International Journal of Engineering Research ISSN:2319-6890)(online),2347-5013(print)
Volume No.3 Issue No: Special 2, pp: 112-115 22 March 2014

Review of Automatic Speech Recognition in Signal Processing

B. Ravi Teja, Asst.Prof in Dept. of IT, RGMCET, Nandyal, raviteja550@gmail.com
M. Suleman Basha, Asst.Prof in Dept. of IT, RGMCET, Nandyal, suleman.ndl@gmail.com
M. Rama Subbiah ,Asst.Prof in Dept. of IT, RGMCET, Nandyal, subhash45229@gmail.com

Abstract : Underlying of speech data refers the speaker features which are useful in speech recognition, speech processing, speech coding, and speech clustering. We described a brief of the area of speaker recognition, speech applications, and their underlying techniques. The review of automatic speech recognition (ASR) will discuss some of the positive and negative aspects of speaker recognition technologies and also outline the potential trends in research, development and applications.

Keywords : ASR, HMMs, ***speech recognition, speech clustering***

1. INTRODUCTION

The speech signal conveys different levels of information to listener. At first level, the speech conveys a message via words. But at next levels the speech will conveys about the language being spoken and the emotion, gender and, generally, the identity of the speaker. The speech recognition aims at recognizing the word spoken in speech, the aim of automatic speaker recognition(ASR) systems is to characterize and identify the information in the speech signal representing speaker identity. The speaker recognition covers two more fundamental tasks. *Speaker identification* is the task of identify who is talking from a set of known voices or speakers. *Speaker verification* (know as speaker authentication) is the task of determining whether a person is who he/she claims to be (a yes/no decision). Since it is generally assumed that imposters (those falsely claiming to be a valid user) are not known to the system, this is called to as an *open-set* task. Adding a "none-of-the-above" option to closed-set identification task.

Merge the two tasks for what is called open-set identification. Depending on the speaker cooperation and control in an application, the speech used for these tasks can be either *text-dependent* or *text-independent*. In text-dependent application, the recognition system requires prior knowledge of the text and it is expected that the speaker will speak this text. Examples of this are a user specific pass-phase or a system prompted phrase. The prior knowledge and constraint of the text can greatly increase performance of the system. In text-independent application, there is no prior knowledge of the text to be spoken by the speaker, text independent recognition system is more complex but also more flexible, for example verification of a speaker while he/she is conducting other speech interactions (background verification). The speaker and speech recognition system merge and speech recognition accuracy increases, the difference between text independent and –dependent applications will decrease. The two basic tasks, text-dependent speaker verification is the most commercially viable and useful technology, although there has been much research conducted on both tasks.

Research on speaker recognition tasks and techniques has been conducted for well over four decade and this continues to be an important area. Approaches have spanned from human hearing and spectrogram comparisons, to simple pattern matching approaches, to more modern pattern recognition approaches, such as neural networks and Hidden Markov Models (HMMs). It is important to note that, although determined to extract and recognize diverse information from the speech signal, many of the same techniques successfully applied to speech recognition and used for speaker recognition.

2. APPLICATIONS

The applications of speaker recognition are quite different and persistently rising. Below is some broad areas where speaker recognition technology has been currently used:

Control Access: Originally for physical facilities, more recent applications are for controlling access to computer networks (add biometric issue to usual password) or websites (thwart password sharing for access sites) and also used for automated password reset.

Transaction Authentication: For telephone banking, superior levels of verification can be used for more responsive transactions. Recent applications are in user verification for remote electronic purchases (e-commerce).

Speech Data Management: voice mail browsing or intelligent react machines, use speaker recognition to tag incoming voice mail with name for browsing. For audio mining applications, for quick indexing and filing interpret recorded meetings or video.

Personalization: Store and retrieve personal setting/preferences using user verification for multi-user site or device.

3. VERIFICATION TECHNOLOGY

The basic structure of modern speaker identification system is shown in below. The system

fundamentally outfits a likelihood ratio test to differentiate between two hypotheses: the test speech arrives from the claimed speaker. characteristics mined from the speech signal in front-end processing and compared to a model representing the claimed speaker. The speaker ratio and imposter match scores is the likelihood ratio statistic (Λ), which is then compared to a threshold (θ) to make decision to accept or reject the speaker. The three main components in identification system, front-end processing, speaker models, and imposter models, are briefly described next.

3.1 Front-end Processing/Feature Extraction

Generally Front-end processing consists of three sub-processes. First, form of speech activity detection is performed to remove non-speech parts from the signal. Next, features are mined from the speech. Although there are no exclusive features conveying speaker identity in the speech signal, from the source-filter theory of speech production it is known that the speech spectrum shape encodes information about the speaker's vocal tract shape via resonances (formants) and glottal source via pitch harmonics. Thus some form of features is used in speaker verification systems. Short-term analysis, typically is used to compute a sequence of magnitude spectra using LPC (all-pole) or FFT analysis. Most commonly the magnitude spectra are then converted to cepstral features after passing through a melfrequency filterbank and time-differential (delta) cepstra are appended. The final front-end processing is channel compensation. It is well known that different input devices will impose different spectral characteristics on the speech signal, such as bandlimiting and shaping. Since verification systems strive to operate independent of the input device. Channel compensation aims to remove these channel effects. Most commonly some form of linear channel compensation, such as long- and short-term cepstral mean subtraction, are applied to features. In addition to channel compensation in the feature domain, there are powerful compensation techniques that can be applied in the model and match domains as well as adaptation techniques to effectively use new data to learn channel characteristics.

3.2 Speaker Modelling

During enrolment, speech is passed through the front-end processing steps and the feature vectors are used for creating a speaker model. popular attributes of a speaker model are: (1) a theoretical model behavior and mathematically approach extensions and improvements; (2) generalize with new data so that the model does not over fit the data and can match new data; (3) economical representation in both size and computation. There are many modelling techniques that have all of these attributes and have been used in speaker identification systems. The model selection is mainly dependent on the type of speech , the expected performance, the simplicity of training and updating, and computation considerations.

Template Matching: In this technique, the model consists of a sequence of feature vectors. During verification a match score is produced by using dynamic time warping (DTW) to align and measure the similarity between the test phrase and the speaker template. This approach is used for text-dependent applications.

Nearest Neighbour: In this, no explicit model is used, in its place all features vectors from the speech are reserved to represent the speaker. During identification, the match is computed by the distance of each test feature vector to its k nearest neighbours in the speaker's training vectors.

Neural Networks: This model can have many forms, such as multi-layer perceptions or radial. these models are explicitly trained to discriminate between the speaker being modelled and some alternative speakers. Training can be computationally pricey and models are sometimes not complex.

Hidden Markov Models: This technique uses HMMs, which encode the evolution of the features and efficiently model statistical variation of the characteristics. During employment HMM parameters are estimated from the speech using recognized automatic algorithms. During confirmation, the likelihood of the test feature sequence is computed against the speaker's HMMs. For text-dependent applications, multi-state left-to right HMMs are used. For text-independent applications, Gaussian Mixture Models (GMMs), are used.

3.3 Imposter Modeling

in spite of the rather late introduction of imposter modeling to speaker recognition (late 1980s), the use of an imposter model is widespread and can be critical to obtaining good performance. Fundamentally it acts as a normalization to help minimize non-speaker related changeability in the likelihood ratio score. There are two foremost approaches used to explain the imposter model in the likelihood ratio test. Usually these approaches can be applied to any speaker modelling technique. The first method, known as likelihood sets, uses a collection of other speaker models to calculate the imposter match score. Normally the imposter match score is computed as a function, such as the maximum or average of the match scores from a set of non-claimed speaker models. The non-claimed models can come from other enrolled speakers or as fixed models from a different database. Different techniques have been observed for the choose

and use of background speaker sets. The second method, called as general, world or universal background model(UBM), uses a single speaker-independent model trained on speech from a huge number of speakers. Here the idea is to signify imposters using a general speech model, this is compared to specific speaker model. The advantage of this method is that only a single model needs to be trained and scored. In addition, this method has been exposed to provide better performance in for some applications (for example in the NIST text-independent evaluations). This approach also allows the use of Maximum A-Posteriori (MAP) training to adapt the claimant model from the background model, which can increase performance and decrease computation and model storage requirements [D.A. Reynolds in 3].

4. PERFORMANCE

It is quite difficult to characterize the performance of speaker verification systems in all applications due to the complexities and differences in the enrollment/testing scenarios. However, in this section we attempt to provide a range of performance for some broad cases. These numbers are not meant to indicate the best performance that can be obtained, but rather a relative ranking of some different scenarios. In Figure 2 we depict a detection error tradeoff (DET) plot, which shows the tradeoff between false-rejects and false-accepts as the decision threshold changes in a verification system. On this DET we show four equal error rate points (EER is a summary performance indicator where FR=FA) for four different verification experiments.

1) Text-dependent using combinations lock phrases (e.g., 35-41- 89). Clean data recorded using a single handset over multiple sessions. Used about 3 min of training data and 2 s test data. (0.1% – 1%)

2) Text-dependent using 10 digit strings. Telephone data using multiple handsets with multiple sessions. Two strings training data and single string verification (1%-5%)

3) Text-independent using conversational speech. telephone data using multiple handsets with multiple sessions. Two minutes training data and 30 s test data. (7%-15%)

4) Text-independent using read sentences. Very noisy radio data using multiple military radios and icrophones with multiple sessions. Thirty sec training and 15 s testing. (20%-35%)

One observed theme in these cases is that performance tends to improve with increasing constraints on the application (more speech, less noise, known channels, text-dependent). Determining acceptable performance for a particular application will depend on the benefit of replacing any current verification procedure, the threat model (claimant to imposter attempts) and the relative costs of errors.

5. STRENGTHS AND WEAKNESSES

It is clear that speaker verification technology is indeed ready for use. But, as stated before, it is not the universal solution. The main strength of speaker verification technology is that it relies on a signal that is natural and unobtrusive to produce and can be obtained easily from almost anywhere using the familiar telephone network (or internet) with no special user equipment or training. This technology has prime utility for applications with remote users and applications already employing a speech interface. Additionally, speaker verification is easy to use, has low computation requirements (can be ported to cards and handhelds) and, given appropriate constraints, has high accuracy. Some of the flexibility of speech actually lends to its weaknesses. First, speech is a behavioral signal that may not be consistently reproduced by a speaker and can be affected by a speaker's health (cold or laryngitis). Second, the varied microphones and channels that people use can cause difficulties since most speaker verification systems rely on low-level spectrum features susceptible to transducer/channel effects. Also, the mobility of telephones means that people are using verification systems from more uncontrolled and harsh coustic environments (cars, crowded airports), which can stress accuracy. Robustness to channel variability is the biggest challenge to current systems. Spoofing of systems is often cited as a weakness, but there have been may approaches developed to thwart such attempts (prompted phrases, knowledge verification). There is current effort underway to address these known weaknesses. Some of these weaknesses may be overcome by combination with a complementary biometric, like face recognition.

6. FUTURE TRENDS

Exploitation of higher-levels of information: In addition to the low-level spectrum features used by current systems, there are many other sources of speaker information in the speech signal that can be used. These include idiolect (word usage), prosodic measures and other long-term signal measures. This work will be aided by the increasing use of reliable speech recognition systems for speaker recognition R&D. High-level features not only offer the potential to improve accuracy, they may also help improve robustness since they should be less susceptible to channel effects.

Focus on real world robustness: Speaker recognition continues to be data-driven field, setting the lead among other biometrics in conducting benchmark evaluations and research on realistic data. The continued ease of collecting and making available speech from real

applications means that researchers can focus on more real-world robustness issues that appear. Obtaining speech from a wide variety of handsets, channels and acoustic environments will allow examination of problem cases and development and application of new or improved compensation techniques.

Emphasis on unconstrained tasks: With text-dependent systems making commercial headway, R&D effort will shift to the more difficult issues in unconstrained situations. This includes variable channels and noise conditions, text-independent speech and the tasks of speaker segmentation and indexing of multi-speaker speech.

REFERENCES

i. *S. Furui. Recent advances in speaker recognition. AVBPA97, pp 237--251, 1997*

ii. *J. P. Campbell, ``Speaker recognition: A tutorial,'' Proceedings of the IEEE, vol. 85, pp. 1437--1462, September 1997.*

iii. *Special Issue on Speaker Recognition, Digital Signal Processing, vol. 10, January 2000. http://www.idealibrary.com/links/toc/dspr/10/1/0*

iv. *R. Teunen, B. Shahshahani, and L. Heck, ``A Model-based Transformational Approach to Robust Speaker Recognition,'' ICSLP October 2000.*