2$ which solves this problem represented in (10). Decision boundaries of multiclass SVM vs confidence measure is shown in Figure 4.

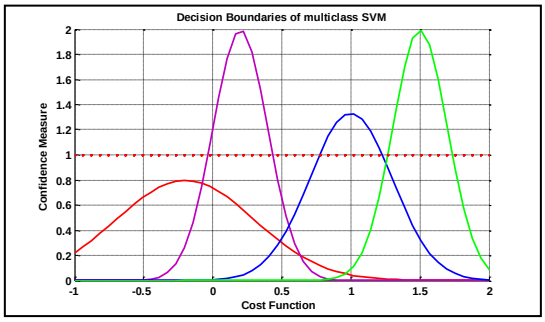

Figure 4. Decision Boundaries of Multiclass SVM

In fact, there are cases where the constraint given by (11) can not be satisfied. In order to overcome this obstacle, the following hinge loss function is defined for alignment in accordance with the soft SVM method. Large margine approach for learning alignment of speaker1 and speakers 2 is shown in Figure 5. Computer simulated cost function $\gamma(\bar{s}_i, \bar{s}')$ assessing the quality of sequences is represented as follows:
0.2423, 0.1406, 0.1031, -0.1844, -0.0233, -0.0162, -0.2758, -0.2269, -0.2046, -0.4617, -0.0764, -0.5951, -0.4095, -0.0877, 0.1738, 0.3080, 0.3705, 0.2608, -0.5248, -0.5161, 0.6893, 0.2658, -1.4624, 0.9442, 1.6271, -3.1532, 3.4593, -3.5774, 1.8691, -0.5880, -0.2175, 0.3431, 0.2986, -0.9581, 1.16863, -1.7063, 0.8625, -1.1150.

Table. 1 represents building large margin distance matrix for learning alignment with respect to cost function, confidance measure, support and up saturation.

Table 1. Building large margin distance matrix for learning alignment

Cost Function	Confidance Measure	#Support	#Up Saturation
1.2450	1.0000	1	0
0.0000	1.0000	2	0
-1.0451	1.0451	3	0
-2.4539	1.3479	4	0
-9.7113	2.9575	5	0
-3.5856	2.6922	4	0
-4.0236	0.1221	5	0
-5.8779	0.4609	6	0
-6.7987	0.1567	7	0
-1.7125	1.5188	6	0
-2.4501	0.4308	5	0
-2.4892	0.0160	6	0
-2.5273	0.0153	7	0
-4.9080	0.9420	8	0
-6.2026	0.2638	9	0
-8.2675	0.3329	8	0
-8.3876	0.0145	9	0
-8.9107	0.0624	8	0
-9.9513	0.1168	9	0

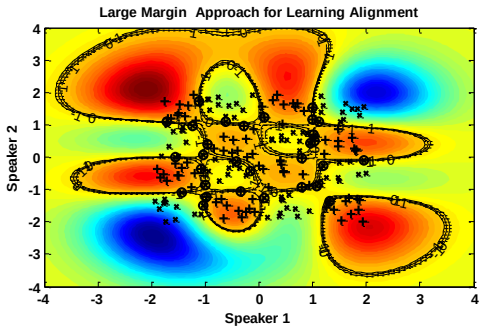

Figure 5. Large Margine Approach for Learning Alignment of Speaker1 and Speakers 2.

5. EXPERMENTAL SETUP FOR ACCURACY AND ROBUSTNESS OF ASR APPLICATION

I performed a speech-to-speech experiment using the TIMIT and HTIMIT corpus originally sampled in the 16 kHz TIMIT corp at 8 kHz. In all, the experiment, the baseline discriminative system and the HMM system were trained on a legible, readable TIMIT corpus with a set of 48 phonemes. I divided the training part of TIMIT expressions of SA1 and SA2 from independent part into three disjoint part excluding the expressions of 500, 100 and 3093 respectively. The second and third sets of utterances 100 and 3096 respectively formed a ASR recognition set and ASR training set for the forced alignment algorithm.

Although TIMIT was originally sampled at 16 kHz, confirmed the discrimination model by training the 8 kHz TIMIT. This was done to evaluate the algorithm's performance at a more realistic sampling rate and to use the trained ASR model on a sub-sampled TIMIT on another 8 kHz sampled corpus. The experimental results of ASR are shown in Table 2. It can be seen that the results of TIMIT sub-samples at 8 kHz still exceed the results reported by Brugnara et al. [20] and Hosom [21].

Table 2. Predefined tolerances of TIMIT corpus, the efficiency of phonemes positioned correctly

	Length of Speech Corpus			
	Less than 10ms	Less than 20ms	Less than 30ms	Less than 30ms
		16KHz TIMIT Utterances		
Discriminative Alignment	76.5	90.77	96.44	99.23
Brugnara et al. (1993)	75.3	88.9	94.4	97.1
Hosom (2002)		92.6		
		8 KHz TIMIT Utterances		
Discriminative Alignment	84.23	94.21	97.12	99.1
		HTIMIT CB1 834 Utterances		
Discriminative Alignment	72.5	89.68	95.89	97.4

6. CONCLUSION

ASR is the use of a machine to identify a person from the spoken word. The ASR system can be used in two modes that recognize the identity required by a particular person or verifier. Basic knowledge of speaker recognition is covered, simple functions and measures for speaker recognition are proposed and compared with conventional recognition using speaker recognition criteria. Proposal of the ASR system to distinguish speakers uses high-level speaker-specific features. This measurement can be interpreted as the shape of information theory measurement by discrimination alignment. In fact, the experiments reported above show that discriminative training requires less sample training than speech-to-phone alignment based on HMM processes. The performance of speaker recognition accuracy of the ASR system proposed in this paper is 99.23% and 99.1% for $< 40\ ms$ of TIMIT utterances respectively.Proposed ASR system shows 1.99%, 2.10%, 2.16% and 2.19 % of improvements compare to traditional ASR system for < 10 ms, < 20 ms, < 30 ms and < 40 ms of 16KHz TIMIT utterances. The proposed ASR system introduced here is actually realized with MATLAB on a moderate personal computer.

REFERENCES

[1] S.Singh, "Forensic and Automatic Speaker Recognition System," *International Journal of Applied Engineering Research* , SSN 0973-4562 Volume 8, Number 5, 2018, pp. 2804-2811, 2018.

[2] S.Singh, Evaluation of Sparsification algorithm and Its Application in Speaker Recognition System," *Journal of Applied Engineering Research*, ISSN: 0973-4562, Volume 13, Number 17, pp. 13015-13021, 2018.

[3] S.Singh, "Support Vector Machine Based Approaches For Real Time Automatic Speaker Recognition System," *International Journal of Applied Engineering Research*, ISSN: 0973-4562, Volume 13, Number 10, pp. 8561-8567, 2018.

[4] S.Singh "The Role of Speech Technology in Biometrics, Forensics and Man-Machine Interface," *International Journal of Electrical and Computer Engineering*, 2018.

[5] S.Singh, Assaf Mansour H, Abhay Kumar and Nitin Agrawal "Speaker Specific Phone Sequence and Support Vector Machines Telephonic Based Speaker Recognition System," *International Journal of Applied Engineering Research*, ISSN 0973-4562 Volume 12, Number 19, pp. 8026-8033, 2017.

[6] A. Eriksson, "Tutorial on forensic speech science," in *Proc. European Conf. Speech Communication and Technology*, Lisbon, Portugal, pp. 40-80, 2005.

[7] Waibel, A., "*Prosody and speech recognition,"* San Mateo: Morgan Kaufmann Publishers, 1988.

[8] Shriberg, E., Stolcke, A., Hakkani-Tur, D., & Tur, G, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, 32, pp. 127-154, 2000.

[9] Nooteboom, S, "The prosody of speech: Melody and rhythm," In The handbook of phonetic sciences. Blackwell handbooks in linguistics Malden, Blackwell Publishers, vol. 5, pp. 640-673, 1997.

[10] Hart, J., Collier, R., & Cohen, A, "A perceptual study of intonation," *Cambridge, UK: Cambridge University Press*, 1990.

[11] Shriberg, E., & Stolcke, A, "Direct modeling of prosody: An overview of applications in automatic speech processing," *In Speech Prosody*, Nara, Japan, pp.1-8, 2004.

[12] Raymond W. M. Ng , Tan Lee, Cheung-Chi Leung , Bin Ma, Haizhou Li, "Spoken Language Recognition With Prosodic Features," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 21, Issue. 9, pp-1841-1853, Sept. 2013.

[13] El Ayadi, M., Kamel, M. S., & Karray, F, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44(3), pp. 572-587, 2011.

[14] Busso, C., Lee, S., & Narayanan, S, " Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17(4), pp. 582-596, 2009.

[15] Luengo, I., Navas, E., Hernáez, I., & Sánchez, "Automatic emotion recognition using prosodic parameters," In *Proceedings of Interspeech*. pp. 493-496, 2005.

[16] Iliou, T., & Anagnostopoulos, C.-N, "Statistical evaluation of speech features for emotion recognition," In *Proceedings of Fourth International Conference on Digital Telecommunications (ICDT'09)*, vol. 1, pp. 121-126, 2009.

[17] Luengo, I., Eva, N., & Hernáez, I, "Feature analysis and evaluation for automatic emotion identification in speech," *IEEE Transactions on Multimedia*, vol. 12(6), 490-501, 2010.

[18] Han, K., Dong, Y., & Tashev, I, "Speech emotion recognition using deep neural network and extreme learning machine," In *Proceedings of Interspeech*, pp. 223-227, 2014.

[19] Zheng, N., Lee, T., & Ching, P.-C, " Integration of complementary acoustic features for speaker recognition," *IEEE Signal Processing Letters*, vol. 14(3), pp.181-84, 2007.

[20] F. Brugnara, D. Falavigna, and M. Omologo, "Automatic segmentation and labeling of speech based on hidden Markov models," *Speech Commun*., vol. 12, pp. 357-370, 1993.

[21] J.-P. Hosom, "Automatic phoneme alignment based on acoustic-phonetic modeling," in *Proc. 7th Int. Conf. Spoken Language Processing*, pp. 357-360, 2002.