Maximum Entropy based Data Selection for Speaker Recognition

Chien-Lin Huang[1,2] *and Bin Ma*[1]

[1]Human Language Technology Department, Institute for Infocomm Research, A*STAR, Singapore
[2]National Institute of Information and Communications Technology, Kyoto, 619-0288, Japan
{clhuang, mabin}@i2r.a-star.edu.sg

Abstract

This paper presents the data selection method for speaker recognition. Since there is no promise that more data guarantee better results, the way of data selection becomes important. In the GMM-UBM speaker recognition, the UBM is trained to represent the speaker-independent distribution of acoustic features while the GMM speaker model is tailored for a specific speaker. In this study of data selection for speaker recognition, we apply the maximum entropy criterion to remove the redundant feature frames in the UBM training and to select the discriminative feature frames in the GMM speaker modeling. The conducted experiments on the 2008 NIST Speaker Recognition Evaluation corpus show that the proposed method outperforms the baseline system without the data selection.

Index Terms: data selection, speaker recognition, maximum entropy

1. Introduction

Speaker recognition is to establish the identity of a person from his/her voice. In speaker recognition, signal processing and statistical modeling techniques are used to characterize a speaker, Gaussian mixture models (GMMs) is a popular approach to model the speaker. A GMM speaker model is typically based on the universal background model (UBM). In the GMM-UBM framework, one makes a speaker detection decision using the log-likelihood ratio between the target speaker and an universal background model [1].

Conventionally, speaker recognition systems prefer more enrollment and testing data for promising the performance. However, The redundant data may hurt the recognition accuracy and are computationally expensive. The way of effective data selection becomes important. There have been many successful attempts to address the problem of data selection. For example, the general case of feature selection based on mutual information criteria can be found in [2]. The segment selection technique was proposed to choose the segments with high target scores and low-variance imposter scores giving a good discrimination ability for recognizing speakers [3]. Recently, the inter-feature Euclidean distance based criterion is used to select feature frames across the speaker acoustic space for efficient universal background model training [4]. This feature frame sub-sampling method reduces the computational cost in UBM training and outperforms the conventional UBM training, which employs excessive amounts of data, in terms of equal error rate. In addition, the feature frame selection according to phonetic information [5] reaffirms that it is worthwhile looking into different ways of feature frame selection. In [5], the feature frames are chosen to have minimum redundancy within selected feature frames and maximum relevancy to speaker models as measured by mutual information.

We extend the previous study of the UBM data selection for the useful information extraction [6]. In this study, we propose a data selection method for speaker recognition by considering distinctive characteristics using the maximum entropy criterion. We conduct the experiments on NIST 2008 Speaker Recognition Evaluation. The results show that the proposed data frame selection is effective.

The rest of this study is organized as follows. Section 2 elucidates the proposed entropy based data selection for speaker recognition. We describe the experimental setup and report a series of experiments in Section 3. Finally, Section 4 concludes this work.

2. Maximum Entropy based Data Selection

Data selection is important in the statistical pattern recognition. In the GMM-UBM speaker recognition, the UBM is trained to represent the speaker-independent distribution of acoustic features while the GMM speaker model is tailored for a specific speaker. In this study of data selection for speaker recognition, we apply the maximum entropy criterion to remove the redundant feature frames in the UBM training and to select the discriminative feature frames in the GMM speaker modeling.

2.1. Information theory

In information theory [7], the information derivable from outcome x_t depends on its probability. A high probability indicates low information because the outcome is well expected. The amount of information, $I(x_t) = \log(1/P(x_t))$, represents uncertainty in the probabilistic framework. Here, X is a discrete random variable and from a finite set of observations x_t with $t = 1,...,T$. One of the most important properties of an information source is the entropy $H(X)$ of the random variable X, defined as the average information [8].

$$\begin{aligned} H(X) &= E[I(X)] \\ &= \sum_T P(x_t) I(x_t) = \sum_T P(x_t) \log \frac{1}{P(x_t)} \end{aligned} \quad (1)$$

This entropy $H(X)$ is the amount of information required to specify what kind of x_t has occurred on average. When all events are with equal probability, the entropy is maximal, because we are completely uncertain which event will occur.

2.2. Data selection for UBM training

We know that there is no promise that more training data guarantee a better result. As the UBM serves as a reference model in speaker recognition, the training data selection is critical to the speaker recognition performance. In this study, we select representative feature frames from speakers based on the maximum entropy criterion for UBM training. In order to have an effective and efficient UBM training, the selected feature frames should have a sufficient speaker coverage and

Copyright © 2011 ISCA

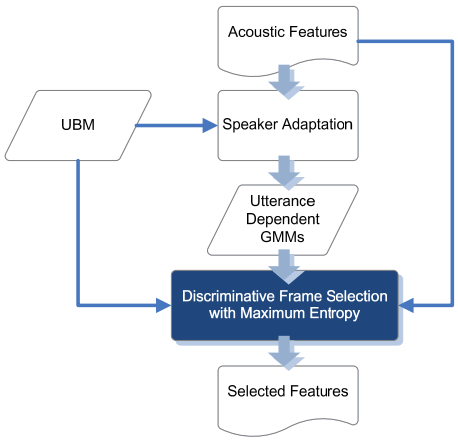

Figure 1: Flowchart of discriminative frame selection based the ME criterion.

have a minimum information redundancy as well.

Suppose there are *S* speech utterances in a speech database, each from a unique speaker *s*. To select the data from these *S* speakers, we have two considerations. One is to keep sufficient information from each of the speakers and another is to remove the redundant feature frames from those speaker who has relative larger number of feature frames. We define a compression ratio α which indicates the ratio of selected feature frames from each speaker. For the speakers who has only a limited training data available, we define a threshold θ which indicates the minimum number of feature frames for each speaker. So the number of selected feature frames from the speaker *s* is as follows:

$$\overline{T}_s = \begin{cases} T_s \times \alpha & if\ T_s > \theta \\ T_s & otherwise \end{cases} \quad (2)$$

where T_s is the original number of feature frames from speaker *s*. In such a way, the characteristics of all the available speakers are well covered through the selection.

To make the data selection from the T_s feature frames of the speaker *s* for the UBM training, the maximum entropy criterion shown in Eq. (1) is applied to select the $T_s \times \alpha$ feature frames which are of the highest estimated entropy against the speaker model. The following objective function is formulated for the selection:

$$H(P(X),\lambda) = \sum_t -P_\lambda(x_t)\log P_\lambda(x_t) \quad (3)$$

where the Gaussian mixture models λ is used to represent the parametric probability densities of the speaker characteristics. For the speaker *s*, the score function $P_\lambda(x_t)$ is estimated by the likelihood of the feature frame x^b given by the speaker-dependent GMM λ_s. The feature frame selection is conducted for every *B* frames, *1<t<B,* while we set *B* to 16 in this study. The likelihood $p(x^b \mid \lambda_s)$ is computed with sum of Gaussian mixtures as follows:

$$p(x_t \mid \lambda_s) = \sum_{m=1}^{M} w_m p(x_t \mid m),\ \ p(x_t \mid m) = N(x_t; \mu_m, \sigma_m) \quad (4)$$

where the normal distribution $N(\cdot;\mu_m,\sigma_m)$ is used with the mean μ_m, diagonal covariance matrix σ_m, and mixture weight w_m. A small number of Gaussian mixtures (*M*=16) is used for the robust estimation of the speaker GMM.

2.3. Discriminative feature selection

With the trained UBM model, we would like to conduct a discriminative feature frame selection for the three datasets of speaker recognition system, the enrollment, T-norm and testing datasets. In speaker recognition, we prefer that the selected feature frames are representative within the speaker and discriminative between speakers. The discriminative feature selection process is based on the maximum entropy criterion shown in Fig. 1. In this study, we use the eigen-channel adaptation [9] for speaker adaptation. A speech utterance of the speaker *s* is adapted from an universal background model λ_{ubm} and represented by the speaker-dependent GMM λ_s. We take into consideration of the discrimination between the target speaker and non-target speakers. The score function $P_\lambda(x_t)$ in Eq. (3) is reformulated as follows:

$$P_\lambda(x_t) = \frac{p(x_t \mid \lambda_{ubm})}{p(x_t \mid \lambda_s)} \quad (5)$$

The purpose of the estimated probability density for speech segments attempts to maximize target speaker characteristics $p(x_t \mid \lambda_s)$ and minimize others $p(x_t \mid \lambda_{ubm})$. The discriminative frames are selected based on the maximum entropy criterion and a threshold θ_{spk} is used to determine the number of feature frames is selected. In other words, θ_{spk} indicates a ceiling which is an upper bound of number of selected feature frames.

3. Experiments

The NIST SRE-2004 one-side data were used to train the gender-independent UBM with 512 Gaussian mixture. The eigen-channel compensation was adopted for the GMM-UBM speaker recognition and the number of channel factors was set to 30. The fast-scoring technique [10] was applied by approximating the likelihood values using only the top 5 mixture components.

3.1. Baseline systems

The Long-Term Feature analysis (LFT) [11] was used to extract the acoustic feature frames of a speech utterance. The LTF analysis was as follows: The conventional short-time feature, Mel-frequency cepstral coefficients (MFCC), was estimated and followed by an auto-regression moving average filter to smooth spikes in the time sequence thereby to reduce noise. The short-time frequency analysis of 16 ms with the long-time window of 6 short-time cepstral frames was used for the long-term feature analysis. The LTF feature vector was composed by 12 coefficients plus their first and second order derivatives for a total size of 36 components. Then, the cepstral mean subtraction and cepstral variance normalization were applied for slowly varying convolution noises. Moreover, feature warping was used to reduce the additive noise and channel effects and to map a feature stream to a standard normal distribution. LFT showed better performances in [11].

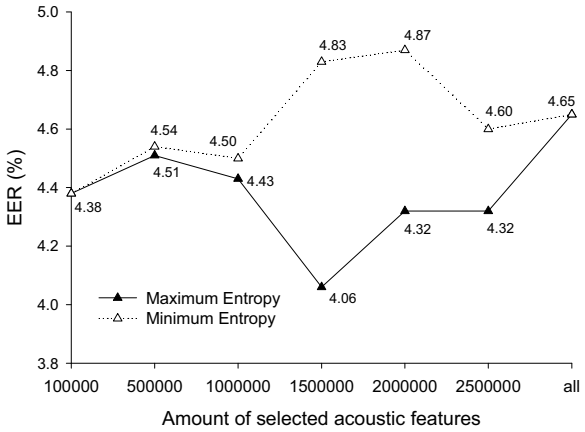

Figure 3: UBM performance with different numbers of selected feature frames from NIST SRE-2004 dataset.

3.2. Evaluation metrics

Two types of errors, false acceptance and false rejection, occurred in speaker recognition. In essence, the equal error rate (EER) reports the system performance when the false acceptance and false rejection rates were equal. The minimum normalized detection cost function (DCF) was a weighed sum of miss detection and false alarm rates defined in the NIST Speaker Recognition Evaluation (SRE) [12] shown as follows:

$$\begin{aligned} DCF &= C_{Miss} \times P_{Miss|\text{Target}} \times P_{\text{Target}} \\ &+ C_{FalseAlarm} \times P_{FalseAlarm|\text{NonTarget}} \times (1 - P_{\text{Target}}) \end{aligned} \quad (6)$$

where $C_{Miss} = 10$, $P_{Target} = 1$ and $C_{FalseAlarm} = 0.01$. Both EER and DCF were reported in this study.

3.3. Data selection with max and min entropy

To evaluate of the proposed data selection based on maximum entropy criterion, the various amounts of feature frames were selected for training the background models shown in Fig. 3. The universal background models were trained with the data selection based on the maximum and minimum entropy to show the effectiveness of the entropy criterion. It is confirmed that the maximum entropy scheme evidently outperformed the minimum entropy scheme. The experimental results in Fig. 3 show that there may not be sufficient training data to significantly discriminate the performance of maximum and minimum entropy when the numbers of selected feature frames vary from 100k to 1000k. However, we can find a significant gain with the data selection when the numbers of selected feature frames exceed 1000k. With the 1500K of selected feature frames, the proposed maximum-entropy data selection scheme achieved 12.69% relative EER reduction (from 4.65% to 4.06%). The results also show that it is not effective to apply all the data for UBM training. The proposed maximum-entropy data selection scheme provided a good representative information and effective training.

3.4. Comparison of different background datasets

In order to consider different background datasets, the UBM was performed using the 1conv4w condition of the NIST SRE-2004, SRE-2005 and SRE-2006 databases. The number of speakers from each of these data sources was summarized in Table 1. Note that each speaker was represented as one speech utterance. The results of different background dataset configurations were detailed in Table 2. Although different, combinations of the datasets have been tried, the UBM training using the NIST SRE-2004 dataset gives the best overall performance.

Table 1: Number of speakers in different background datasets.

Gender	NIST04	NIST05	NIST06	NIST08
Female	368	372	462	795
Male	248	274	354	470
Total	616	646	816	1265

Table 2: Results of the proposed data selection for UBM training using different background datasets (in %).

Dataset	Female		Male		All	
	EER	100xDCF	EER	100xDCF	EER	100xDCF
NIST04	4.30	2.07	3.84	1.66	**4.06**	**1.92**
NIST05	4.62	2.20	3.61	1.72	4.48	2.11
NIST06	4.32	1.96	3.81	1.49	4.11	**1.90**
NIST04+05	4.62	2.04	4.05	1.62	4.42	1.96
NIST04+05+06	4.78	1.96	3.59	1.63	4.48	1.96

3.5. Discriminative frame selection

To evaluate the discriminative frame selection, the experiment was conducted on the conversational telephone English speech of NIST SRE-2008. The eigen-channel compensation was trained by the NIST SRE-2004 telephone data. In the NIST evaluation, one two-channel telephone conversation, of approximately 5 minutes total duration, with the target speaker channel designated [12] was used for the enrollment and testing. Ideally, there is a speech utterance of about 2.5 minutes available for each speaker in the 5 minutes telephone conversation, but the durations of evaluation data vary much shown in Fig. 4. The histogram of feature frames was illustrated for a speech utterance of the NIST SRE-2008 core test telephone condition, after voice activity detection (VAD). All speech utterances were divided into fifty scales. The x-axis denoted the number of feature frames in a speech utterance and the y-axis was the number of corresponding speech utterances. The minimum and maximum of feature frame were 191 and 9256, respectively. The different threshold θ_{spk} was applied for the discriminative frame selection for each of the speakers.. Table 3 showed the speaker recognition results, EER and DCF, on the NIST SRE-2008 core test telephone condition and all English trials. The EER was reduced when θ_{spk} was from 6000 to 4000 for each speaker. In Fig. 4, the black bars denoted those speech utterances with speech frames more than 4500. The 13.68% frame reduction was obtained when we set $\theta_{spk} = 4500$. The relative reductions of EER and DCF were 4.69% and 6%, respectively. Note that 4500 feature frames corresponded to 2.4 minutes of speech.

3.6. Evaluation for the long-long condition

In order to understand the performance of the discriminative frame selection in the longer speech utterances experiments were conducted on the NIST SRE-2008 long-long evaluation condition, which was with a single channel microphone recorded conversational segment of eight minutes or longer duration for enrollment and testing. In addition, the eigen-channel compensation was trained by the NIST SRE-2005 and

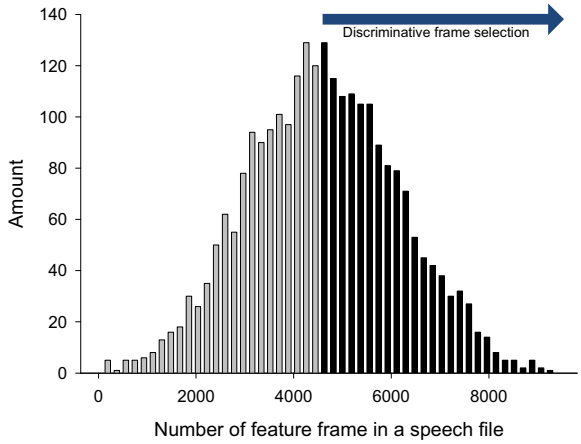

Figure 4: Accumulated histogram of feature frames after VAD on NIST SRE-2008 core test telephone condition.

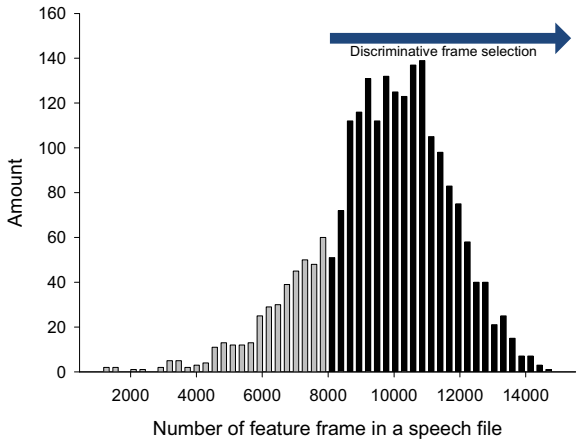

Figure 5: Accumulated histogram of feature frames after VAD on NIST SRE-2008 long-long condition.

Table 3: Results with different frame ceiling on NIST SRE-2008 core test telephone condition, all English trials (in %).

θ_{spk}	baseline	6000	5000	**4500**	4000	3000
EER	4.05	4.00	3.89	**3.86**	4.05	4.21
100xDCF	1.73	1.70	1.67	**1.63**	1.68	1.86
Frame Reduction	-	3.12	8.89	**13.68**	19.91	36.16
EER Reduction	-	1.23	3.95	**4.69**	0.0	-3.95
DCF Reduction	-	1.97	3.70	**6.00**	3.12	-7.26

Table 4: Results with different frame ceiling on NIST SRE-2008 long-long condition (in %).

θ_{spk}	baseline	12000	10000	**8000**	6000	4000
EER	7.88	7.70	7.64	**7.74**	7.82	8.47
100xDCF	3.24	3.17	3.15	**3.16**	3.34	3.62
Frame Reduction	-	0.84	6.45	**20.18**	38.62	58.78
EER Reduction	-	2.28	3.05	**1.78**	0.76	-7.49
DCF Reduction	-	2.10	2.72	**2.41**	-3.15	-11.8

SRE-2006 microphone data. The results of the discriminative frame selection with different thresholds θ_{spk} on the long-long condition were summarized in Table 4. Figure 5 is the histogram of feature frames after VAD for each speaker while the minimum and maximum of feature frames were 1267 and 14693, respectively. The black bars denoted those speech utterances with the speech frames more than 8000, about 4.28 minutes of speech. By setting the threshold $\theta_{spk} = 8000$, we obtained a substantial feature frame reduction (20.18%) as well as 1.78% EER reduction and 2.41% DCF reduction. Table 3 and Table 4 show that the proposed discriminative feature frame selection can reduce the redundant feature frames and accomplish an effective and efficient speaker recognition.

4. Conclusions

In this study, the data selection approach based on maximum entropy criterion is proposed for efficient UBM training and effective speaker recognition. The proposed method selects feature frames across the speaker acoustic space using the maximum entropy criterion. The selected training data minimizes the redundancy and is a representation that allows a useful information extraction from a source. Experiments were conducted on NIST SRE-2008. Based on experiments, we confirm that the proposed maximum entropy based data selection is helpful in UBM training and improves the speaker recognition performance. An EER reduction of 12.69% has been achieved using the data selection for UBM training. The discriminative frame selection also shows a substantial performance improvement both in EER and frame reduction.

5. References

[1] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I.Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovsk-Delacrtaz, and D. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP J. Appl. Signal Processing*, vol. 4, pp. 430–451, 2004.

[2] H. C. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min- redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol 27, pp. 1226–1238, 2005.

[3] N K. P. Li and J. E. Porter, "Normalizations and Selection of Speech Segments for Speaker Recognition Scoring," in *P Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, pp. 595–598, 1988.

[4] T. Hasan, Y. Lei, A. Chandrasekaran and J. H. L. Hansen, "A novel feature sub-sampling method for efficent universal background model training in speaker verfification," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, pp. 4494–4497, 2010.

[5] C.-S. Jung, M. Y. Kim, H.-G. Kang, "Selecting feature frames for automatic speaker recognition using mutual information," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 18, no. 6, 2010.

[6] C.-L. Huang and H. Li, "UBM Data Selection for Effective Speaker Modeling," in *Proc. ISCSLP*, Tainan, Taiwan, 2010.

[7] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, p. 379–423/623–656, 1948.

[8] X. Huang, A. Acero, H.-W. Hon, "Spoken Language Processing: A Guide to Theory, Algorithm and System Development," Prentice Hall, 2001.

[9] P. Kenny, G. Boulianne, P. Ouellet and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[10] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[11] C.-L. Huang, H. Su, B. Ma, H. Li, "Speaker Characterization Using Long-Term and Temporal Information," in *Proc. Interspeech*, pp. 370–373, Makuhari, Japan, 2010.

[12] NIST SRE [Online]: http://www.nist.gov/speech/tests/spk/