Automatic Speaker Recognition: Current Approaches and Future Trends[•][1]

Douglas A. Reynolds

MIT Lincoln Laboratory, Lexington, MA USA
dar@ll.mit.edu

ABSTRACT

In this paper we provide a brief overview of the area of speaker recognition, defining terminology, discussing applications, describing underlying techniques and providing some indications of performance. Following this overview we compare speaker verification to other biometrics and discuss some of the strengths and weaknesses of speaker verification technology. Finally we outline some potential future trends in research and development.

1. INTRODUCTION

The speech signal conveys many levels of information to the listener. At the primary level, speech conveys a message via words. But at other levels speech conveys information about the language being spoken and the emotion, gender and, generally, the identity of the speaker. While speech recognition aims at recognizing the word spoken in speech, the goal of automatic speaker recognition systems is to extract, characterize and recognize the information in the speech signal conveying speaker identity.

The general area of speaker recognition encompasses two more fundamental tasks. *Speaker identification* is the task of determining who is talking from a set of known voices or speakers. The unknown person makes no identity claim and so the system must perform a 1:N classification. Generally it is assumed the unknown voice must come from a fixed set of known speakers, thus the task is often referred to as *closed-set* identification. *Speaker verification* (also know as speaker authentication or detection) 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 referred to as an *open-set* task. By adding a "none-of-the-above" option to the closed-set identification task one can merge the two tasks for what is called open-set identification. In general, most compelling applications of speaker recognition technology use open-set speaker verification.

Depending on the level of user cooperation and control in an application, the speech used for these tasks can span from *text-dependent* to *text-independent*. In a pure text-dependent application, a speaker speaks the same text during enrollment and verification and the recognition system has prior knowledge of this text. An example of this would be a common or user-specific pass-phase (e.g., "Open sesame"). Without being as rigid, a *text-constrained* application allows a speaker to use text from a limited vocabulary, such as the digits. The system has prior knowledge of the constrained vocabulary to be used and may have exact knowledge of the text to be spoken, as when using prompted phrases. In both text-dependent and text-constrained applications it is expected that the user will cooperatively speak the fixed text or words from the prescribed vocabulary. The prior knowledge and constraint of the text can greatly boost performance of a recognition system. In a text-independent application, there is no prior knowledge by the system of the text to be spoken, such as when using extemporaneous speech. Text-independent recognition is more difficult but also more flexible, for example allowing verification of a speaker while he/she is conducting other speech interactions (background verification). As speaker and speech recognition system merge and speech recognition accuracy improves, the distinction between text-independent and –dependent applications will decrease. Of the two basic tasks, text-dependent speaker verification is currently the most commercially viable and useful technology, although there has been much research conducted on both tasks.

Research and development on speaker recognition methods and techniques has been undertaken for well over four decades and it continues to be an active area. Approaches have spanned from human aural and spectrogram comparisons, to simple template matching, to dynamic time-warping approaches, to more modern statistical pattern recognition approaches, such as neural networks and Hidden Markov Models (HMMs). It is interesting to note that, although striving to extract and recognize different information from the speech signal, many of the same features and techniques successfully applied to speech recognition have also been used for speaker recognition.

Over this same time, research and development corpora have evolved from small, private corpora (5-10 speaker) under laboratory clean, controlled conditions (single session, read speech) to large, publicly available corpora (500+ speakers) reflecting more realistic and challenging conditions (extemporaneous speech from landline and cellular telephone channels). Benchmark evaluations using common corpora and paradigms have been conducted for several years (e.g. YOHO, CAVE project, NIST http://www.nist.gov/speech/tests/spk) allowing comparison of technical approaches and focusing effort on common challenges. The field has matured to the point that commercial applications of speaker recognition have been steadily increasing since the mid-1980s, with a large number of companies currently offering this technology.

Most of the commercialization has focused on using speaker verification as a biometric to control access to information,

[•] This work is sponsored by the Department of Defense under Air Force Contract F19628-00-C-0002. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Government

[1] This paper is based in part on the tutorial "Speaker Verification: From Research to Reality" by D.A. Reynolds and L.P. Heck, ICASSP 2001.

services, or computer accounts. As with other biometrics, speaker verification offers the ability to replace or augment PINSs and passwords with something that cannot be forgotten lost or stolen. There are two main factors that make speaker verification a compelling biometric:

- Speech is a natural signal to produce that is not considered threatening by users to provide. In many applications speech may be the main (or only, e.g., telephone transactions) modality, so users do not consider providing a speech sample for authentication as a separate or intrusive step.

- The telephone system provides a ubiquitous, familiar network of sensors for obtaining and delivering the speech signal. For telephone based applications, there is no need for special signal transducers or networks to be installed at application access points since a cell phone gives one access almost anywhere. Even for non-telephone applications, sound-cards and microphones are low-cost and readily available.

Despite these strengths, speaker verification, as all biometrics, has limitations and is certainly not the best tool for every application. With the increased interest, urgency and hype in applying biometrics to improve security, it is quite important to understand the pros and cons of any biometric before proceeding with deployment.

In this paper we provide a brief overview of the area of speaker recognition, describing applications, underlying techniques and some indications of performance. This is not a comprehensive review and readers should see, for example, [1], [2] and papers in [3], [4], [5], [6] for more details and references. Following this overview we discuss some of the strengths and weaknesses of current speaker recognition technologies and outline some potential future trends in research, development and applications.

2. APPLICATIONS

The applications of speaker recognition technology are quite varied and continually growing. Below is an outline of some broad areas where speaker recognition technology has been or is currently used (this is not an exhaustive list). A search on the web will produce numerous pointers to companies and products (some lists can be found at, http://www.biometrics.org/html/voice_speaker.html http://www.biometricscatalog.org/).

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

Transaction Authentication: For telephone banking, in addition to account access control, higher levels of verification can be used for more sensitive transactions. More recent applications are in user verification for remote electronic and mobile purchases (e- and m-commerce).

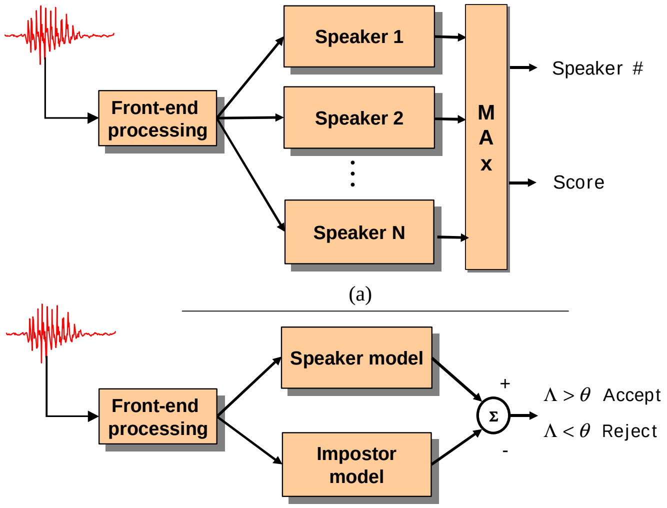
Law Enforcement: Some applications are home-parole monitoring (call parolees at random times to verify they are at home) and prison call monitoring (validate inmate prior to outbound call). There has also been discussion of using automatic systems to corroborate aural/spectral inspections of voice samples for forensic analysis.

Speech Data Management: In voice mail browsing or intelligent answering machines, use speaker recognition to label incoming voice mail with speaker name for browsing and/or action (personal reply). For speech skimming or audio mining applications, annotate recorded meetings or video with speaker labels for quick indexing and filing.

Personalization: In voice-web or device customization, store and retrieve personal setting/preferences based on user verification for multi-user site or device (car climate and radio settings). There is also interest in using recognition techniques for directed advertisement or services, where, for example, repeat users could be recognized or advertisements focused based on recognition of broad speaker characteristics (e.g. gender or age).

3. RECOGNITION TECHNOLOGY

The basic structure for a speaker identification and a verification system is shown in Figure 1 (a) and (b), respectively. In both systems, the speech signal is first processed to extract features conveying speaker information. In the identification system these features are compared to a bank of models, obtained from previous enrollment, representing the speaker set from which we wish to identify the unknown voice. For closed-set identification, the speaker associated with the most likely, or highest scoring model is selected as the identified speaker. This is simply a maximum likelihood classifier.



(a)

(b)

Figure 1 Basic structure of (a) speaker identification and (b) speaker verification systems.

The verification system essentially implements a likelihood ratio test to distinguish between two hypotheses: the test speech comes from the claimed speaker or from an imposter. Features extracted from the speech signal are compared to a model representing the claimed speaker, obtained from a previous enrollment, and to some model(s) representing potential imposter speakers (i.e., those *not* the claimed speaker). The ratio (or difference in the log domain) of speaker and imposter match scores is the likelihood ratio statistic (Λ), which is then compared to a threshold (θ) to decide whether to accept or reject the speaker.

The general techniques used for the three main components of these systems, namely, front-end processing, speaker models, and imposter models, are briefly described next.

3.1 Front-end Processing/Feature Extraction

Front-end processing generally consists of three sub-processes (see Figure 2). First, some form of speech activity detection is performed to remove non-speech portions from the signal. Next, features conveying speaker information are extracted 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 spectral based features is used in most speaker verification systems. Short-term analysis, typically with 20 ms windows placed every 10 ms, is used to compute a sequence of magnitude spectra using either LPC (all-pole) or FFT analysis. Most commonly the magnitude spectra are then converted to cepstral features after passing through a mel-frequency filterbank and time-differential (delta) cepstra are appended. Typical feature vectors will have 24-40 elements.

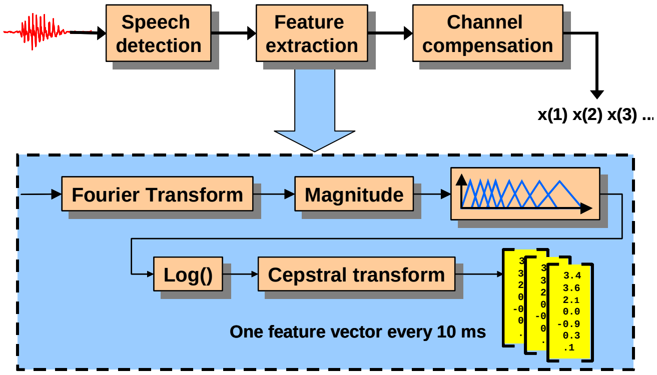


Figure 2 Front-end signal processing to extract features from speech

The final process in front-end processing is some form of channel compensation. It is well known that different input devices (e.g., different telephone handsets) will impose different spectral characteristics on the speech signal, such as bandlimiting and shaping. Since recognition systems strive to operate independent of the input device used (e.g., enroll on office telephone and verify using cell or pay telephone) channel compensation aims at removing 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 also powerful compensation techniques that can be applied in the model and match score domains (e.g., [7] and Aukenthaler et. al in [3]) as well as adaptation techniques to effectively use new data to learn channel characteristics. Details of these approaches are beyond the scope of this paper.

3.2 Speaker Modeling

During enrollment, speech from a speaker is passed through the front-end processing steps described above and the feature vectors are used to create a speaker model[2]. Desirable attributes of a speaker model are: (1) a theoretical underpinning so one can understand model behavior and mathematically approach extensions and improvements; (2) generalizable to new data so that the model does not over fit the enrollment data and can match new data; (3) parsimonious representation in both size and computation. There are many modeling techniques that have some or all of these attributes and have been used in speaker verification systems. The selection of modeling is primarily dependent on the type of speech to be used, desired performance, the ease of training and updating, and storage and computation considerations. A brief description of some of the more prevalent modeling techniques is given next.

Template Matching: In this technique, the model consists of a template that is a sequence of feature vectors from a fixed phrase. 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 almost exclusively for pure text-dependent applications.

Nearest Neighbor: In this technique, no explicit model is used; instead all features vectors from the enrollment speech are retained to represent the speaker. During verification, the match score is computed as the cumulated distance of each test feature vector to its k nearest neighbors in the speaker's training vectors. To limit storage and computation, feature vector pruning techniques are usually applied [8].

Neural Networks: The particular model used in this technique can have many forms, such as multi-layer perceptions or radial basis functions. The main difference with the other approaches described is that these models are explicitly trained to discriminate between the speaker being modeled and some alternative speakers. Training can be computationally expensive and models are sometimes not generalizable.

Hidden Markov Models: This technique uses HMMs, which encode the temporal evolution of the features and efficiently model statistical variation of the features, to provide a statistical representation of how a speaker produces sounds. During enrollment HMM parameters are estimated from the speech using established automatic algorithms. During verification, the likelihood of the test feature sequence is computed against the speaker's HMMs. For text-dependent applications, whole phrases or phonemes may be modeled using multi-state left-to-right HMMs. For text-independent applications, single state HMMs, also known as Gaussian Mixture Models (GMMs), are used. From published results, HMM based systems generally produce the best performance.

[2] Often the term *Voice Print* is used to refer to the speaker model in verification/biometric systems.

3.3 Imposter Modeling

One problem in using a speaker model's raw match score to a test utterance in a verification system is that non-speaker related variability (e.g., text, microphone, noise) can cause large variations in the score from test to test making it very difficult to set a fixed threshold for accept/reject decisions. By using an imposter model to create a likelihood ratio score, the system can use this relative score to allow more consistent threshold settings. Surprisingly, the idea of using an imposter model was not introduced until the early 1990s, but its use in speaker verification systems is widespread and can be crucial for good performance.

For closed-set identification this normalization is not as important since the classification is made by rank ordering speaker model match scores relative to each other. If, however, the application is open-set, then some form of normalization is required to allow stable threshold setting.

There are two dominant approaches used for representing the imposter model in the likelihood ratio test. Generally these approaches can be applied to any speaker modeling technique[3]. The first approach, known as likelihood sets, cohorts or background sets, uses a collection of other speaker models to compute the imposter match score [8]. The imposter match score is usually computed as a function, such as the max or average, of the match scores from a set of non-claimant speaker models. The non-claimant speaker models can come from other enrolled speakers or as fixed models from a different corpus. Various techniques have been examined for the selection and use of background speaker sets.

The second approach, known as general, world or universal background modeling, uses a single speaker-independent model trained on speech from a large number of speakers to represent speaker-independent speech [9]. The idea here is to represent imposters using a general speech model, which is compared to a speaker-specific speech model. The advantage over the cohort approach is that only a single imposter model needs to be trained and scored. Additionally, this approach has been shown 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 (Reynolds et. al in [3]).

4. PERFORMANCE

It is quite difficult to characterize the performance of speaker recognition 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.

Some of the broad factors that can affect the performance of a speaker recognition system are:

- *Speech quality*: types of microphones used, ambient noise levels, types of noise, compression of speech, etc.
- *Speech modality*: text-dependent or text–independent
- *Speech duration*: amount to train and test data, temporal separation of training and testing data.
- *Speaker population:* number and similarity of speakers.

No one data set will match the factors for all applications, but to get realistic results from the evaluation of a system, the data and design should match the target application as much as possible. Knowing performance using clean fixed-text speech will indicate little about performance using free-text, telephone speech.

In Figure 3 we show performance of closed set speaker identification for a text-independent system on three corpora. Performance is shown here as the percent error (misidentifications of the true speaker) for increasing speaker set sizes. TIMIT is an unrealistically pristine corpus of 630 speakers, reading sentences in a sound booth. NTIMIT is the TIMIT corpus re-played through the telephone network. And Switchboard (SWBI) is conversational speech recorded over the telephone network. Two general points can be made from this plot. First, as the speech quality degrades, in this case by passing through the telephone system, performance also degrades. Second, error rate increases with the size of the speaker set. This occurs because there are more speakers to distinguish among as the speaker set size increases.

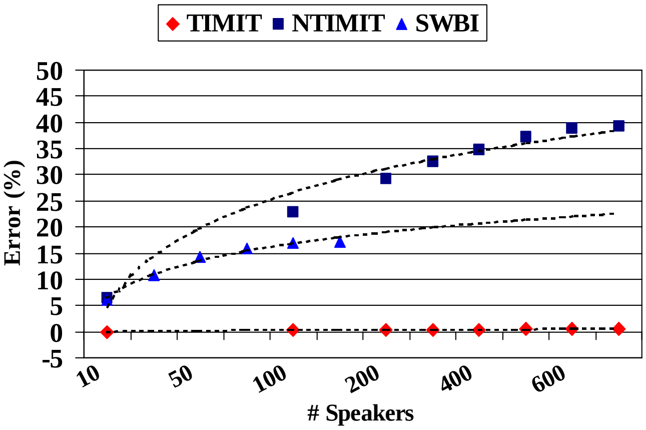


Figure 3 Closed set identification performance on three corpora.

In Figure 4 we depict a detection error tradeoff (DET) plot, which shows the tradeoff between false-rejects (FR) and false-accepts (FA) as the decision threshold changes in a verification system. A verification system is making a single comparison of test speech to a claimed speaker model, so results are not a function of speaker set size.

On this DET we show four equal error rate points (EER is a summary performance indicator where FR=FA) for four different verification experiments.

[3] Even though neural network models already use imposter speech for discriminative training, they too have shown improvement using these approaches.

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 microphones with multiple sessions. Thirty sec training and 15 s testing. (20%-35%)

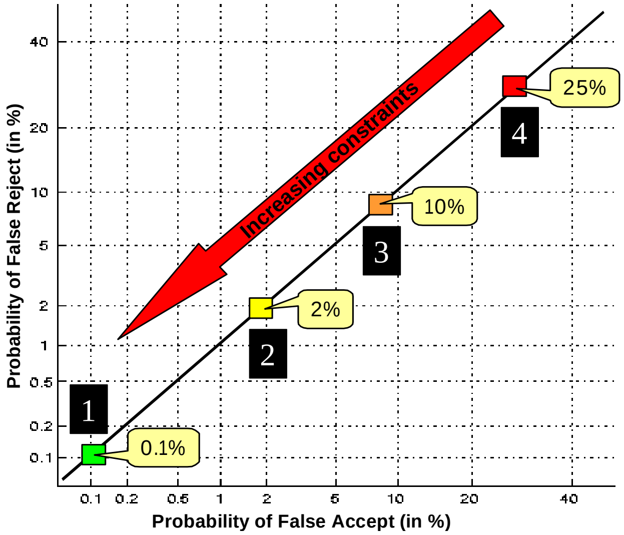


Figure 4 Range of speaker verification performance.

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. COMPARISON TO OTHER BIOMETRICS

It is commonly asked, how does speaker verification compare to other biometrics, such as iris, fingerprint or face recognition? There really is no complete way to compare different biometrics since there are so many dimensions on which to evaluate a biometric (accuracy, suitability for application, ease of use, recognition time, cost, etc). However, in this section we discuss some of the strengths and weaknesses of speaker verification and point to one study which attempted to compare several biometrics based on accuracy.

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 smartcards 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 acoustic 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 many approaches developed to thwart such attempts (prompted phrases, knowledge verification). There is current efforts underway to address these known weaknesses and some of these weaknesses may be overcome by combination with a complementary biometric, like face recognition.

Finally, we show some results from a study by the United Kingdom's Communications-Electronics Security Group (CESG) that attempted to compare performance of several biometrics. The complete report can be found in [10]. In Figure 5 we show a DET plot for eight systems (1 face, 3 fingerprint, 1 hand, 1 iris, 1 vein and 1 voice). While it is debatable that a test can be conducted to compare all these biometrics, it is interesting to note that voice verification performed quite well. Readers, however, should read the report to get all the details of the test.

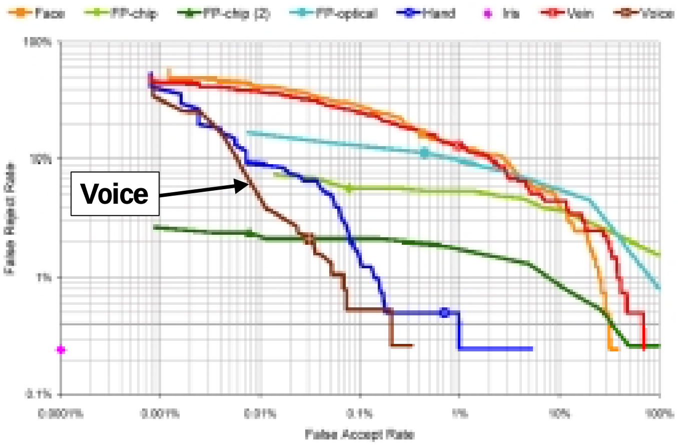


Figure 5 DET curves from CESG study comparing several biometrics. (Best of three attempts Figure 6 [10]).

6. FUTURE TRENDS

In this section we briefly outline some of the trends in speaker recognition research and development.

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.

In recent work, Doddington has shown that a speaker's idiolect can be used to successfully verify a person [11], and Andrews et. al [12] have used n-grams of phonetic sequences for verifying speakers.

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. Currently NIST conducts annual speaker verification evaluations in which participation is open to any interested parties [13].

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. Increasingly speaker segmentation and clustering techniques are being used to aid in adapting speech recognizers and for supplying metadata for audio indexing and searching. This data is very often unconstrained and may come from various sources (for example broadcast news audio with correspondents in the field).

There has been significant progress made in speaker recognition technology and applications, but there still remain many current problems to overcome, such as channel and noise robustness, as well as new areas to explore.

REFERENCES

[1] S. Furui. Recent advances in speaker recognition. AVBPA97, pp 237--251, 1997

[2] J.P. Campbell, ``Speaker recognition: A tutorial," Proceedings of the IEEE, vol. 85, pp. 1437--1462, September 1997.

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

[4] Proceedings of the 2001 Odyssey Workshop, 18-22 June 2001, Crete, Greece http://www.isca-speech.org/archive.html

[5] Reconnaissance du Locuteur et ses Applications Commerciales et Criminalistiques (RLA2C), Avignon, France, 1998 [proceedings in English]

[6] ESCA Workshop on Automatic Speaker Recognition, Identification, and Verification, Martigny, Switzerland 1994 http://www.isca-speech.org/workshops.html

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

[8] A. Higgins, L. Bahler, and J. Porter, ``Speaker Verification using Randomized Phrase Prompting,'' Digital Signal Processing, vol. 1, pp. 89--106,1991

[9] M. Carey, E. Parris, and J. Bridle, ``A Speaker Verification System using Alphanets," ICASSP, pp. 397--400, May 1991

[10] CESG Biometrics Website http://www.cesg.gov.uk/biometrics

[11] G. R. Doddington, "Speaker Recognition based on Idiolectal Differences between Speakers," Eurospeech 2001

[12] A. D. Andrews, M. A. Kohler and J. P. Campbell, "Phonetic Speaker Recognition," Eurospeech 2001

[13] NIST speaker recognition website http://www.nist.gov/speech/tests/spk