Voice Activity Detection for GMM-based Speaker Recognition

Ivan A. Rakhmanenko

Department Of Complex Information Security
Tomsk State University of Control Systems and Radioelectronics, Tomsk
ria@keva.tusur.ru

Abstract

This paper overviews significance of voice activity detection for speech related problems, the application sphere of speaker identification systems and illustrates the use of the simple energy-based voice activity detector in an automatic text-independent speaker identification task. The experimental evaluation of the voice activity detector and the Gaussian mixture model (GMM) for speaker recognition is conducted on a 50 speaker set and a result is presented.

1. Introduction

Automatic speaker recognition task is one of the biggest problems in speech processing field. Methods that are used in modern speaker recognition systems are not perfect.

There are models that work effectively in acoustically clean environment but not in low signal-noise ratio environment. Requirements for voice identification accuracy are constantly increasing because of the growing spreading of biometric multi-factor authentication systems. These systems include remote voice authentication banking account management systems, access control systems and others. All these systems require high accuracy of speaker recognition in order to satisfy customers' needs.

Equal error rate value (EER) is one of the most common values of such accuracy measuring used nowadays. EER is used both for text-dependent and text-independent automatic voice authentication systems. By now the best speaker recognition systems are characterized by 3-5% EER values [1]. This accuracy is absolutely insufficient for modern speaker verification systems because if there are a lot of speakers working with such systems errors, then mistakes will occur definitely, and such mistakes are unacceptable in systems granting access rights to confidential data or banking accounts.

If there are impulse or background noises, segments without speech signal in source signal used for speaker recognition accuracy of recognition could deteriorate. Therefore, voice activity detection techniques could be applied to speaker recognition task increasing effectiveness of speaker recognition module.

The application field of currently developed voice authentication systems includes multi-factor (biometric) authentication and access restriction systems, banking account management systems using voice biometrics in order to give speaker access to his banking account, national security and anti-terrorism issues. The use of speaker recognition systems that have even small possibility of mistake in such a sensitive application areas could be very dangerous.

Another application field of speaker recognition is forensic examinations. This is a special case of speaker recognition task where speaker verification is used. In this case the speaker does not want to cooperate with system because it could prove his guilt. A lot of attention should be paid to verification accuracy while developing such forensic speaker verification systems. If there was an error in such systems, an object of a forensic examination would be falsely accused if he is not guilty or a guilty person would be released of accusations.

Speaker recognition encompasses verification and identification. Automatic speaker verification (ASV) is a verification of a person's claimed identity from his voice. In automatic speaker identification (ASI), there is no a priori identity claim, and the system decides who the person is, what group the person is a member of, or (in the open-set case) that the person is unknown [2]. Automatic speaker identification system that is presented in this paper works with the closed-set identification problem, deciding who is the owner of the speech signal presented to the system. Existence of speakers that are not registered in the system is not taken into consideration.

2. Voice activity detection

Important step of the speech signals preprocessing is detecting what segments contain speech and what does not. The task of separating speech from background noise is not a trivial one except in the case of acoustic environment with extremely high signal-noise ratio [3]. Voice activity detectors use short time processing due to non-stationary nature of the speech signal.

Short Term Processing of speech can be performed either in time domain or in frequency domain. Short term energy, short term zero crossing rate and others can be computed from the time domain processing of speech. Alternatively, short term Fourier transform can be computed from the frequency domain processing of speech. Mel frequency cepstral coefficients (MFCC) could be used as the frequency domain parameters. Each of these parameters gives different information about speech that can be used for automatic processing.

2.1. Short Term Energy Parameter

By the nature of production, the speech signal consist of voiced, unvoiced and silence regions. Further the energy associated with voiced region is large compared to unvoiced region and silence region will not have least or negligible energy. Thus short term energy can be used for voiced, unvoiced and silence classification of speech.

Short term energy can be calculated as follows

$$E(n) = \sum_{m=-\infty}^{\infty} (s(m) \cdot w(n-m))^2 \qquad (1)$$

where $w(n)$ represent the windowing function of finite duration.

2.2. Short Term Zero Crossing Rate (ZCR)

Zero Crossing Rate gives information about the number of zero-crossings present in a given signal. Intuitively, if the number of zero crossings are more in a given signal, then the signal is changing rapidly and accordingly the signal may contain high frequency information. On the similar lines, if the number of zero crossing are less, hence the signal is changing slowly and accordingly the signal may contain low frequency information. Thus ZCR gives an indirect information about the frequency content of the signal.

The ZCR in case of non-stationary signal is defined as,

$$z = \sum_{n=-\infty}^{\infty} \left| \mathrm{sgn}(s(n)) - \mathrm{sgn}(s(n-1)) \right| \qquad (2)$$

where *sgn (s(n))* = 1 if *s(n)* ≥ 0, *sgn(s(n))* = 0 if *s(n)* < 0.

In case of unvoiced sounds like |s|, the ZCR value is significantly high compared to the region of voiced sounds like |a| and hence can be used for distinguishing voiced and unvoiced regions. Also, ZCR value of voiced sounds is relatively high compared to regions of silence region that contains background noise.

3. Features extraction

Mel frequency cepstral coefficients (MFCC) are used as feature vectors in this specific paper as well as in many other scientific works dedicated to speaker recognition.

Mel frequency cepstral spectrum transform method was first introduced in [4]. MFCC are used for speaker recognition, speech recognition and other speech related applications from 10 to 30. In some systems a delta and a double delta features related to the change in cepstral features over time are added. Besides MFCC, the energy from the frame is added to feature vector. The energy in a frame is the sum over time of the power of the samples in the frame.

Feature vectors extraction process is shown in Fig. 1. First step of feature vector extraction is windowing – taking a small part of speech signal instead of the whole signal. Hamming windows were used for MFCC calculation. Window length is 20 ms, window shift is 10 ms. Discrete Fourier Transform (DFT) is performed after windowing.

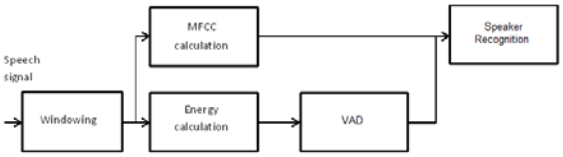

Figure 1: Features extraction and preprocessing diagram.

Next step of feature vector extraction is warping frequencies outputted by DFT to the mel scale defined as,

$$f_{mel} = 1125 \ln (1 + f / 700) \qquad (3)$$

The mapping between frequency in hertz and the mel scale is linear below 1000 Hz and logarithmic above 1000 Hz [5]. A bank of triangular filters is created for implementing this scaling (Fig. 2) and the log energy is collected from each of these frequency bands [4]. The final step of MFCC extraction is the inverse Discrete Fourier Transform (IDFT).

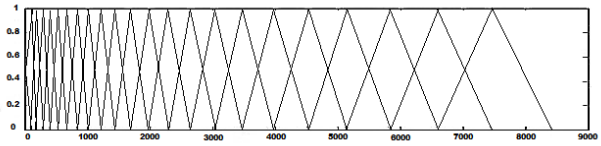

Figure 2: The mel filter bank.

Despite the fact that there are no features in spectrum and cepstrum that could help distinguish speakers, nevertheless it could be effectively used in automatic speaker recognition (ASR) task [6]. This is possible due to the fact that the spectrum reflects the structure of the human vocal tract, which is the main physiological factor that allows us to distinguish voices.

4. Gaussian mixture model

The choice of decision rules composition method is very important in ASR task. The most common methods are Gaussian Mixture Model (GMM), Support Vector Machines (SVM), Hidden Markov Models (HMM), neural networks and factor analysis modifications. GMM is used in ASI system presented in this paper.

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities [7]. A GMM with M component Gaussian densities can be presented by the equation

$$p(x \mid \lambda) = \sum_{i=1}^{M} w_i g(x \mid \mu_i \Sigma_i), \qquad (4)$$

where x is a D-dimensional continuous-valued data vector (i.e. measurement or features), w_i, i = 1,...,M, are the mixture weights, and $g(x|\mu_i, \Sigma_i)$, i = 1,...,M, are the component Gaussian densities with mean vector μi and covariance matrix Σ_i. The complete GMM is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. It could be represented by the equation,

$$\lambda = \{w_i, \mu_i, \Sigma_i\} \qquad (5)$$

Each speaker is represented by his Gaussian mixture λ for speaker identification task.

There are two reasons for using Gaussian mixture densities as a representation of speaker identity [8]. The first reason is the intuitive notion that the individual component densities of the GMM may model some underlying set of acoustic classes, reflecting some general speaker-dependent vocal tract configurations. The second reason is the empirical observation that a linear combination of Gaussian basis functions is capable of representing a large class of sample distributions. A GMM can form smooth approximations to arbitrarily-shaped densities.

5. Experimental evaluation

The experiments were conducted using speech database containing collection of speech from 25 male and 25 female

speakers. This speech database includes records of proverbs and sentences from science fiction stories read without preparation. The total length of speech for each speaker is at least 6 minutes. Each speaker was recorded using medium-quality microphone, 8000 Hz sampling rate, 16 Bit sample size. All records have high level background noise, making recognition task more challenging.

During the experiment 10-fold consecutive random sampling validation was conducted. In each iteration of validation, speech was randomly divided on training and testing data. 3 various lengths of training data were used: 30, 60 and 90 seconds for each speaker. Remaining data that was not used in training was involved in testing. Length of one testing speech sample is 10 seconds.

One test segment is considered to be identified correctly if speaker's number that was identified by the system matches actual speaker's number of test segment. The final performance evaluation was calculated as the percent of correctly identified test segments over all test utterances. Results of the experiment using only the Gaussian mixture are given in Table 1, where N is number of mixture components, T_{train} – training data length.

Table 1: Average identification accuracy [%] depending on the number of GMM components and the length of training data

T_{train},sec / N	30	60	90
8	74,3	79,4	81,2
16	80,9	87,7	89,4
32	84,2	92,4	93,6
64	79,9	92,2	94,8
128	77,8	92,5	95,6

As the Table 1 shows, there is a correlation between the length of the training sample and identification accuracy. Obviously, the longer is training data set, the better the correctness of identification is. Furthermore, it is noticeable that experimental results are better on a small training set using small number of Gaussian mixture components. The dependence is more straightforward when using larger training data lengths.

6. Conclusions

Speaker identification system based on Gaussian mixture model using simple energy-based voice activity detection was created. Identification accuracy is more than 95%. This accuracy is not high enough for a real-world application of such a system. It is planned to conduct a study of methods that could improve the accuracy of speaker identification and evaluate the impact of acquiring speech data from different channels on the accuracy of identification.

7. References

[1] V.N. Sorokin, V.V. Viugin, A.A. Tananykin Speaker identification: analytical review, *Information Processes Vol. 12, Num 1 (2012) 1-30.*

[2] J. Campbell, Speaker recognition: a tutorial, *Proc. IEEE Vol.85, N9 (1997) 1437–1462.*

[3] L. R. Rabiner, M. R. Sambur, An Alghorithm for Determining the Endpoints of Isolated Utterances, *The Bell System Technical Journal, Vol. 54, № 2, 1975.*

[4] S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *IEEE Trans. Acoustics, Speech, Signal Process v. 28 № 4 (1980) 357–366.*

[5] D. Jurafsky, J.H. Martin, *Speech and language processing, secon ed.*, Pearson Education, New Jersey, 2009.

[6] B. Atal, Automatic recognition of speakers from their voices, *Proc. IEEE v. 64 (1976) 460-475.*

[7] D. Reynolds, *Gaussian Mixture Models. Encyclopedia of Biometric Recognition*, Springer, Heidelberg (2008).

[8] D. Reynolds, R. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models, *IEEE Trans. on speech and audio processing v.3 №1 (1995) 72-83.*