International Journal of Engineering Research & Management Technology ISSN: 2348-4039
May-2016 Volume 3, Issue-3
Email: editor@ijermt.org www.ijermt.org

SPEAKER RECOGNITION- A FACT BASED APPROACH USING NEURAL NETWORK AND HMM/GMM

Rohit Saini
M.Tech Scholar
Quantum School of Technology
Roorkee

Vishal Vig
M.Tech Guide
Quantum School of Technology
Roorkee

Deepak Painuli
M.Tech Co-Guide
Quantum School of Technology
Roorkee

ABSTRACT

Automatic speaker recognition is a machine to recognize a person from a spoken phrase. These systems can operate in two modes: to identify a particular person or to verify a person's claimed identity. Speech processing and the basic components of automatic speaker recognition systems are shown and design tradeoffs are discussed. Speaker recognition technologies, with an emphasis on front end features for robust speaker recognition. There is several way of characterizing the communication potential of speech. According to information theory, speech can be representing in term of its message content. Speaker recognition has been studied actively for several decades. The applications of speaker recognition technology are quite varied and continually growing. This technique makes it possible to use the speaker's voice for verification of their identity and thereafter enable the control access to services such as voice dialing and voice mail, tele-banking, telephone shopping, database access related services, information services, security control for confidential information areas, forensic applications, and remote access to computers. Speaker recognition technology is expected to create a host of new services that will make our daily lives more convenient. We give an overview of both the classical and the state-of-the-art methods. We start with the fundamentals of automatic speaker recognition, concerning feature extraction and speaker modeling I have developed an approach for finding the individual speaker who will be difficult for the system, using a set of feature statistics calculated over regions of speech. We also provide an overview of this recent development and discuss the evaluation methodology of speaker recognition systems.

KEYWORDS-*component;Speaker identification, Speaker verification, HMM/GMM, Neural network.*

I. INTRODUCTION

Speaker recognition refers to recognizing persons from their voice. The recognition of speaker and speech recognition are very closely related. While speech recognition sets its goals at recognizing the spoken words in speech, the aim of speaker recognition is to identity the speaker by extraction, characterization and recognition of the information contained in the speech signal [1]. Speech processing is a diverse field with many applications [2]. It a few of these areas and how speaker recognition relates to the rest of the field; this paper focuses on the three boxed areas [4].

The most significant factor affecting automatic speaker recognition performance is the variation in the signal characteristics (intersession variability and variability over time). Variations arise from the speakers themselves as well as from the recording and transmission channels, such as:

- Short-term variation due to the speaker's health and emotions
- Long-term changes due to aging
- Different microphones
- Different background noises (closed environment vs. open environment etc.)

It is well known that samples of the same utterance recorded within session are much more highly correlated than samples recorded in separate sessions [3]. This is due to the fact that the speaker and channel effects are bound together in spectrum and hence speaker and channel characteristics are both

involved in the features that are used in speaker recognition systems. Therefore anything that affects the spectrum can cause problems in speaker recognition. Unlike speech recognition systems, which may average out these effects using large amounts of speech?

In this paper, we carry out research to improve the robustness for speaker recognition on distant microphones from two levels: to improve robustness for the traditional system based on low-level acoustic features and to improve robustness using high-level features [3]. From the low-level, we introduced a reverberation compensation approach and applied feature warping in the feature processing of the distant signals. We proposed multiple channel combination approaches to alleviate the issues of acoustic mismatches on far-field speaker recognition. From the high-level, we explored phonetic speaker recognition, in which we try to capture high-level phonetic speaker information and model speaker pronunciation dynamics using such information.

2. SPEAKER IDENTIFICATION AND VERIFICATION

There are a number of tasks that fall into the category of speaker recognition. Speaker recognition encompasses verification and identification. Automatic speaker verification (ASV) is the use of a machine to verify a person's claimed identity from his voice [5].

2.1 Speaker identification Speaker verification is the process of rejecting or accepting the identity claim of a speaker. In most of the applications, voice is used as the key to confirm the identities of a speaker and is classified as speaker verification. The literature abounds with different terms for speaker verification, including voice verification, speaker authentication, voice authentication, talker authentication, and talker verification. Speaker identification aims to identify a speaker who belongs to a group of users through a sample of his speech. In speaker identification, a speech utterance from an unknown speaker is analyzed and compare with models of known speakers. The unknown speaker is identified as the speaker whose model best match the input utterance. Speaker verification is aim to verify the identity of the speaker of the speaker through a compression of some samples of his speech with the reference of the speaker he claim to be. If the match is above the certain threshold, the identity claim is verified. A high threshold makes it difficult for imposters to be accepted by the system, but at the risk of rejecting the generous person. Conversely a low Threshold ensures that the generous person is accepted consistently, but at the risk of accepting impostors [6].

2.2 SPEAKER VERIFICATION

Verifying the identity of a speaker by his voice is called speaker verification. This technique is used generally for access control purposes. The speaker claims his identity and speaks to the system, the system compare the speech data with the model of claimed speaker and give access right or reject him/her according to a certain threshold of acceptance [7].In speaker verification, the task is easier since the speaker is claiming his identity and the system knows which model will be used for comparison. In the test phase, the only thing to do is measuring the similarity between the model of speaker and the speech data, then comparing it to a threshold. Training phase of speaker verification include, unlike speaker identification, threshold determination which consists in fixing a threshold value for each speaker, or a general threshold, which will be used in test phase to take a decision. As the identity of speaker is known", the output is access" or reject" decision.

3. PROPOSED APPROACH

Virtually all state-of-the-art speaker recognition systems use a set of background speakers or cohort speakers in one form or another to enhance the robustness and computational efficiency of the recognizer. In the enrollment phase, background speakers are used as the negative examples in the training of a discriminative model or in training a universal background model from which the target speaker models are adapted. In the recognition phase, background speakers are used in the normalization of the speaker match score. There are several broad areas of prior work relevant to this dissertation. I begin in Section by setting up the speaker recognition problem, while in Sections and I provide details about features, system approaches, relevant speech corpora, and measures of system performance, respectively. There are a number of intrinsic speaker qualities, which account for intra-speaker variability, as well as

differences between speakers, that I describe in Section. The most directly related work involves error analysis pertaining to speaker recognition systems, which I discuss in Section [8].

3.1 HMM/GMM APPROACH

A literal application of this framework to connectionist models would be to test the ratio of phrase-level likelihoods (i.e. the overall cost of the best alignment found by a Vitter HMM decoder) from the recognition of a given utterance by speaker-adapted and speaker-independent models. This, however, turns out to be quite useless: The connectionist acoustic models have been trained to estimate the posterior probabilities of a given phone class, p(qk/X) where qkj are the phone labels and X represents the acoustic features.

3.2 ARTIFICIAL NEURAL NETWORK (ANN) APPROACH

An artificial neural network (ANN), often just called a neural network (NN), is an interconnected group of artificial neurons that uses a mathematical model or computational model for information processing based on a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network. The particular model used in this technique can have many forms, such as multi-layer perceptions or radial basis functions. The MLP is a type of neural network that has grown popular over the past several years. A MLP with one input layer, one hidden layer, and one output layer is shown in paper. MLP's are usually trained with an iterative gradient algorithm known as back propagation [11].

The MLP is convenient to use for problems with limited information regarding characteristics of the input. However, the optimal MLP architecture (number of nodes, hidden layers, etc.) to solve a particular problem must be selected by trial and error, which is a drawback. In addition, the training time required to solve large problems can be excessive, and the algorithm is vulnerable to converging to a local minima instead of the global optimum.

4. EXPERIMENTAL FRAMEWORK

The experimental framework begins with an explanation of the speaker recognition procedure to be employed, followed by a discussion of the system development process. Lastly, the valuation paradigm is presented and the evaluation procedure for each task detailed. These provide solutions for someone who would like to start research in speaker recognition. It may also be useful for speech scientists to have a glance at the current trends in the field. We assume familiarity with basics of digital signal processing and pattern recognition [9].

4.1 FEATURE EXTRACTION

Feature Extraction is a process of reducing data while retaining speaker discriminative information. The amount of data, generated during the speech production, is quite large while the essential characteristics of the speech process change relatively slowly and therefore, they require less data [10].

MFCC feature extraction is usually a non-invertible (lossy) transformation, as the MFCC described pictorially in Fig Making an analogy with filter banks, such transformation does not lead to perfect reconstruction, i.e., given only the features it is not possible to reconstruct the original speech used to generate those features. Computational complexity and robustness are two primary reasons to allow loosing information. Increasing the accuracy of the parametric representation by increasing the number of parameters leads to an increase of complexity and eventually does not lead to a better result due to robustness issues. The greater the number of parameters in a model, the greater should be the training sequence.

5. EVALUATION RESULTS AND ANALYSIS

In this paper, we conducted research work to improve speaker recognition using Neural network on Matlab. Firstly we calculate the MFCC feature for sampling, modeling and pattern matching, The speech signal conveys many levels of information which incorporate: linguistics (e.g., text, language, accent/dialect), speaker specific (e.g., gender, emotion, speaker identity), and environmental information (e.g., communication channels, background noises). This dissertation focuses on addressing speech-pattern recognition for detection of foreign accent and speaker identity information. This describes

classification of speech from native and nonnative speakers, enabling accent-dependent automatic speech recognition. The audio signal cannot be treated as a whole because this would require a lot of calculations for the machine, so the signal is sliced

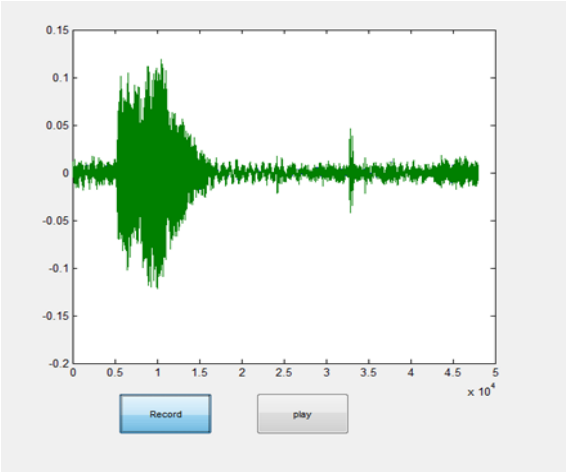

Fig1: input signal

into windows that have the particularity of overlap in half with the aim to have a better treatment for FFT (Fast Fourier Transform). It typically uses a window of N samples,
N is a number that is a power of 2, it is because the FFT [12] algorithm is much faster for these numbers.
5.1 NEURAL NETWORK TRAINING
An artificial neural network (ANN) [14], often just called a neural network (NN), is an interconnected group of artificial neurons that uses a mathematical model or computational model
for information processing based on a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network. The particular model used in this technique can have Fig: Recorded voice

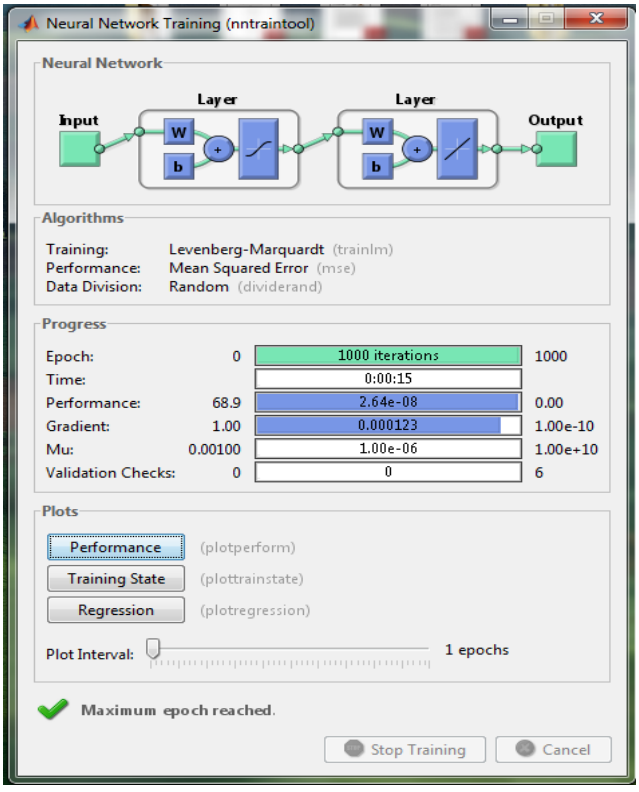

Fig2: Neural network training

many forms, such as multi-layer perceptions or radial basis functions. The MLP is a type of neural network that has grown popular over the past several years. MLP's are usually trained with an iterative gradient algorithm known as
back propagation.
In this paper, we have chosen to use a back propagation neural network since it has been successfully applied to many pattern classification problems including speaker recognition and our problem has been considered to be suitable with the supervised rule. Evaluation of neural network [15] provides performance, training set and Regression. Best validation performance result is 2.8564e-008 at epochs 1000 that show in fig.
Our models achieved 84.0% accuracy on the 1000 epoch test segments from the 2001 term. The 16 errors (2.0%) are all "false errors"; either these segments contain significant overlaps Between different speakers or high background noise, or they are too short (much less than one second). The test on the 100 turns from the 1995 to 2004 terms also showed perfect results, although this test data used different recording devices and was digitalized at different sampling rates. Three of the 100 turns were not correctly identified, but all of them are "false errors".

6. FUTURE WORK AND CONCLUSION

In this thesis we have presented an overview of the classical and new methods of automatic text-independent speaker recognition. The recognition accuracy of current speaker recognition systems under controlled conditions is high. However, in practical situations many negative factors are encountered including mismatched handsets for training and testing, limited training data, unbalanced text, background noise and non-cooperative users. The techniques of robust feature extraction, feature normalization, model-domain compensation and score normalization methods are necessary. The technology advancement as represented by NIST [13] evaluations in the recent years has addressed several technical challenges such as text/language dependency, channel effects, speech durations, and cross-talk speech. However, many research problems remain to be addressed, such as human-related error sources, real-time implementation, and forensic interpretation of speaker recognition scores. Early applications of the technology have achieved varying degrees of success. The promise for the future is significantly higher performance for almost every speech recognition technology area, with more robustness to speakers, background noises etc. This will ultimately lead to reliable, robust voice interfaces to every telecommunications service that is offered, thereby making them universally available. Speech recognition technology has migrated from mini- and mainframe Computers to workstations and personal computers, and applications are already running on them. As embedded digital signal processors become more prevalent on workstations (such as the Next computer), we expect to see much wider use of speech and speaker recognition. Current applications depend on the use of various simplifying constraints that make speech recognition feasible, as discussed above. The dependence on them means that, although useful practical applications of speech recognition exist, we have not yet achieved comfortable and natural communication with computers through voice. Text-dependent speaker recognition exists in the form of operational systems, but accurate text-independent speaker recognition remains a target.
Improvements in the speech and speaker recognition techniques discussed here will no doubt advance the performance of recognition systems, but it seems likely that we will also need natural language understanding before we can achieve comfortable and natural communication with computers through voice.

7. REFERENCES

1. A. Adami, R. Mihaescu, D. A. Reynolds, and J. Godfrey.Modeling Prosodic Dynamics for Speaker Recognition. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Hong Kong, China, 2003.
2. Sheryl R. Young, \Detecting Misrecognitions and Out-Of-Vocabulary Words", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1994.
3. M. Indovina, U. Uludag, R. Snelik, A. Mink, and A. Jain. Multimodal Biometric Authentication Methods: A COTS Approach. In ACM SIGMM2003Multimedia Biometrics Methods and Applications Workshop, Berkeley, CA, 2003.
4. joseph p. Campbell "Speaker Recognition: A Tutorial" IEEE, VOL. 85, NO. 9, September 1997
5. Johan Lindberg and HakanMelin, \Text-prompted versus sound-prompted passwords in speaker veri_cation systems", EUROSPEECH, 1997.

6. Douglas A. Reynolds, \Speaker identi_cation and veri_cation using Gaussian mixture speaker models", Speech Communication, 17(1-2):91-108, August 1995.
7. Toshihiro Isobe, Jun-ichi Takahashi, \A New Cohort Normalization Using Local Acoustic Information For Speaker Veri_cation" ICASSP, 1999.
8. S. Dupont, H. Bourlard, O. Deroo, V. Fontaine, and J. M. Boite, (1997), "Hybrid HMM/ANN Systems for Training Independent Tasks: Experiments on 'Phonebook' and Related Improvements,", Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (Munich, Germany), pp. 1767-1770, 1997.
9. Te-Won Lee and Anthony J. Bell. Blind source separation of real world signals. In Proceedings of Int. Conf.On Neural Networks, pages 2129-2134, 1997. Houston, TX.
10. S. S. Kajarekar, L. Ferrer, E. Shriberg, K. Sonmez, A. Stolcke, A. Venkataraman, and J. Zheng, "SRI's 2004 NIST Speaker Recognition Evaluation System," *Proc. IEEE ICASSP*, vol. 1, pp. 173-176, Philadelphia, 2005
11. Moonasar, V., Venayagamoorthy, G., 2001. A committee of neural networks for automatic speaker recognition (ASR) systems. In: Proc. Internat. Joint Conf. on Neural Networks (IJCNN 2001), Washington, DC, USA, July 2001, pp. 2936–2940.
12. McLaughlin, J., Reynolds, D., Gleason, T., 1999. A study of computation speed-ups of the GMM–UBM speaker recognition system. In: Proc. Sixth European Conf. on Speech Communication and Technology (Eurospeech 1999), Budapest, Hungary, September 1999, pp. 1215–1218.
13. Niemi-Laitinen, T., Saastamoinen, J., Kinnunen, T., Fra¨nti, P., 2005. Applying MFCC-based automatic speaker recognition to GSM and forensic data. In: Proc. Second Baltic Conf. on Human Language Technologies (HLT'2005), Tallinn, Estonia, April 2005, pp. 317–322. NIST 2008 SRE results page, September 2008.
14. J. Oglesby and J. *S.* Mason, "Optimization of neural models for speaker identification," in *Proc. ICASSP*, 1990, pp. 261-264.
15. *L* Fausette, "Fundamentals of Neural Networks-Architecture, Algorithm, and Applications", Prentice Hall, 1994.