DOI: 10.2298/CSIS120208053F

Speech Unit Category based Short Utterance Speaker Recognition

Nakhat Fatima[1], Xiaojun Wu[1] and Thomas Fang Zheng[1]*

[1] Center for Speech and Language Technologies, Division of Technical Innovation and Development, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, 100084, Beijing, China
fatima@cslt.riit.tsinghua.edu.cn, xjwu@tsinghua.edu.cn
* Corresponding Author: fzheng@tsinghua.edu.cn

Abstract. Information of speech units like vowels, consonants and syllables can be a kind of knowledge used in text-independent Short Utterance Speaker Recognition (SUSR) in a similar way as in text-dependent speaker recognition. In such tasks, data for each speech unit, especially at the time of recognition, is often not enough. Hence, it is not practical to use the full set of speech units because some of the units might not be well trained. To solve this problem, a method of using speech unit *categories* rather than individual phones is proposed for SUSR, wherein similar speech units are put together, hence solving the problem of sparse data. We define Vowel, Consonant, and Syllable Categories (VC, CC and SC) with Standard Chinese (Putonghua) as a reference. A speech utterance is recognized into VC, CC ad SC sequences which are used to train Universal Background Models (UBM) for each speech unit category in the training procedure, and to perform speech unit category dependent speaker recognition, respectively. Experimental results in Gaussian Mixture Model-Universal Background Model (GMM-UBM) based system give a relative equal error rate (EER) reduction of 54.50% and 40.95% from minimum EERs of VCs and SCs, respectively, for 2 seconds of test utterance compared with the existing SUSR systems.

Keywords: Short Utterance Speaker Recognition, Vowel Categories, Universal Background Vowel Category Model.

1. Introduction

In real-time speaker recognition there can be circumstances where obtaining appropriate speech data might become difficult. Conditions like noisy background, faulty devices or unwilling speakers, are a few of the many factors that might reduce the available amount of speech data. Hence, it becomes essential to make the most of the available data efficiently by employing short utterances of speech to identify a speaker. Short Utterance Speaker Recognition (SUSR) has recently emerged as an important area of research. SUSR technology attempts to use small amount of data from short utterances of speech for recognition purpose. In conventional speaker recognition systems, a large amount of data is required for a successful speaker identification system. However, SUSR attempts to use smaller amount of data for the recognition task. Different research endeavors have described different lengths for short utterance from around a minute to less than 10 seconds. However, recently the meaning of short utterance has taken a turn to mean less than 3 seconds of speech.

When utterance lengths are short, there is not enough variation is speech for an efficient speaker recognition task. Therefore, the performance of speaker recognition systems deteriorates sharply when utterance lengths are shorter than 10 seconds.

It is because of the declining performance of speaker recognition with short utterances that we were motivated to investigate the characteristics of sounds at phoneme level, which is the smallest meaningful unit of speech. The pronunciation idiosyncrasies of a speaker begin at phoneme level. Therefore we study the phonemes in different languages. To begin our study we limited ourselves to the vowels. We created a set of vowel categories to use them as phoneme sequences and then perform speaker recognition on 1-3 seconds of the sequence lengths. We extend our research onto other speech units like consonants and later the combination of vowels and consonants i.e. syllables.

Ultimately this research is intended to have the following contributions:

1. To describe the importance of phonemes and their combination (speech units like vowels and syllables) in Short Utterance speaker recognition.
2. Making the SUSR system essentially language and text independent by designing speech unit categories.
3. Improving the performance of speaker recognition by employing speech unit category sequences for training speaker models and for performing recognition with utterance lengths of 1-3 seconds.
4. Reduction of comparative training length by using speech unit categories.

1.1. Related Work

Generally, speaker recognition comprises of three steps: feature extraction, statistical modeling and score calculation. Most of the conventional speaker recognition systems perform feature extraction using various frame sizes shifts or vocal source features called wavelet octave coefficients of residues (WOCOR) [1, 2, 3]. For statistical modeling, eigenvoice and factor analysis subspace estimation are applied [4]. CGMM-UBM (Clustered GMM-UBM) and sub GMM-UBM methods are also used in many system, which find their basis in clustering and subspace estimation [5]. Though conventional, the improvements built on these techniques are widely used in speaker recognition. These methods, however, require a large amount of speech data for training as well as recognition purposes. This is the reason why these methods do not perform well when utterance length becomes shorter than a minute.

The innovative method of Phonetic or prosodic speaker recognition makes use of idiosyncrasies in a person's speech to identify a speaker e.g. pronunciation habits [6, 7]. This method, too, requires a large amount of speech processing and training data, again making it quite difficult to be performed when speech data is not enough.

In order to solve the problem of large data required to perform speaker recognition, research has lead the use of Factor Analysis, Joint Factor Analysis (JFA), Support Vector Machine (SVM) and I-vector based technologies [1, 6, 7]. In other works performing short utterance speaker recognition, Dimension decoupled GMM is applied [8]. Training and testing with 10 seconds of speech on variations of GMM and SVM have also shown improvements in results [7]. Since SUSR is a complicated task, it has also been used in combination with video aid for better results [9]. Mel-frequency cepstrum coefficients (MFCC) and Hidden Markov Models (HMM) based system has been employed in using phonemes for speaker recognition [10] which suggests that using larger amount of training data; phoneme level speaker recognition can be improved. A study of formant contours at consonant-vowel boundary has shown that a speaker can be identified using information at the consonant to vowel transition boundary, which makes use of the distinctive "speaker-style" [11]. The biannual National Institute of Standards and Technology (NIST) Speaker Recognition Evaluation (SRE) also presents the optional test of 10-second evaluation [12]. The length of utterance in most of these cases is around 10 seconds.

I-vector based technology has proven its vast advantage in recent studies [6]. The i-vector based SUSR achieves a minimum of 21.98% of EER for 2 seconds of test utterance. In performing SUSR, most of the research has been conducted using 10 seconds of utterances..

1.2. Performance of Existing SUSR Techniques

The term "Short Utterances" has taken various meanings and hence evolved in recent years, ranging from more than a minute to only a few seconds. Different studies have taken a different measure for short utterance. Previously most of the works defined short utterance to be around 10 seconds of speech [12]. However, as the need of short utterance speaker recognition grew, utterance lengths were defined to be shorter. Of late, research has taken a turn and utterances as short as 2-3 seconds are being used in order to recognize, validate and discern a speaker.

Working in the domain of short utterances, it has been observed that the overall performance decreases significantly when utterance length becomes smaller than 20 seconds. A comparison of different researches on SUSR [6, 7, 13, 14, 15, 16, 17] was made and it was observed that although performance of speaker recognition in terms of EER% remains close to 5-6% on average (minimum EER of 3.13% and maximum being 13.47% for more than a minute of test utterances) when utterance lengths are more than a minute long down to 20 seconds, there is a considerable drop in performance when utterance lengths go below 20 seconds of speech (EER reaching around 12% with 20 seconds of speech). The decline in performance is more marked when utterances are shorter than 10 seconds (average EER at 10 seconds from different systems being 18.32%). The average EER% of SUSR systems using 4 seconds of speech decline to 24.15% (min: 17.38%, max: 31.3%) and when 2 seconds of speech is used, the performance of different methods give an average of 28.94% EER (min: 21.98%, max: 36.16%). Hence, there is an exponential drop in performance when utterance lengths are below 10 seconds for test speech.

Considering the dwindling efficiency below 10 seconds, it becomes intriguing as well as incumbent that shorter utterances than 10 seconds be investigated for better results. Our motivation to strive for a solution to the problem of short utterances in speaker recognition comes from Phonetic Speaker Recognition (PSR). We attempt to find the usefulness of phoneme similarities in speaker recognition when using short utterances. Hence, we propose a category based short segment speaker recognition that can use vowel, consonant and syllable categories in SUSR.

The rest of the paper is organized as follows. In Section 2 we describe the framework of the Speaker Recognition system used. Experimentation details on continuous speech divided in random segments are given in Section 3. Section 4 discusses Vowel Categories and their performance in SUSR. This discussion and performance analysis is continued in Section 5 with Syllable, Consonant and improved Vowel Categories. Finally, we draw our conclusions and present future research prospects in Section 6.

2. Speaker Recognition System Framework

2.1. Proposal and Framework Overview

In our previous work [18], we devised an innovative design of speaker recognition, which was based on phoneme-classes for speaker recognition, creating a Phoneme Category Based Short Utterance Speaker Recognition (PCBSUSR). For this purpose, a vowel-class set was defined, i.e. a language-independent set of vowel categories. The vowel-class set was defined using linguistic knowledge of vowels. Consonants were not used for the study because consonants are more complex to recognize and categorize. The vowels categories were defined to help cover maximum number of the vowels in most languages. In training phase a group of vowel-class models was built with respect to each speaker making a speaker vowel-class model. In test phase the test utterance was first recognized into a sequence of phonemes and then speaker recognition was performed on the utterance using the vowel categories from the recognized phonemes. The system performance was not very efficient as it gave 46% EER. There were, however, many constraints during the experiments including quality of speech segmentation and recognition. In the subsequent study, initially the speech segmentation problem was addressed. We decided to use speech segments such that each segment was a complete unit of meaningful sound, vowels in this case. In this study, we devised five VCs. In [19], we presented Syllable categories (SCs), wherein utterances were recognized into syllables and then assigned to appropriate categories. The results of this study showed that the syllables are very significant speech with respect to speaker recognition and they contain a large amount of speaker information.

In the current research, we present how improvement in category classification can be beneficial to speaker recognition. Hence we propose new vowel categories and their improvement, and a new set of syllables categories (SC). In addition, we present consonant categories (CC) and compare their performance with vowel and syllable categories so that we can determine which type of speech units give better performance.

This research would help in understanding that though speech recognition plays an important role in text dependent speaker recognition, when the duration gets shorter than one second, exact phone recognition is not important. Instead, broader phoneme-classes/categories can provide good results as well. This makes the system reasonably text independent. The use of phoneme properties helps bring together similar phonemes under single category for speaker recognition. Also, this research will help in understanding that there are speaker related phonemic idiosyncrasies which can be put to use in SUSR by exploring the role of vowels and other speech units in speaker recognition. Furthermore, this study will strive to show that in speaker recognition with short utterances, it is important to use complete segments of phonemes rather than continuous speech cut in random segments. The random segments lose speaker information where a complete speech unit retains it.

2.2. Baseline System

Our baseline system has been organized as follows. Feature extraction is performed on a 20-millisecond frame every 10-milliseconds. The pre-emphasis coefficient is 0.97 and hamming windowing is applied to each pre-emphasized frame. Voice activity detection based on energy is performed with each frame, resulting in the labeling of either valid or invalid. 16-dimensional MFCC features are extracted from the utterances only for those valid frames with 30 triangular Mel filters used in the MFCC calculation. For each frame, the MFCC coefficients and their first delta coefficients form a 32-dimentional feature vector. Gender-independent UBM consists of 1,024 mixture components and are trained using EM algorithm [20]. For the MAP training [20], only mean vectors were adapted with a relevance factor of 16. The baseline system is a speaker verification system based on the conventional

GMM-UBM. Note, that previously [18], the EER of this system was 46% using phoneme categories.

2.3. Database

The experiments were performed using the “Annotated Speech Corpus of Chinese Discourse (ASCCD) from Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, China” [21]. The dataset consists of sentences uttered by 10 Chinese speakers (5 males and 5 females). Each speaker spoke 47 sentences of average duration of 25 seconds, in normal speech. 30 utterances of the speakers were selected to train the UBM while the remaining (17) utterances were used for testing purpose.

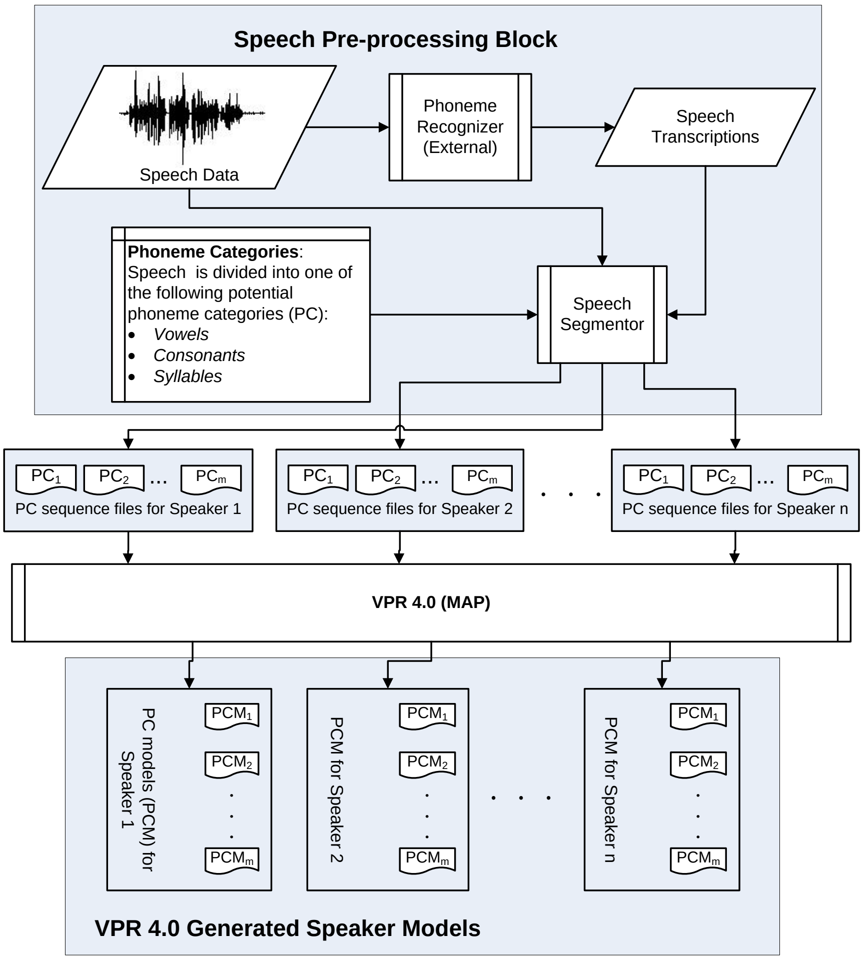

Speech Pre-processing Block
Speech Data
Phoneme Recognizer (External)
Speech Transcriptions
Phoneme Categories: Speech is divided into one of the following potential phoneme categories (PC):
• Vowels
• Consonants
• Syllables
Speech Segmentor
PC1 PC2 ... PCm
PC sequence files for Speaker 1
PC sequence files for Speaker 2
PC sequence files for Speaker n
VPR 4.0 (MAP)
PCM1 PCM2 PCMm
PC models (PCM) for Speaker 1
PCM for Speaker 2
PCM for Speaker n
VPR 4.0 Generated Speaker Models

Fig. 1. Block diagram of Model Training in SUSR using Phoneme Categories (PC)

2.4. Universal Background Model Training and Testing

In our experiments, a Universal Background Model (UBM) [7] is built based on the EM algorithm and the GMM modeling using MAP algorithm. During speaker recognition, the test utterance is scored against its corresponding UBM, giving *n* scores. These *n* scores are fused to obtain the final score.

We propose a SUSR framework which is described as follows:

The speech would first be passed through a phoneme recognizer, which would recognize the speech into phoneme categories sequences. This would result into the phoneme category sequence files for each speaker. For example category ‘a’ for “speaker A” would contain ‘a’ type sounds from the training utterance of “speaker A”. Similarly phoneme category sequence file of each phoneme category for every speaker would be created. These sequences would then be given as input to the GMM-UBM system wherein a Speaker Phoneme Category Model would be created for each Speaker Phoneme Category, resulting in speaker models against each phoneme category.

Later the same process of recognition of speech into phoneme categories would be applied during testing. The resulting phoneme category files of specified length would then be matched against the corresponding models and scored to give the recognition result.

Fig. 1 and Fig. 2 are block diagram representation of our proposed SUSR system with Phoneme Categories.

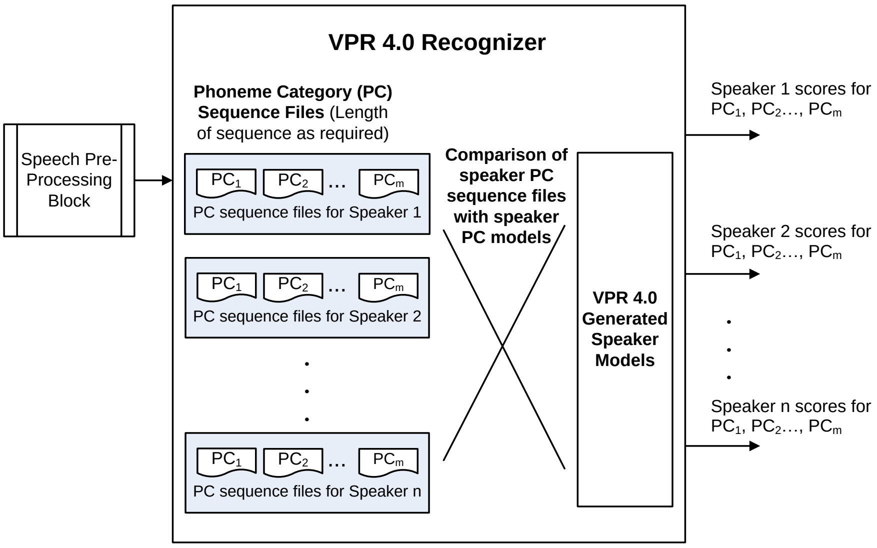

VPR 4.0 Recognizer
Speech Pre-Processing Block
Phoneme Category (PC) Sequence Files (Length of sequence as required)
PC1 PC2 ... PCm
PC sequence files for Speaker 1
PC1 PC2 ... PCm
PC sequence files for Speaker 2
PC1 PC2 ... PCm
PC sequence files for Speaker n
Comparison of speaker PC sequence files with speaker PC models
VPR 4.0 Generated Speaker Models
Speaker 1 scores for PC1, PC2..., PCm
Speaker 2 scores for PC1, PC2..., PCm
Speaker n scores for PC1, PC2..., PCm

Fig. 2. Block Diagram of Recognition and Scoring in SUSR

2.5. Designing Phoneme Categories

Each speech utterance, most likely a sentence, a phrase or an exclamation, comprises of words. Each word is formed of syllables and syllables are made up of vowels, consonants or both. Therefore, it is important to take into account these meaningful units of sounds to design meaningful categories for SUSR.

We propose the following process flow for designing phoneme categories. This process flow has been shown in Fig. 3:

1. Studying the types of sound.
2. Learning the types of vowels and their properties.
3. Studying the vowel structure of most common languages spoken in the world.
4. Designing the vowel categories which would cover most languages.
5. Studying the consonant types.
6. Designing consonant categories in the most generic form.
7. Designing syllable categories based upon the vowel and consonant categories.

3. Performance of Speaker recognition with continuous speech divided randomly

3.1. Importance of Meaningful Segmentation of Speech

When we are developing a speaker recognition system to make use of short utterances, speech segmentation is an important factor. It is important that each segment of sound should be meaningful. According to Adami et al. [22], segmentation of speech in an appropriate manner is needed when prosodic/high level features are being used for speaker recognition. During normal speech, there can be many clicks, hiss and random noise sounds, like ingressive sounds at an expression of regret or laughter or other similar expressions which carry little speaker information. Also, random picking up of segments might give an unknown mix of sounds like a combination of a palatal stop and low vowel, none of which, if chosen at random, would give a consistent result. A model developed from such random sounds would vary dramatically when segmentation of speech would be changed. To elaborate this fact, we show in Section 3.2 that a system based on utterances segmented in random, does not perform well when utterance lengths are as short as 3, 2 and 1 second.

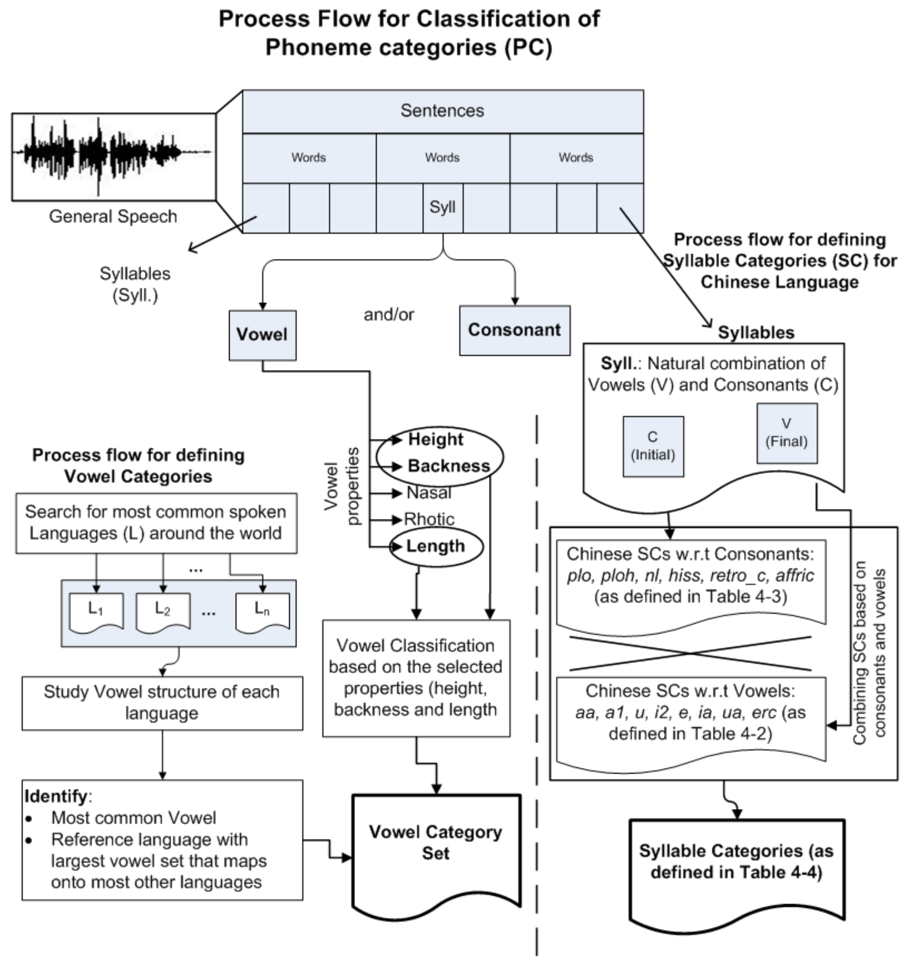

Process Flow for Classification of Phoneme categories (PC)
Sentences
Words
Words
Words
Syll
General Speech
Syllables (Syll.)
Vowel
and/or
Consonant
Process flow for defining Syllable Categories (SC) for Chinese Language
Syllables
Syll.: Natural combination of Vowels (V) and Consonants (C)
C (Initial)
V (Final)
Process flow for defining Vowel Categories
Vowel properties
Height
Backness
Nasal
Rhotic
Length
Search for most common spoken Languages (L) around the world
...
L1
L2
...
Ln
Study Vowel structure of each language
Identify:
Most common Vowel
Reference language with largest vowel set that maps onto most other languages
Vowel Classification based on the selected properties (height, backness and length
Vowel Category Set
Chinese SCs w.r.t Consonants: plo, ploh, nl, hiss, retro_c, affric (as defined in Table 4-3)
Chinese SCs w.r.t Vowels: aa, a1, u, i2, e, ia, ua, erc (as defined in Table 4-2)
Combining SCs based on consonants and vowels
Syllable Categories (as defined in Table 4-4)

Fig. 3. Process flow of Designing Phoneme Categories

3.2. Experimentation and Results

In order to test the system performance, we decided to first conduct speaker recognition experiments using our system without dividing speech into phoneme categories, so that a comparison could be made between categorical and continuous speech based speaker recognition systems.

We conducted an experiment with 2.5 minutes of training data for each speaker, which was used to train the GMM-UBM models for each speaker. Recognition was performed on full length of test utterances (9 minutes on average), and then by dividing the utterances into segments of 3 seconds, 2 seconds and 1 second in testing phase. The training of speaker models and testing was done with the system and database described in Section 2.

The results of this experiment are shown in Table 1.

Table 1. System Performance of SUSR with continuous Speech with Training length of 2.5 minutes

Utterance Length (Training-Testing)	EER%
2.5min-full (8min)	9.09
2.5-3 sec	19.46
2.5-2 sec	22.53
2.5-1 sec	29.56

It can be seen from Table 1 that although the system performs considerably well with long test utterance, there is a sharp decline in performance in terms of EER% when test length is equal to or less than 3 seconds. The high values of EERs show that the system performance remains unreliable and it can be improved by investigating further into the short utterance properties.

Hence the SUSR using continuous speech poses a concern for speaker recognition systems. This is why we propose categorical division (i.e. segmentation of sounds) of speech units for speaker recognition system.

This also shows that randomly dividing speech into segments does not give good results. Hence, investigating further into speech units when the utterance lengths are short, becomes important. This is why we propose categorical division of speech units for speaker recognition system as we describe in Section 4.

3.3. Importance of Phoneme based Short Utterance SR

Due to the several difficulties presented by using continuous speech and low performance of current SUSR systems , we decided to look deeper into using phonetic cues for speaker recognition. There are the following potential disadvantages of using continuous speech for speaker recognition with short utterances:

1. Using continuous speech, the performance of speaker recognition falls exponentially below 10 seconds of speech.
2. The random segments generated from continuous speech can result in unexpected and unintelligible mix of sounds when segment lengths become smaller.
3. There is not much continuous speech available in emergency situations where there might be a lot of bursts and breaks in speech.

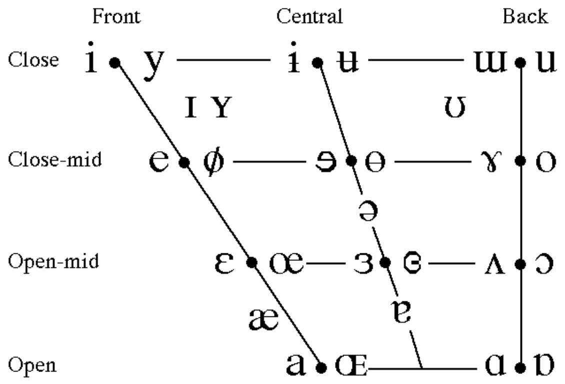
We believe phonemes are the correct, meaningful and practical unit which should be employed in speaker recognition using short utterances. We believe phonemes based Short utterance SR will have vast advantage over continuous speech for the following reasons:

1. When recognizing a speaker using short utterances of sound, it is important that the units of sounds should be meaningful. Phonemes are the smallest meaningful unit of sound.
2. For SUSR, the utterances of speech should be practical. The random segmentation of sounds do not allow intelligible and practical utterances.
3. It has been shown by study of phonetic speaker recognition that phonetic idiosyncrasies improve speaker recognition vastly [24].
4. Various studies have shown that high level characteristics in speech have vast advantage in speaker recognition [25].
5. High level characteristics can be mapped onto the low level features like formant frequencies [11]. Therefore, high level idiosyncrasies can be mapped onto the acoustic features. This is only possible with short utterances of speech.
6. Due to physiology of phoneme production, the phonemes contain distinct wave patterns which vary from speaker to speaker. This is because of the habit of speaking e.g. style of tongue rolling, or lisping etc.
7. Phoneme combinations (like syllables) can cover more speaker specific information because of speaker habits in uttering a specific combination of sounds.

4. Vowel Categories for PCBSUSR

As described in Section 3, continuous speech, if randomly divided into segments, does not perform well when utterance lengths are short. This is why we study smaller speech units by taking the whole phoneme/syllable into account, such that the speech segment is not random, but a meaningful unit of sound.

For an SUSR system that is language independent, we studied the phonology (systematic organization of sounds in languages) of phoneme production. The detailed study was made in order to figure out those phonemes which overlap in most commonly spoken languages of the world. The kind of sounds studied were mainly the egressive sounds (sounds produced by pushing the air out of the vocal tract), e.g. plosives, affricates, fricatives, continuants and vowels. Like in our previous study [18], we initially confine ourselves to the study of vowels and their properties to choose a universal vowel set for languages. Fig. 4 is the illustration of vowels according to International Phonetic Association [26].



Where symbols appear in pairs, the one to the right represents a rounded vowel.

Fig. 4. Vowels in languages

4.1. Vowel Characteristics

Height, Backness, Roundedness, Nasalization and R-coloration are the features that define a vowel [27]. Considering height and backness to be the main features, vowels were searched for in main languages of the world. For the purpose of our experimentation, we selected two languages i.e. English and Chinese, keeping them as baselines for the future system. The reason for choosing English is because it has a fairly large vowel inventory, which can be mapped onto most other common languages.

Diphthongs or gliding vowels are those, which glide from one quality to another during s production, e.g. “cow”. They are different from unilateral vowels, such that boundary where change takes place is not distinctly recognizable. English and Chinese, both contain many diphthongs. It is hard to ignore them altogether, since a huge amount of speech consists of vowels and a large portion of them are diphthongs. Diphthongs, however, pose a problem since they are vowels which might contain two very different vowel types. Most of the times, one vowel is dominant. Hence we propose to solve this problem by categorizing vowels into categories in a way that diphthongs are categorized on the basis of the first dominant vowel that appears in the diphthong.

4.2. Proposed Vowel Categories

Previously in [18], we defined eight categories of vowels. However we now define the following five categories of vowels:

1. Category a: low centre vowels, and those vowels which have dominant lower central vowels.
2. Category e: mid centre.

3. Category i: front high followed by neutral vowel (e.g. /a/), front high i followed by front vowel or none (e.g. /i/, /e/) and front high i followed by back vowel (e.g. /o/, /u/).
4. Category o: mid back rounded and those vowels which have dominant mid back vowels.
5. Category u: high back and (yu): high back unrounded.

The new categories are defined because of the following reasons:

1. To make a considerably uniform distribution of vowels across categories.
2. Some previous categories did not give very good results; it is assumed that along with the quality of segmentation, there was less number of phones in those categories.

The categories are defined based on dominant vowels in case of diphthongs. This distribution can be subject to scrutiny and subsequent change.

4.3. Phoneme Extraction

The speech data from “Annotated Speech Corpus of Chinese Discourse (ASCCD) from Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, China”, was used in this study. The speech data was segmented into phones based on the labels and placed into separate wave files. For the purpose of training and testing, the phones from each Vowel Category (VC) were concatenated into single wave file, representing one VC.

4.4. Universal Background Vowel Category Model (UBVCM) Training and Testing

Our study confines to vowel categories only, therefore, in our experiments, a Universal Background Vowel Category Model (UBVCM) [7] is built in a UBM fashion. A batch of vowels is obtained from the annotated speech data. The vowels are then used to estimate UBVCM based on the EM algorithm and the GMM modeling, giving a group of UBVCMs.

Vowel category (VC) models replace conventional speaker models. Vowel sequence and then VC sequences are obtained from training utterances of the speaker. Each VC model is estimated from UBVCM using the MAP algorithm with features of the vowels corresponding to the VC. During speaker recognition, the speech annotated test utterance is used to obtain VCs. Each VC is scored against its corresponding UBVCM, giving n scores. These n scores are fused to obtain the final score.

4.5. Results and Discussion

Table 2 represents the results of individual categories. Compared with the EER% of our previous work [18], which was of no less than 42%, this experiment has shown a huge improvement.

Table 2. Results Comparison Of UBVCM with [18]

EER %				
VCs	Complete length	3 sec	2 sec	1 sec
Category “a”	10.00	13.90	14.27	17.46
Category “e”	17.78	20.53	20.42	21.81
Category “i”	10.56	13.76	14.03	16.18
Category “o”	10.56	15.85	15.09	18.5
Category “u”	17.78	15.44	18.63	18.9
Average	13.336	15.90	16.49	18.57
Min	10.00	13.76	14.03	16.18
Max	17.78	20.53	20.42	21.81

The following measures have been different in this approach compared to the baseline system, yielding better results.

1. **Vowel Categories**: In the previous unilateral vowels [a, e, i, o, u, v] represented the main categories. Diphthongs were divided into the categories based on the first vowel. e.g. “iang” was put under category i1 (i followed by neutral vowel). However, in the present work, vowel categories have been merged into 5 broad level categories, such that the unilateral vowels come under the same name category except the unrounded high-back vowel “v”. “v” has been merged with “u”, the rounded high-back vowel category, hence giving the categories [a, e, i, o and u]. The diphthongs are categorized based on the dominant/open vowel in it, e.g. “iang” is placed in category “a” because /a/ is the dominant vowel in it. The categories of mid vowels e.g. “e” have not been changed in order to keep a fairly balanced distribution of phoneme across the vowel categories.
    a. **Segmentation method:** Segmentation of speech has been a big difference between our current and previous work. In the previous work [18], speech was segmented randomly before phoneme recognition and categorization. In the current work, however, annotated speech data was used on which phoneme level recognition had been performed earlier. The speech was subsequently segmented using the labels and each category was separated. Later, training and testing was performed on those vowel segments combined into one wave (.wav) file.

Following observations are made from the results in Table 2:

1. Data segmentation plays an important role. It shows that when performing SUSR, it is necessary to have a meaningful unit of sound like a complete phoneme. If speech is segmented randomly, as

described in Section 3, the individual information from each phoneme is wasted. Using a complete vowel, on the other hand, allows a speaker recognition system to make use of idiosyncrasies of a speaker in pronouncing that vowel.

2. If an entire phoneme is taken, it does not remain necessary to have accurate speech recognition. Instead broad phoneme categories can provide good results.
3. The vowel segments may be shorter than 500 ms, but being complete segments, they contain a large amount of speaker related information, pronunciation habits in particular.
4. Category “a” has given the best performance as shown in Tables 2. It is important to note that “a” is the most open vowel category in the set of VCs. Therefore, in articulating open vowels as those that fall in VC “a”, speaker specific voice attributes are represented in the speech un-constricted. This is the reason vowel /a/ and other similar vowels can be a preferred choice in speaker recognition based on phonemes.
    a. Categories “e” and “u” have higher EER as shown in Table 2.
    b. For category “e”, the reason could be the presence of phoneme “er” in category "e". The r-colored “e” or 儿 (in Chinese) has frequent occurrence in speech. The presence of the retroflex “r” brings about a constriction in the airway and consonant-like properties to the vowel, which is a variation from pure vowel. This causes a variation in speech in which one type of speech is different from other. This is why mismatch might have occurred and affected the results. This further strengthens our hypothesis that phonemes that fall under the same category provide higher chances of better results.
    c. Similarly, for category “u”, the reason of low performance could be the relative closeness of /u/. The narrowed airway due to backness of tongue and roundedness of mouth allow less amount of vocal tract information to enter the speech signal.
5. When taking phoneme categories all by themselves, pure vowels contain a larger amount of speaker information compared to diphthongs because their open and unrestricted properties allow vocal tract information to enter the speech signals.
6. Lastly but most importantly, it is worthy to note that the average training length of speech data used in our method was about 1 minute on average for each vowel category compared to 2.5 minutes of training length with continuous speech as described in Section 3. This shows that vowel categories perform better than continuous speech with smaller training durations hence reducing the computation cost and time of training process significantly.

The results from our method using VC sequences has shown an improvement of 37.73% of relative EER reduction compared to our baseline system using continuous speech and a relative EER reduction of 36.17%

compared to the minimum value of EER from the existing works (Section 1.2) using 2 seconds of speech.

5. Speech Unit Category based SUSR

In order to extend our research of phonemes from vowels to other units of speech, we initially gave a Syllable Category bases SUSR system in [19]. In Section 4, we presented improvement in vowel categories, which yielded vastly better results compared with [18]. In this section we present VCs, SCs as well as CCs for short utterance speaker recognition. Universal Background VC, CC and SC Models in Universal Background Model (UBM) fashion [7] as we did in our previous works [18, 28]. We use phonemic knowledge of phonemes to bring together similar speech units under single category. The purpose of this research is to determine which type of speech units provide better speaker recognition information.

5.1. Vowel, Consonant and Syllable Categories

Tables 3 and 4 respectively describe vowel and consonant categories (Based on Standard Chinese) used in current experiments.

Table 3. Vowel categories

VC	Representation	Vowels
Long /a/ type	aa	a, an, ang, ai, ao
Short /a/ type	a1	e, en, eng, i (zi)
/u/ type	u	u, ou, un, ong, iong, iou
/i/ and /v/ type	i2	i, v, in, ing, vn
/e/ type	e	ei, ie, a, ian, ve, van, uei, ven
/i/ diphthongs	ia	ia, iao, iang, iu,
/u/ diphthongs	ua	ua, uai, uan, uang, o, uen, ui, uo, yuan
retroflex	er	i (zhi), er

We also devise the following Syllable Categories by combining vowel and consonant categories. Since there is limitation to the valid combination of initials (consonants) and finals (vowels) [29], all vowel categories do not appear under each consonant category. liq category represents liquids, e.g. no-initial syllables and syllable starting in /l/.

plo_aa, plo_a1, plo_u, plo_i2, plo_e, plo_ua, ploh_aa, ploh_a1, ploh_u, ploh_i2, ploh_e, ploh_ua, nl_aa, nl_a1, nl_u, nl_i2, nl_e, hiss_aa, hiss_a1, hiss_u, hiss_ua, retro_aa, retro_a1, retro_u, retro_ua, affric_aa, affric_a1, *affric_u, affric_i2, affric_e, affric_ia, affric_ua, liq_aa, liq_a1, liq_u, liq_i2, liq_e, liq_ia, liq_ua, and retro_er*. The retro_er category incorporates all syllables ending in retroflex vowels.

Table 4. Consonant categories

CC	Representation	Consonants
Plosives	plo	b, d, g
Aspirated Plosives	ploh	p, t, k
Nasals and Laterals	nl	m, n, l
Hiss sounds	hiss	h, f
Retroflex	retro_c	zh, ch, sh, r
Affricates and Fricatives	affric	s, z, c, j, q, x

Knowledge based approach is used to devise the categories because rules of Chinese Language limit the total number of syllables and there are some syllables against which very scant data is obtained.

5.2. Experimental Results and Discussion

With the similar training and testing conditions as described in Section 2, we perform Speaker Recognition on each of the vowel, consonant and syllable categories. Each speech unit based UBM corresponds to its specific vowel, consonant or syllable category.

The results for syllable, vowel and consonant categories have been described in Tables 5 to Table 9.

Tables 5 and 6 present the results of syllable categories and Table 7 gives statistical summary of syllable category results.

Tables 8 and 9 respectively show results of vowel and consonant categories.

Tables 5, 6 and 7 show the syllable category results. Minimum values of EER% are 9.9%, 10% and 15.38% for 3 seconds, 2 seconds and 1 second respectively. In addition, 1.1% of EER has been obtained for Retro_er category for an average 18 seconds of test length.

The following observations have been made from these results:

1. It is observed from the results in Tables 5, 6 and 7 that longer test lengths are generally associated with better performance.
2. /a/ and /a/ based diphthongs, e.g. ia and ua based categories have shown better results proving that openness in vowels contributes to better performance.
3. The lower performance in case of /u/ based categories show that performance can deteriorate with closeness of vocal tract.
4. In some cases there is a sudden degradation of performance when sequence length becomes smaller (3 seconds or less). However, generally, length of test utterance, when smaller, does not deteriorate greatly.

5. In some instances, the EERs reduce when shorter utterances are used (e.g. SC retro_ua, SC affric_u). The reason could be that when smaller utterances are used, the effect of potentially unfavorable speech units is reduced.
6. Overall, long /a/ vowels (aa and ia), /e/ based categories, plosive based syllables and affricate/fricative categories in general have performed better.
7. Retroflex categories, especially the syllables ending in a retroflex (e.g. retro_er) show the promise of extremely good results. With training length ranging between 33 seconds to 44 seconds and test length of no longer than 22 seconds, the EER% was 1.11 with ID rate of 100%. This is because the articulation of retroflex sounds varies a lot from one person to another.

Table 5. Syllable category results in terms of EER% - Part 1

SC	Full Length		3 Sec	2 Sec	1 Sec
	Length (sec)	EER%			
plo_aa	0:09-0:12	10.00	17.37	16.56	17.82
plo_a1	0:08-0:11	18.33	25.00	24.17	26.85
plo_u	0:06-0:08	12.22	19.44	20.22	20.83
plo_i2	0:02-0:03	18.88	14.81	15.00	17.85
plo_e	0:03-0:08	21.11	20.55	26.08	29.29
plo_ua	0:10-0:13	10.00	15.78	17.27	19.39
ploh_aa	0:05-0:06	18.33	17.85	16.97	18.78
ploh_a1	0:03-0:04	21.67	26.11	29.39	32.80
ploh_u	0:03-0:03	21.11	34.25	30.00	38.23
ploh_i2	0:02-0:02	18.33	18.33	26.31	24.13
ploh_e	140ms-500ms	29.44	29.44	29.44	29.44
ploh_ua	0:01-0:02	20.00	20.00	27.77	25.92
nl_aa	0:03-0:04	22.22	34.31	25.00	29.62
nl_a1	0:03-0:03	28.33	30.40	30.00	32.43
nl_u	688ms-941ms	37.78	37.78	37.78	37.78
nl_i2	0:01-0:02	21.11	21.11	25.25	20.10
nl_e	0:07-0:09	9.44	11.80	12.76	16.40

The results show that speech unit categories perform very well in SUSR. Even with conventional GMM-UBM system, the results have shown that idiosyncrasies of speaker at phoneme level have its advantages in SUSR and they do not go wasted at short segments, rather each unit of speech; vowel, consonant or syllable, contains significant information about speaker’s identity.

Table 6. Syllable category results in terms of EER% - Part 2

SC	Full Length (sec)		3 sec	2 sec	1 sec
	Length	EER%			
hiss_aa	0:06-0:08	12.22	23.33	25.00	25.00
hiss_a1	0:08-0:09	27.78	28.25	27.03	27.71
hiss_u	0:04-0:06	27.78	27.27	22.77	33.53
hiss_ua	0:05-0:07	10.00	17.85	21.05	22.22
retro_aa	0:08-0:10	10.00	15.78	19.66	19.56
retro_a1	0:22-0:28	10.00	15.56	17.87	23.75
retro_u	0:11-0:13	10.00	21.49	20.87	21.66
retro_ua	0:03-0:04	18.88	28.89	26.26	26.57
affric_aa	0:09-0:10	8.33	17.83	16.88	18.42
affric_a1	0:05-0:06	11.11	21.75	26.76	28.74
affric_u	0:08-0:11	40.00	30.55	32.33	38.97
affric_i2	0:31-0:39	2.22	10.74	13.00	16.90
affric_e	0:12-0:15	10.00	12.50	14.39	15.38
affric_ia	0:15-0:19	18.33	17.56	15.73	18.25
affric_ua	0:03-0:04	11.11	14.91	10.00	21.05
liq_aa	0:06-0:09	12.22	15.79	17.77	21.69
liq_a1	0:06-0:08	20.00	26.66	26.42	32.85
liq_u	0:06-0:09	18.88	19.44	20.80	21.58
liq_i2	0:16-0:23	10.00	14.86	17.03	18.00
liq_e	0:08-0:11	9.44	13.73	13.62	17.88
liq_ia	0:04-0:05	18.33	20.00	20.68	20.40
liq_ua	985ms-0:01	20.55	20.55	20.55	22.83
retro_er	0:16-0:22	1.11	9.93	11.00	15.95

Table 7. Summary of Syllable category results in terms of EER%

	Full Length	3 sec	2 sec	1 sec
Average	16.91	20.99	21.69	24.16
Minimum	1.11	9.93	10.00	15.38
Maximum	40.00	37.78	37.78	38.97

It can be seen from Tables 8 that Vowels present an extremely feasible option for SUSR.

Minimum values of EER% for 3 seconds, 2 seconds and 1 second of test length has been 13.72, 12.93 and 16.09 respectively. The Category ia has been seen to be consistent with the best results.

In conformity with results in Section 4.5, vowel categories have shown good results. This proves that with careful selection of phoneme categories, performance of speaker recognition improves significantly when phoneme category sequences are used for speaker recognition with short utterances, even when they are as little as 1 second.

Table 8. Vowel category results in terms of EER%

VC		Complete Length (sec)	3 sec	2 sec	1 sec	
			Utterance length		EER%	
aa	0:35	0:43	10.00	14.72	15.36	17.40
a1	0:34	0:47	18.33	21.00	21.36	22.58
U	0:31	0:42	10.55	17.16	18.47	20.29
i2	0:39	0:52	10.56	14.74	14.97	16.09
E	0:30	0:38	12.22	16.41	15.85	18.73
Ia	0:11	0:15	11.11	13.72	12.93	17.28
ua	0:20	0:28	12.22	18.84	18.98	21.39
er	0:09	0:14	18.88	20.51	19.54	21.30
Average			12.98	17.14	17.18	19.38
Min			10.00	13.72	12.93	16.09
Max			18.88	21.00	21.36	22.58

Table 9. Consonant category results in terms of EER%

CC		Complete Length (sec)	3 sec	2 sec	1 sec	
	Utterance length		EER%			
plo	0:04	0:09	40.55	45.83	42.10	45.58
ploh	0:05	0:08	30.00	30.65	31.70	33.33
nl	0:07	0:11	30.00	29.72	28.30	30.80
hiss	0:08	0:14	30.00	30.23	38.00	36.20
retro_c	0:22	0:31	30.00	35.41	34.79	36.87
affric	0:33	0:43	30.00	29.99	32.98	34.40
plo	0:04	0:09	40.55	45.83	42.10	45.58
ploh	0:05	0:08	30.00	30.65	31.70	33.33
Average			31.76	33.64	34.65	36.20
Min			30.00	29.72	28.30	30.80
Max			40.55	45.83	42.10	45.58

Table 9 shows the results of speaker recognition performance with consonant categories. With minimum EER% of 29.72, 28.3% and 30.8% for 3 seconds, 2 seconds and 1 second, consonant categories have not shown promising results. The minimum values have been obtained with the nasal-lateral category. The reasons behind poor performance of SUSR with consonant categories are:

1. Consonants generally have very little duration. The duration does not allow speaker variations in speech signals to be captured to their full extents.
2. Many consonants are produced without vibration of vocal cords i.e. unvoiced consonants. Due to this factor, the properties of sound are reduced to white noise like hiss, bursts and silences. Such sounds do not have any significant speaker related information in them.
3. Consonants, when on their own, do not have active signals and idiosyncratic information in them.
4. The constriction in vocal tract, when uncoupled with vowels, contributes to the lower performance of SUSR.
5. The minimum values of EER% have been obtained from nasal-lateral consonant category. This is the consonant category with most voicing activity as well as vowel like properties of laterals allow some speaker variation in speech signals.

It is because of the above factors that we do not recommend consonant sequence based speaker recognition i.e. when consonants are used on their own without their combination with vowels. However when coexisting with vowels, they impact the properties of vowels (especially at the transition boundary). This fact can be proven by the good results shown by syllable categories in Table 5, 6 and 7. Therefore, although the importance of consonants in speech is undeniable and their impact on speaker recognition can be considerable, they are not a feasible choice when used alone.

Using Syllable Categories for our SUSR system there has been a relative EER reduction of 55.61% and 54.50% compared to our baseline system with continuous speech and minimum EER recorded in literature (Section 1.2) respectively for 2 seconds of test utterance length.

6. Conclusions and Future Work

Our research discusses how phone unit categories can be used in SUSR. We used conventional GMM-UBM models for different category types. We show from an experiment on randomly divided continuous speech that a complete speech unit is very important when we are performing speaker recognition on units as small as 1 second. Our results on the experiments based on speech units like vowels, and syllables show that text dependent speaker recognition based on speech units gives significantly better results in SUSR. Although consonants are extremely important in speech, when they are used alone, they do not provide good results. However in combination with vowels i.e. syllables, they provide quite good results. We conclude from our experiments that when short utterances are used, it is better to use phonemes rather than continuous speech. Phonemes can use knowledge of utterances, while using categories allowing the system to use rudimentary speech recognition, not requiring high accuracy. Our results show that different speech units have different performances. So, in performing real time speaker recognition, priority can be given to specific speech units while poor performing speech units can be avoided to ensure recognition.

As part of future work, we propose the following potential research directions:

1. Applying data driven statistical method for determining the phone units as an alternative of current knowledge based approach.
2. Study of formant dynamics at speech units and their role in SUSR.
3. Exploring speaker recognition in complete absence of a specific speech unit category due to fewer amounts of training data.
4. Cross speech unit category experimentation in order to understand which categories perform well against each other.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant No. 61271389 and the National Basic Research Program (973 Program) of China under Grant No. 2013CB329302.

References

1. Vogt, R., Baker, B.,Sridharan, S.: Factor analysis subspace estimation for speaker verification with short utterances. In Proceedings of INTERSPEECH 2008. 853-856. (2008).
2. Chan, W. N., Zheng, N.,Lee, T.: Discrimination Power of Vocal Source and Vocal Tract Related Features for Speaker Segmentation. IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSIONG, Vol. 15, No. 6, 1884-1892. (2007).
3. Jayanna, H. S.,Prasanna, S. R. M.: Multiple Frame Size and Rate Analysis for Speaker Recognition under Limited Data Condition. IET SIGNAL PROCESSING, Vol. 3, No. 3, 189-204. (2009).
4. Kenny, P., Boulianne, G.,Dummouchel, P.: Eigenvoice Modeling with Sparse Training Data. IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, Vol. 13, No. 3, 345-354. (2005).
5. Wu-fu, Lu, X. Y., Bei-qian, D.,Hui, L.: Speaker Recognition based on the phonal speech Short CGMM-UBM. Journal of Circuits and Systems, Vol. 12, No. 5, 131-136. (2007).
6. Kanagasundaram, A., Vogt, R., Dean, D. B., Sridharan, S.,Mason, M. W.: i-vector based speaker recognition on short utterances. in Proceedings of the 12th Annual Conference of the International Speech Communication Association, (2011).
7. Fauve, B., Evans, N., Pearson, N., Bonastre, J.-F.,Mason, J.: Influence of task duration in text-independent speaker verification. In Proceedings of Interspeech, 2007. 794- 797. (2007).
8. Stadelmann, T.,Freisleben, B.: Dimension-Decoupled Gaussian Mixture Model for Short Utterance Speaker Recognition. in 20th International Conference on Pattern Recognition, (2010).

9. Larcher, A., Bonastre, J.-F.,Mason, J. S. D.: Short Utterance-based Video Aided Speaker Recognition. in International Workshop on Multimedia Signal Processing, Cairns, Australia, (2008).
10. Fattah, M. A.: Phoneme Based Speaker Modeling to Improve Speaker Recognition. Information, Vol. 9, No. 1, 135-147. (2006).
11. Fatima, N., Aftab, S., Sultan, R., Shah, S. A. H., Hashmi, B. M., Majid, A.*, et al.*: Speaker Recognition Using Lower Formants. In Proceedings of Proceedings of IEEE INMIC 2004. Lahore, 125-130. (2004).
12. (NIST), N. I. o. S. a. t.: (2012). [Online]. Available: http://www.nist.gov/index.html (current February 2012).
13. Li-fu, W., Yan-lu, X., Bei-qian, D.,Hui, L.: CGMM-UBM based speaker Verification using short telephone speech. Journal of Circuits and Systems, Vol. 12, No. 5, 131-136. (2007).
14. Lin, L., Shu-xun, W.,Gang, G.: Novel Speaker Recognition Method Based on Little Speech Data. Journal of System Simulation, Vol. 19, No. 10, 2272-2275. (2007).
15. McLaren, M., Vogt, R., Baker, B.,Sridhara, S.: Experiments in SVM-based Speaker Verification Using Short Utterances. In Proceedings of Odyssey: The Language and Speaker Workshop. (2010).
16. Vogt, R., Lustri, C.,Sridharan, S.: Factor Analysis Modelling for Speaker Verification with Short Utterances. In Proceedings of In Proceedings Odyssey 2008: The Speaker and Language Recognition Workshop. South Africa. (2008).
17. Xiao-han, L., Nan-chen, H., Bei-qian, D.,Zhi-qiang, Y.: Research on the HMM-UBM and Short Text Based Speaker Verification. Information and Control, Vol. 33, No. 6, 762-764. (2004).
18. Fatima, N., Wu, X., Thomas Fang Zheng, Chenhao Zhang,Wang, G.: A Universal Phoneme-Set Based Language Independent Short Utterance Speaker Recognition. in 11th National Conference on Man-Machine Speech Communication (NCMMSC '11), Xi'an, China, (2011).
19. Fatima, N.,Zheng, T. F.: Syllable Category Based Short Utterance Speaker Recognition. in International Conference on Audio, Language and Image Processing, Shanghai, China, (2012).
20. Reynolds, D. A.: Channel robust speaker verification via feature mapping. In Proceedings of ICASSP. 53-56. . (2003)
21. Aijun, L., Maocan, L., Xiaxia, C., Yiqing, Z., Guohua, S., Wu, H.*, et al.*: Speech corpus of Chinese discourse and the phonetic research. In Proceedings of International conference of speech and language processing (ICSLP). 13-18. (2000)
22. Adami, A. G.,Hermansky, H.: Segmentation of speech for speaker and language recognition. In Proceedings of 8th European Conference on Speech and Communication and Technology (Eurospeech). (2003)
23. Hagiwara, R.: Monthly mystery spectrogram webzone (2009). [Online]. Available: http://home.cc.umanitoba.ca/~robh/archives/arc0601.html (current March 4 2012)
24. Kohler, M. A., Andrews, W. D., Campbell, J. P.,Hernandez-Cordero, J.: Phonetic Refraction for Speaker Recognition. In Proceedings of Workshop on Multilingual Speech and Language Processing. Aalborg, Denmark. (2001)
25. Kinnunen, T.,Li, H.: An Overview of Text-Independent Speaker Recognition: from Features to Supervectors. Speech Communication, Vol. 52, No. 1, 12-40. (2010).
26. IPA: Vowel Types in languages (2005). [Online]. Available: http://www.langsci.ucl.ac.uk/ipa/vowels.html (current March 4 2012)
27. Wikipedia: Vowel (2012). [Online]. Available: http://en.wikipedia.org/wiki/Vowels (current march 4 2012)

28. Fatima, N.,Zheng, T. F.: Vowel Category Based Short Utterance Speaker Recognition. in International Conference on Systems and Informatics, YanTai, China, (2012).
29. Jiyong, Z., Fang, Z., Mingxing, X.,Shuqing, L.: Intra-syllable dependent phonetic modeling for Chinese speech recognition. In Proceedings of International Symposium on Chinese Spoken Language Processing (ISCSLP'00). Beijing, China, 73-76. (2000).

Nakhat Fatima received the Bachelor's and Master's degree in Computer Science from National University of Computer and Emerging Sciences (NUCES-FAST), Lahore, Pakistan in 2004 and 2007 respectively. She worked as a Software Development Engineer in Centre for Research in Urdu Language Processing (CRULP). She passed her defense for Ph.D. in Computer Science in December 2012 from Tsinghua University, Beijing, China. Her interests include Speaker Recognition and Text to Speech processing.

Xiaojun Wu received the Ph.D. degree in computer science and technologies from Tsinghua University, Beijing, China, in 2004. He is currently an associate professor in the Research Institute of Information Technology, Tsinghua University. His research interests include speaker recognition, natural language understanding and dialogue management.

Thomas Fang Zheng received the Ph.D. degree in computer science and technology from Tsinghua University, Beijing, China, in 1997. He is now a Research Professor and Director of the Center for Speech and Language Technologies, Tsinghua University. His research focuses on speech and language processing. He has published more than 210 papers. He plays active roles in a number of communities, including Vice President of the Asia-Pacific Signal and Information Processing Association (APSIPA), IEEE Senior Member, Chair of the Chinese Corpus Consortium, Chair of the Standing Committee of China's National Conference on Man-Machine Speech Communication, Council Member of Chinese Information Processing Society of China, and Council Member of the Acoustical Society of China. He is an Associate Editor of the IEEE Transactions on Audio, Speech and Language Processing and APSIPA Transaction on Signal and Information Processing, he is on the editorial boards of Speech Communication, Journal of Signal and Information Processing, and the Journal of Chinese Information Processing. (Please refer to http://cslt.riit.tsinghua.edu.cn/~fzheng for details).

Received: February 08, 2012; Accepted: November 27, 2012.