Short Utterance Speaker Recognition

A research Agenda

Nakhat Fatima and Thomas Fang Zheng
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
Email: fatima@cslt.riit.tsinghua.edu.cn, fzheng@tsinghua.edu.cn

***Abstract*— Short Utterance Speaker Recognition (SUSR) is an important area of speaker recognition when only small amount of speech data is available for testing and training. We list the most commonly used state-of-the-art methods of speaker recognition and the significance of prosodic speaker recognition. A short survey of SUSR is hereby conducted, highlighting various methodologies when using short utterances to recognize speakers. We also specify future research directions in the field SUSR which, together with modern technologies and the ongoing research in prosodic speaker recognition, can lead to better results in speaker recognition.**

Keywords; Short Utterance Speaker Recognition, Prosodic Speaker Recognition, Phoneme Categories

I. INTRODUCTION

A. Background

Research in the area of Speaker Recognition began decades ago. There has been research in this area ranging from theoretical to applied linguistics and signal processing. Voice is both physiological as well as behavioral biometric in nature. Speech carries physiological aspects of a speaker because it is affected by the unique shape and size of his/her vocal tract, mouth, nasal cavity etc. It is a behavioral aspect because accent and involuntary changes in acoustics when. one shifts from one phone to the other, prosody etc. are learned behaviors. Voice is the most natural way of communication, hence making it natural choice in recognizing a person [1].

With the advancement in technology, various methods started to develop for Automatic Speaker Recognition. Speaker recognition has its main advantages in security systems and forensics. There are two applications of Speaker Recognition i.e. Speaker verification and Speaker Identification. Speaker Verification is to confirm the claim of identity and declaring the person to be true or impostor. It can be used in security systems. Speaker Identification means recognizing a speaker from a pool of speakers. This area has its main application in forensics and investigation where a given voice sample can be used to determine the identity of a person. Similarly, speaker recognition can be text dependent or text independent. In text dependent systems, speaker recognition is performed coupled with speech recognition. In text independent systems, a speaker is recognized independent of the speech or language.

Speaker recognition generally consists of several steps [2]:

- Parameter/feature extraction
- Statistical Modeling; Training scheme is applied at this point
- Normalization and Score Computation

Speaker recognition can make use of acoustic as well as prosodic aspects of speech. This paper surveys the contemporary approaches to speaker recognition in order to have a brief look at the two aspects. In doing so we highlight how these two aspects can be combined to address the research problems related to Short Utterance Speaker Recognition (SUSR), which is now becoming a major consideration of modern speaker recognition research. Most of the speaker recognition methods require a large amount of speech data for training of models. SUSR becomes important because of the difficulty in acquiring large amount of appropriate speech. Most of the times background noise gets into the way; other times a faulty recording reduces most of the voice to glitches and chirps, leaving behind only a few seconds of intelligible speech; sometimes voice overlaps become hard to manage. For such problems, it becomes necessary to take into account the speaker specific information in short utterances of speech so that speaker recognition can be performed even when there is only a little amount of data available. SUSR can be addressed by using both acoustic as well as the high level information in speech, which includes pronunciation idiosyncrasies, prosody, phoneme dynamics etc.

The rest of the paper is organized as follows. Section II and III, respectively overview the state-of-the-art research in Acoustic and Prosodic Speaker Recognition. In Section IV, we present current research in the field of SUSR, followed by discussion on SUSR research directions in Section V. We draw our conclusions in Section VI.

II. ACOUSTIC SPEAKER RECOGNITION TECHNIQUES

There are various techniques involved in speaker recognition. In order to make speaker recognition as close to human perception as possible, methods such as filter banks were introduced to simulate human perception [2]. Many technologies have been used, including Dynamic Time Warping (DTW) [3], Vector Quantization (VQ) [4], Artificial Neural Networks (ANN) and Hidden Markov Models (HMM), Linear Predictive Coding (LPC) [2], [5], to name a few, for the purpose of speaker recognition.

 978-1-4673-0199-2/12/$31.00 ©2012 IEEE

As technologies progressed and statistical methods started to evolve, modeling and adaptation of acoustic models advanced. Complex calculations like norms, Gaussian Mixture Models (GMM), Universal Background Model (UBM) and Factor Analysis (FA) have emerged in recent years [2], [5]. Most of the successful Speaker Recognition systems employ GMM training, UBM and simplified FA. Silence detection and multi-speaker segmentation are added in these systems to enhance speaker recognition efficiency. Techniques such as noise removal are also used in order to make these systems robust to channel and noise. Support Vector Machine (SVM) based systems have proven to provide even better results when combined with GMM-UBM based systems [2], [5]. Since in real time systems, speaker recognition products can have any amount of data, unknown to the systems, there is a need to incorporate lengthy, complicated calculations like Joint Factor Analysis (JFA) to handle the situations of having large or small amounts of speech data [6]. Hence, today a speaker recognition system is a combination of many functions, each with high statistical complexities.

III. Prosodic Speaker Recognition

Pronunciation and prosody are very elemental factors for speaker recognition using high level cues [7]. Prosodic features include F0, duration (e.g. pause statistics, phone duration), speaking rate, energy distribution/modulations, formant trajectories, phone sequences etc. [5].

Relatively new and innovative Phonetic Speaker Recognition saw its pioneers in Kohler, Andrews and Campbell in 2001 [8]. According to [8] traditional systems have several drawbacks. The channel effects can dramatically change the acoustic properties of a particular individual, e.g. acoustically varying landline and cell phone channels. The traditional systems rely upon methods quite different from human recognition. Humans use intonation, prosody, word choice, pronunciation, accent, and other speech habits when recognizing speakers. It is because of using these properties of a speaker, which are called the higher level features, human listeners are not significantly affected by variations in channels. Automatic speaker recognition using high level features attempt to simulate human perception in speaker recognition. How phone sequences can be used to exploit differences in pronunciation in two individuals for speaker recognition is described in [7]. In order to perform this, phone sequences are generated using phone recognizers and then recognition is performed on those sequences. Phonetic Language Modeling (PLM) [9] is the continuation to this idea. For each phone sequence, the PLMs are created with the help of phoneme recognizers. The whole collection of phone-models can then be used in the recognition task.

Selective use of phones, i.e. keeping the phone sequences occurring frequently and ignoring those that occur sparsely has shown to improve the results in N-gram computation of phone sequences [10]. Gender and language dependent system outperform gender fused and language fused systems [10].

Since GMM does not utilize phonetic information, the training set for GMM simply contains all the spectral features of different phonetic classes pooled together. By using separate GMM for each phone or syllable class, this problem is addressed in several researches [5].

SVM-based phonetic speaker recognition has been described in [8] which halved the error rates compared to the previous speaker recognition. In [11] the method of phonetic speaker recognition has been shown to overcome forgery by voice transformation.

Among other features in the prosodic domain, F0 has been found to give the best accuracy so far since it conveys both physiological as well as learned characteristics [12]. Also, F0 contour calculations are used for speaker recognition to determine energy slopes [12]. Another research explores N-gram feature approach for prosody measurement in speaker recognition [13]. AFCC-HMM framework with adaptive frequency scale has been used in [14] for pronunciation evaluation in speaker recognition. Similarly broad categories of phones for speaker recognition were explored in [15]. According to their findings, the highest identification rate was found to be given by vowels followed by nasals, fricatives, semi vowels and then stops.

Working on phoneme durations, Charl has described in his thesis [1] that phoneme duration can be used as a high level feature for speaker recognition. The research has been carried out at word and sentence level. Modeling was done using HMM. A study of formant contours at consonant-vowel boundary shows that a speaker can be identified using the consonant to vowel transition information. The speaker information at the transition was named the “speaker-style” [16]. Mohamed Abdel Fattah et al have explored the importance of phonemes in speaker recognition [17]. They used speaker dependent models by segmenting each utterance into phoneme segments, and creating an HMM speaker model with it. Later using the phonemes for recognition they determined which phonemes are most suitable for recognition. They found that vowels and nasals have a good result in speaker recognition whereas fricatives and stops have little use in speaker recognition purpose. Out of the many methods employed for speaker recognition using phonetic cues, use of phoneme dynamics provides a novel yet successful way of recognizing speaker. According to the research conducted by DyVis project – Cambridge University, formant dynamics are a successful way of recognizing a speaker which covers different speaking styles as well as background conditions. Regression can be used to classify speakers and then recognition based on formant dynamics can be used [18].

IV. Short Utterance Speaker Recogntion

At present, a number of researches are going on to determine methods to identify a speaker when either the speech given is too small or to use less amount of speech to cut computation costs.

In their investigation to unify languages, [19] have looked for a way to develop universal phone models. They have used speech attribute detection to achieve this goal e.g. place of articulation mapped with phone etc. Similarly broad categories of phones for speaker recognition were explored through [15]. According to their findings, the highest identification rate was found to be given by vowels followed by nasals, fricatives,

semi vowels and then stops. They have shown that the phonetic content of speech is more important than the quantity of speech.

According to recent research [20], background atmosphere or noise etc also contribute to errors in speaker recognition. The influence of background increases when the test segment of speech is very small. Due to this reason, feature vectors have to be carefully selected. Selective use of feature vectors enables the system to select only those vectors which would later be useful for recognition purpose, rejecting the ones too much influenced by the overlap of background noise of silence.

Speaker recognition requires a large amount of speech data, making use of huge files and complicated processing. This has hampered the speaker recognition technology to be used widely. Research has thus lead to JFA, SVM and I-vector based technologies [21]. As the utterances get shorter, results deteriorate. However factor analysis has shown better results for segments shorter than 10 sec. Factor analysis approaches to speaker verification were originally intended to model the intersession variability directly in the construction of the super-vectors used for scoring verification trials, such as in the standard JFA approach [21]. I-vector is a form of front end analysis that represents the GMM super-vector by a single total-variability space. The EER of speech more than 10 seconds is quite low, although it increases when speech gets smaller in duration. Also, in [22], [23], factor analysis is performed for short utterance speaker recognition.

In [24] dimension decoupled GMM's are used. This system achieved more than 80% recognition accuracy with less than 5.5 seconds of training- and 1.3 seconds of evaluation data.

Training and testing with 10 seconds of speech on variations of GMM and SVM have been presented in [25]. A short utterance speaker recognition system based on GMM-UBM along with HMM for linguistic modeling makes use of video aid [26]. The utterances go as short as 3 seconds. In [27], 15-45 seconds of speech utterances have been used to determine a speaker's identity.

V. SUSR - Research Agenda

SUSR is an emerging field. It is open to a lot of research by combining the modern technologies and prosodic speaker recognition aspects.

In our previous work we proposed a text dependent speaker recognition system making use of short utterances to recognize a speaker [28]. We devised an innovative design of speaker recognition, which was based on phoneme-classes for speaker recognition. For this purpose, a language independent vowel-category set was defined. The vowel class set was defined using linguistic knowledge of vowels. Consonants were not used for the study because there are varying speaker-related information in consonants, thereby hampering the overall recognition process. Those vowels categories were defined that would help cover maximum number of the vowels in most of the languages. In training phase a group of vowel-category models was built with respect to each speaker making a speaker vowel model. In test phase the test utterance was first recognized into a sequence of phonemes and then text-dependent 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.

Data segmentation plays an important role in SUSR. When performing SUSR, it is necessary to have an entire phoneme. If speech is segmented randomly, the individual information in each phoneme is wasted. However, if an entire phoneme is taken, it does not remain necessary to have accurate speech recognition. Instead broad phoneme categories can provide with desired results. After addressing the problem of speech segmentation, we defined new vowel categories (VC), merging similar vowels in one category and redefining diphthongs using the knowledge of openness and closeness of vowels. The results varied vastly after that, reducing EER of the base line system from 42% to 13.76% for one of the vowel categories. We shall be presenting the details of the VC based approach for SUSR in a separate manuscript.

Based on our survey in the domain of Speaker recognition, with particular focus on SUSR, we believe there are a number of research areas still open when phoneme-classes are used in SUSR, and hence propose the following research directions in this field:

A. Cross category comparison

In phoneme category SUSR when using VCs, we suggest a comparative study of vowel categories. Each VC should be tested against all others to examine their consistency and to determine which other VCs can be brought together. This study would determine which categories can be used in the absence of sufficient data in one of the categories. This would also help to determine if vowels should indeed be put into separate categories or all vowels should be placed together. If results of cross-categories do not degrade much, then all vowels can be put in one category. If, however, the difference in performance is huge, it will show that length of vowel has different importance.

B. Comparison of long and short vowel categories

Another potential research lies in investigating whether short VCs can also present desired results and to compare them with long VCs. This can also help to determine if short vowels can be used against long vowel models in the absence of long vowel data. Also, it will show if short vowels have the same level of speaker idiosyncratic information as that of long vowels. It will further determine if similar type of long and short vowels can be put together in one category or they should be kept apart.

C. Detailed analysis of nasal vowels

Nasal tract information in nasal vowels provides added knowledge about the speaker. The shape and size of nasal tract as well as a person's habits can change the audio signals. A detailed analysis of nasal vowels can be made in order to determine their effectiveness in SUSR. Similarly they can be tested against non-nasal vowels to examine variations. It can further be explored if all nasal vowels have similar results or they perform differently and have to be placed in varying categories based on their base-vowel category.

D. Consonant categories for SUSR

Consonants are important to each language. Although vowels constitute a larger portion of speech, a speech cannot be complete without consonants. There are various types of consonants depending upon the type of constriction they form in the vocal tract. A study of consonants in SUSR can be beneficial in determining which consonants/consonant categories can perform better for SUSR and to merge those consonants, which score better when cross tested. It has been shown in various studies that vowels generally perform better than consonants in phoneme based speaker recognition [17]. This study would show if likewise is true when short utterances are used for speaker recognition.

E. The use of biphone categories

In order to determine the effect of consonants and vowels put together, biphone categories can be defined and then cross tested. As shown in [16], this information might provide further speaker specific knowledge and that if consonant-vowel transition information is indeed valuable for speaker recognition with short utterances. In order to test the biphone categories, diphones, syllables and word levels can be tested. Because syllables are the most natural segmentation of speech, it is most suitable to use syllables. Syllable categories can be defined by multiplying vowel categories with consonant categories. The categories can be further refined by taking only legitimate consonant-vowel combinations according to a given language.

This study is potentially important. This would show if biphones/triphones or n-grams retain speaker information at short utterances. It would also show which type of consonant-vowel combinations provide with best performance. This study can show that speaker style when speaking a biphone and his/her habits and peculiarities play an important role in speaker recognition. Also, such idiosyncrasies are retained even when speech units are small.

F. Biphone based formant dynamics for SUSR

The use of formant trajectories and dynamics in co-articulation of sounds is another challenging area in SUSR. This study can provide with information about formant dynamics at phone transition point. This study might also help to determine new features for speaker recognition and potentially reduce computational costs.

Along with the aforementioned topics, SUSR can have many other facets. Being relatively new and still under research, SUSR with phonemes and other speech-units can be a revolutionary study in speaker recognition. In combination with modern methods, phoneme based study can help determine the minute aspects of speech that help human listeners identify a person in everyday interaction. The results of this type of study can both be unexpected as well as enlightening. This would help determine the phonetics of speaker recognition and the importance of such knowledge in SUSR.

VI. CONCLUSIONS

Short Utterance Speaker Recognition is a challenging area of speaker recognition. It is emerging rapidly in the contemporary technologies. We have overviewed a short survey of state-of-the-art speaker recognition and prosodic methods to identify a speaker. The contemporary ventures in SUSR have also been listed. SUSR is important when only small amount of speech data is available. In light of the current state of research in SUSR, we also propose future research directions, which together with modern technology and prosodic information in speech can yield better results when using short utterances of speech.

REFERENCES

[1] C.J.V. Heerden, "Phoneme duration modelling for speaker verfication", Facutly of Engineering, the Built Environment and Information Technology, University of Pretoria, Pretoria, 2008.

[2] F. Bimbot, "A Tutorial on Text-Independent Speaker Verification", EURASIP Journal on Applied Signal. Processing vol. 4, no., pp. 430-451.

[3] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition", IEEE Trans. Acoustics,Speech and Signal Processing, ASSP vol. 26, no. 1, Feb. 1978, pp. 43-49.

[4] A. Gersho and R. Gray, Vector Quantization and Signal Compression, Kluwer Academic Publishers, Boston, 1992

[5] T. Kinnunen and H. Li, "An Overview of Text-Independent Speaker Recognition: from Features to Supervectors", Speech Communication vol. 52, no. 1, pp. 12-40.

[6] P. Kenny, P. Ouellet, N. Dehak, V. Gupta and P. Dumouchel, "A Study of Inter-Speaker Variability in Speaker Verification", IEEE TRANS. AUDIO SPEECH AND LANGUAGE PROCESSING vol. 16, no. 5, pp. 980-988.

[7] M.A. Kohler, W.D. Andrews, J.P. Campbell and J. Hernandez-Cordero, "Phonetic Refraction for Speaker Recognition", in Workshop on Multilingual Speech and Language Processing, Aalborg, Denmark, 2001.

[8] W.M. Campbell, J.P. Campbell, D.A. Reynolds, D.A. Jones and T.R. Leek, "Phonetic Speaker Recognition with Support Vector Machines", in Advances in Neural Information Processing Systems (NIPS), 2003.

[9] Q. Jin, Tanja Schultz and Alex Waibel, "Phonetic Speaker Identification", in In proceedings of the International Conference of Spoken Language Processing (ICSLP-2002), Denver, CO, September 2002.

[10] W. Andrews, M. Kohler and J. Campbell, "Phonetic Speaker Recognition", in Proceedings of Eurospeech 2001, Aalborg, Denmark, September 3-7, 2001.

[11] Q. Jin, A.R. Toth, A.W. Black and T. Schultz, "Is voice transformation a threat to speaker identification?", in Acoustics, Speech and Signal Processing, ICASSP 2008, 2008., April 4, pp. 4845 –4848.

[12] A. Adami, R. Mihaescu, D.A. Reynolds and J. Godfrey, "Modeling Prosodic Dynamics for Speaker Recognition", in ICASSP, Hong Kong, China 2003.

[13] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman and A. Stolcke, "Modeling Prosodic Feature Sequences for Speaker Recognition", Speech Communication vol. 46, no. 3-4, pp. 455-472.

[14] Z. Ge, S.R. Sharma and M.J.T. Smith, "Adaptive frequency scale HMM method for pronunciation evaluation", in IEEE ICASSP 2011, 2011.

[15] M. Antal and G. Toderean, "Broad Phonetic Classes Expressing Speaker", Individuality, Studia Informatica vol. 51, no. 1, pp. 49-58.

[16] N. Fatima, S. Aftab, R. S6ultan, S.A.H. Shah, B.M. Hashmi, A. Majid, et al., "Speaker Recognition Using Lower Formants", in Proceedings of IEEE INMIC 2004, Lahore, 2004, pp. 125-130.

[17] M.A. Fattah, "Phoneme Based Speaker Modeling to Improve Speaker Recognition", Information vol. 9, no. 1, pp. 135-147.

[18] K. McDougall and F. Nolan, "Discrimination of Speakers Using the Formant Dynamics of /u:/ in British English", in 16th International Congress of Phonetic Sciences, Saarbrücken, 2007, 6-10 August 2007, pp. 1825-1828.

[19] C.-H. Lee, "Attribute-Based Universal Phone Modeling for Multilingual Automatic Speech Recognition (MASR)", in Oriental COCOSDA, Kyoto, Japan 2008, Nov. 25, 2008.

[20] S. Kwon and S. Narayanan, "Robust speaker identification based on selective use of feature vectors", Pattern Recogn. Lett. vol. 28, no. 1, pp. 85-89.

[21] A. Kanagasundaram, R. Vogt, D.B. Dean, S. Sridharan and M.W. Mason, "i-vector based speaker recognition on short utterances", in Proceedings

of the 12th Annual Conference of the International Speech Communication Association, 2011.

[22] R. Vogt, B. Baker and S. Sridharan, "Factor analysis subspace estimation for speaker verification with short utterances", in INTERSPEECH 2008, 2008, pp. 853-856.

[23] W.N. Chan, N. Zheng and T. Lee, "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), August 2007, pp. 1884-1892.

[24] T. Stadelmann and B. Freisleben, "Dimension-Decoupled Gaussian Mixture Model for Short Utterance Speaker Recognition", in 20th International Conference on Pattern Recognition, 2010.

[25] B. Fauve, N. Evans, N. Pearson, J.-F. Bonastre and J. Mason, "Influence of task duration in text-independent speaker verification", in Interspeech, 2007, 2007, pp. 794- 797.

[26] A. Larcher, J.-F. Bonastre and J.S.D. Mason, "Short Utterance-based Video Aided Speaker Recognition", in International Workshop on Multimedia Signal Processing, Cairns, Australia 2008.

[27] B. Xiang, U.V. Chaudhari, G.N. Ramaswamy and R.A. Gopinath., "Short-time gaussianization for robust speaker verification", in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Orlando, Florida 2002.

[28] N. Fatima, X. Wu, Thomas Fang Zheng, Chenhao Zhang and G. Wang, "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, Oct. 16-18, 2011.