i Vector used in Speaker Identification by Dimension Compactness

Soumen Kanrar
Department of Computer Science
Vidyasagar University
West Bengal, India
rscs_soumen@mail.vidyasagar.ac.in

ABSTRACT

The automatic speaker identification procedure is used to extract features that help to identify the components of the acoustic signal by discarding all the other stuff like background noise, emotion, hesitation, etc. The acoustic signal is generated by a human that is filtered by the shape of the vocal tract, including tongue, teeth, etc. The shape of the vocal tract determines and produced, what signal comes out in real time. The analytically develops shape of the vocal tract, which exhibits envelop for the short time power spectrum. The ASR needs efficient way of extracting features from the acoustic signal that is used effectively to makes the shape of the individual vocal tract. To identify any acoustic signal in the large collection of acoustic signal i.e. corpora, it needs dimension compactness of total variability space by using the GMM mean supervector. This work presents the efficient way to implement dimension compactness in total variability space and using cosine distance scoring to predict a fast output score for small size utterance.

Categories and Subject Descriptors

H 1.1 [Information System]: Models and Principles – System and Information Theory

General Terms

Algorithms, Design, Management, Measurement, Verification.

Keywords

Dimension Compactness, Feature Vector, Spectral Analysis, Supervector, Cepstral coefficient, Cosine scoring.

1. INTRODUCTION

In current technology of an automatic speech recognition (ASR) system produce good prediction in controlled environment i.e. the collected samples of utterance from the clean environment. The most implementable areas of ASR system are more noise full environments such as target advertising, forensic science, and service customization. The accuracy among the previously implemented ASR system based on pure GMM (Gaussian mixture model) or HMM (Hidden Markov model) [9, 10, 11, 12, 15] brings poor performance, particularly in noise full environment. Another major drawback in old ASR procedure consumes large computation time. Number of methods has been

proposed over the last decade to give an efficient technique to solve this issue in ASR system or even in the other Avenue like multimedia streaming [16, 17], but still remain it is challenging one. The previous method used EM (expectation maximization) module that consumes a lot of computation time during the final prediction about the unknown utterance. The current start of art, dimension compactness successfully reduces the computation time and efficiently works in noise full environments. In old ASR system also used the score normalization for the predicted score, based on the log(likelihood) ratio test [8]. The proposed work used the cosine kernel to predict the closeness among the test utterance and target utterance that gives very fast and efficient prediction in compare to old ASR system. Najim Dehak et.al., proposed new modeling about the low dimensional speaker and channels dependent space [1]. Deep Neural Networks technique being proposed by P.Kenny et.al.,[2] for extracting statistics for speaker recognition. The environmental sound classifications are based on acoustic feature extraction been proposed by Takumi [3]. The delta spectral coefficient proposed by kshitiz kumar et .al., for robust speech recognition, [4]. Maximum likelihood estimates of the supervector covariance matrix that effectively extended speaker adaption for Eigen voice estimation [5]. The accent recognition by i-vector based on Gaussian means supervector improved the performance of ASR system [6]. The cosine based on distance scoring is currently proposed to improve the computation time for predict the existence of utterance in a collected sample [7]. This paper is structured as follows. Section I introduces about the problem. Section II presents the spectral analysis. Section III presents the dimension compactness of total variability space. Section IV presents the cosine based distance scoring between the test utterance and target utterance in the corpora of acoustic signal with the conclusion at the end.

2. SPECTRAL ANALYSIS

Speech waves are band limited to 4kHz, 8kHz and 16kHz respectively and sampled at 8kHz to 32kHz and windowed by the Hamming window of 20ms long with the 10ms shift.

As the audio signal is constantly changing, so it is required to simplification, in this regards we expect that on short time, audio signal doesn't change statistically. Here we can assume. It is statistically stationary, but obviously the samples are constantly changing on even in the short time scale. It is very much required that frame length of the signal should be kept optimized. If the frame size is much shorter then we don't have sufficient samples to get a reliable spectral estimation. On the other hand, if the frame size is longer than the signal changes too much throughout the signal. A pre emphasis filter of the form $H(z) = 1 - (0.97)z^{-1}$ is applied first. The FFT (Fast Fourier Transform) analysis is performed using Hanning windows of duration 20ms, with the 10ms shift between frames for a sampling frequency of 16 kHz. A band pass filter of 40 gamma tone of channel is considered in the filter bank. A band pass filter is used in auditory modeling to approximate the frequency for small selected portion. The linear predictive coefficient and the estimated energy are transformed to LPC cepstrum and logarithmic energy respectively. The LPC cepstrum coefficients are depended on the peak-weighted. The smoothed spectral envelop is obtained by Fourier transformation. The time functions of the LPC (linear predictive coding) cepstrum coefficients are called as cepstrum coefficient is used in the feature extraction. Therefore, the logarithmic spectral envelop, which corresponds to the dynamics emphasized cepstrum be obtained as follows.

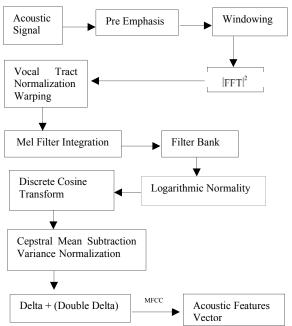


Figure 1. Feature Extraction

The k-dimension cepstrum vector, each of which consists of the 1^{st} through the k^{th} cepstrum coefficients for the adjacent n – frames are denoted by C_J for $(J \ 1, \dots, n)$. The 1^{st} and 2^{nd} order polynomial expansion coefficients with their time functions are considered. If S(w,t) and

 $C_{[m]}(t)$ are the power spectral envelop and the m^{th} order cepstrum coefficient at time t then it can be expressed as

$$log(S(w,t))$$
 $\sum_{-k}^{k} C_{[m]}(t)e^{-jmw}$. The data Cepstral features

capture dynamic speech information that improves the speech recognition procedure. The Delta Cepstral and double-Delta Cepstral coefficients are appended to 13 MFCC features and 26 Delta- Cepstral coefficients. The Cepstral sequence is $C_{\rm [m]}$ for short time interval. The

Delta-Cepstral features are defined as,

$$\Delta_{\lceil m \rceil} = C_{\lceil m+s \rceil} - C_{\lceil m-s \rceil} \tag{1}$$

The index m is related to considered frame for analysis and s is an integer constant. The symbol Δ operator is stood for Delta operation. The double –Delta Cepstral features are obtained by the Delta operation on the expression (1) again. The acoustic feature vectors are obtained according to the procedure presented in figure 1.

3. DIMENSION COMPACTNESS OF TOTAL VARIABLITY SPACE

Joint factor Analysis (JFA) considers the K-number of acoustic feature vectors that is obtained according to the flow diagram 1. The JFA is implemented on the m probabilistic pattern of the linear Gaussian mean supervector. Here K is the number of observed feature vectors obtained according to the flow diagram 1, these K vectors are independent and identically distributed.

Now, $Q \supset \{q_t\}_{t=1}^K$ with $q_t \in R^F$, Q is the collection of all possible acoustic features and F is the dimension of the acoustic class. The observed vector q_t is presented by L-components soft Gaussian Mixture model (GMM), for the utterance model $\lambda = (\{\omega_i\}, \{\theta_i\}, \{\sum_i\})$.

The general d – variate GMM is expressed as,

$$p_{\lambda}\left(\boldsymbol{q}_{t}\mid\left\{\boldsymbol{\theta}_{i}\right\}\right)\!=\!$$

$$\sum_{i=1}^{L} \omega_{i} \frac{1}{(2\pi)^{d/2} \left| \sum_{i} \ \right|^{1/2}} exp \left\{ \frac{1}{2} \Big(\big(q_{t} - \theta_{i} \, \big)^{\! /} \sum_{L}^{-1} \! \big(q_{t} - \theta_{i} \, \big) \! \right) \right\}$$

The means θ_i for i 1,2,...,L are random vector

$$\theta_i \in R^F$$
, with associated weight, $\omega_i \in R$ and $\sum_{i=1}^L \omega_i = 1$

with covariance is $\sum_{i} \in R^{F \times F}$ for i = 1, 2, 3, ...L.

The Universal Background Model (UBM), which is a big size GMM, is built with likelihood function,

$$p(Q \mid \lambda) = \sum_{j=1}^{J} \omega_j p_{\lambda}(q_t \mid \theta_i, \sum_j), q_t \text{ is the acoustic vector at}$$

time t, and ω_j is the mixture weight for the j^{th} mixture component. Here, $p_{\lambda}(q_t\mid\theta_i,\sum_i)$ is a Gaussian probability

function with mean θ_i and covariance matrix is \sum_j . Now, J is the total number of Gaussian present in the in the mixture UBM [14] and optimized value of J is the number $(1024 \times i)$, for i=1,2,3,...

To adopt the Gaussian means of UBM, we consider 'Maximum A Posterior' (MAP) method according to the flow diagram 2. The mean supervector is constructed by appending together with the means of each mixture component. It is expressed as $\theta = \left\{\theta_1^{\prime}, \theta_2^{\prime}, \cdots, \theta_L^{\prime}\right\}^{\prime} \in R^{F^*L} \text{, the notation '/'} \text{ is the transpose of vector and } E(\theta) \approx m.$

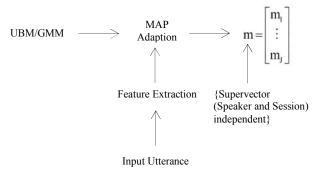


Figure 2. Gaussian means Supervector

The total variability modeling assumes the GMM mean super vector is M . It is optimized and presented by a set of feature vectors with the expression

$$M = m + \phi \tag{2}$$

Here, m is the speaker and session independent supervector from the universal background model (UBM), \$\phi\$ is an additive function of the speaker subspace and channel subspace. The JFA helps to reduce the subspace size. Prediction about the matching of input utterance based on "dimension compactness" of total variability space for the acoustic features are consider as the cosine value between the test utterance and target utterance. In this regards the support vector machine wouldn't much effective of enhancement for speech identification. The additive space is considered as a single space. This single space is considered as the total variability space. The variability space consists of the speaker variability and channel variability. The total variability space is presented by a total variability matrix T(say). The total variability matrix is obtained from the collection of Eigen vectors for the corresponding larger Eigen values of the total variability covariance matrix of acoustic feature data-set. The data-set is obtained from the acoustic feature class data Q. The variability matrix T is described as

$$T \begin{bmatrix} \sum X_{1}^{2} / N & \sum X_{1} X_{2} / N & \cdots & \sum X_{1} X_{c} / N \\ \sum X_{2} X_{1} / N & \sum X_{2}^{2} / N & \cdots & \sum X_{2} X_{c} / N \\ \vdots & \vdots & \vdots & \vdots \\ \sum X_{c} X_{1} / N & \sum X_{c} X_{2} / N & \cdots & \sum X_{c}^{2} / N \end{bmatrix}$$

Here, $\sum X_i^2/N$ is the variance for the i^{th} component in the low dimensional total variability space and $\sum X_i X_j/N$ is the covariance of the i^{th} and j^{th} component of the total variability space, for $c, N \in I^{\succ 0}$.

Now the expression (2) is modified to expression (3).

$$M = m + T\omega + \varepsilon \tag{3}$$

T is the low dimension square matrix as it is obtained from the selected Eigan vectors of the corresponding larger Eigen values of the total variability space. The Eigan vector gives the direction along the maximum variability in low dimensional space. The matrix T having the low ranks and ω is the random vector such that $\omega = \{\omega_i\}_{i=1}^c$, ω follows the standard normal distribution. The components associated with each vector ω_i are the feature factors collected from the matrix T. Those are collected according to the flow diagram 1. Each vector of $\{\omega_i\}_{i=1}^c$ follows identical type distribution, and hence it is called 'i-vector'. M is normally distributed with mean vector $\mathbf{m} \{\mathbf{m}_1, \dots \mathbf{m}_1\}$, M is session and channel dependent supervector with square covariance matrix $\sum TT'$. The '/' notation presents the transpose, and the notation ε is the residual noise such that ε is normally distributed with standard deviation \sum . The impact of noise is reduced on the total variability compact space. Hence, the equation (3) is redefined as

$$M \approx m + T\omega$$
 (4)

4. PREDICTED COSINE SCORE MEASUREMENT

The i-vector extraction is based on factor analysis. It is extended the session and speaker variabilities of supervector to Joint Factor Analysis (JFA) [1, 13, 14]. The extracted 'i-vector' simultaneously and efficiently capture the speaker and channel variabilities. To identify any utterance in a target list, the cosine kernel based predicated procedure is implemented, between the test speaker 'i-vector' and target speaker'i-vector'. This procedure produces more optimized results. The cosine mapping based scoring procedure is effectively implemented as follows. Consider, two distinct speaker's utterances are (x) and (y) respectively. Those utterances are approximated by two distinct multinomial polynomials

based on extracted acoustic features [7], say $P_x(i)$ and $P_y(i)$. Each multinomial polynomial consists of K number of features and each feature has at most F dimensions. The derived probabilities of feature are considered as $p_x(i-1)$,..., $p_x(i-K)$ and $p_y(i-1)$,..., $p_y(i-K)$ for two utterances. Since, $p_x(i)$ and $p_y(i)$ are the probability distributions for the corresponding utterances, clearly $\sum_{i=1}^K p_x(i) = \sum_{i=1}^K p_y(i) - 1$. The divergence measurement between multinomial polynomials in the space l(p) is defined a mapping $d: l(p) \times l(p) \rightarrow R^+$.

The mapping is expressed by

$$d\left(P_{x}, P_{y}\right) \quad \cos^{-1}\left\{\left(\sum_{i} p_{x}(i) p_{y}(i)\right)^{q_{k}}\right\}^{1/S} \tag{5}$$

Let $\{p_k\}$ is bounded sequence of strictly positive real number and present the derived probability of features. For the two separate utterances we consider k=x,y.

As, $\sup\{p_k\}$ 1, since q_k is the derived probability So, $S = \max(1, H)$ and $0 \prec p_k \le \sup\{p_k\} = H$.

For simplicity, we consider $q_k = \frac{1}{2}$ and S = 1.

Now,

i)
$$d(P_x, P_x) = \cos^{-1} \left\{ \left(\sum_i p_x(i) p_x(i) \right)^{1/2} \right\} = \cos^{-1} \left\{ 1 \right\} = 0$$

ii)
$$d(P_x, P_y) = 0 \Leftrightarrow x = y$$

iii)
$$d(P_x, P_y) = d(P_y, P_x)$$

iv) According to figure 3.

We get
$$d(P_x, P_z) \prec (d(P_x, P_y) + d(P_y, P_z))$$

According to figure 4.

We get
$$d(P_x, P_z) = (d(P_x, P_y) + d(P_y, P_z))$$

The above two expression converges to

$$d(P_x, P_z) \le \left(d(P_x, P_y) + d(P_y, P_z)\right)$$

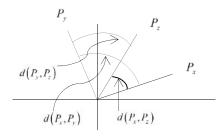


Figure 3. Divergence measurement for 1st orientation

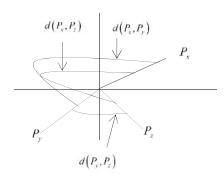


Figure 4. Divergence measurement for 2nd orientation

So, l(p) is a metric space with metric d, (divergence measurement) that is expressed by the expression (5). According to the equation (4), the total variability modeling considers the GMM mean super vector is M. Those are the conversation side supervector, and it is clearly speaker and session dependent. The set of feature vector is decomposed according to expression (4). The mean supervectors from UBM are {m} and those are speaker and session independent. The total variability matrix T spans a low dimensional subspace. Now, ω are the factors describing the utterance-dependent mean offset Tω. The set of low dimensional total variability factors ω presents each conversation side. Each factor controls the separate Eigen dimension of the total variability matrix (T). The distance scoring on channel compensated, 'i-vector' i.e. ω for a pair of conversation sides is the cosine between the target speaker, i-vector and the test speaker, i-vector in l(p)metric space. The predicated score is obtained based on the cosine based prediction algorithm.

Algorithm: Cosine based Prediction

Input: $W_{t \operatorname{arg} et} W_{test}$

Var A: real

Compute:

$$A \quad \cos^{-1}\left(\frac{\left\langle w_{target}, w_{test} \right\rangle}{\parallel w_{target} \parallel * \parallel w_{test} \parallel}\right)$$

$$if\left(0 \le A \le \frac{\pi}{2}\right) then$$

score cos(A);

elseif
$$\left(\frac{\pi}{2} \prec A \leq \frac{3\pi}{2}\right)$$
 then

score 0;

else

 $score = cos(2\pi - A);$

Output: Score

The acceptance or rejection based on user controlled decision threshold angle.

5. RESULT AND DISCUSSION

The above methodology being tested in the collected utterances from the noise full environment of telephone captured signal, mobile captured recording in rail station, busy bus stoppage from nonnative English spoke person. The corpus is the collected utterance from Native Indian languages likes Hindi, Bengali, Teague, and Oriya. In the testing purpose, we have randomly used English utterance as monologue or in the conversation even as a mixed spoke person. The used test utterance is of 45-second duration, and we consider the target list of 30 people. In this regards, the considered i-vector with dimensions is 400 and feature dimension is equal to the 39 MFCC features. The tested results are presented in figure 5, figure 6 and figure 7. The figure 5 presents predicted cosine score about the probable matches for the speaker 1 speaker 2 and speaker 3. The horizontal axis presents the model identifier number that is already in the target list. The vertical axis presents the predicated cosine score between the test and target utterance in the range of [0, 1] with the scale 0.01. The score more than 0.9 to be considered as a good match with considering the false accept, and below 0.8 are not matched with considering false reject.

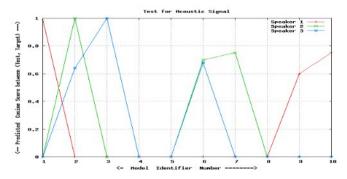


Figure 5. Voice test score for speaker 1, 2, 3

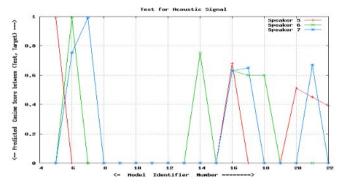


Figure 6. Voice test score for speaker 5, 6, 7

According to figure 7 if we consider 0.8 is the threshold level for acceptance, then the utterance of the speaker 9 that matches with the model identifier of the speaker 9. But the utterance of speaker-ID 9 is also matched to the model identifiers 11,25,26,27. Clearly, these 4 speaker models are falsely accepted. The equal error rates among the false accept and false reject controlled by the decision

threshold. The decision threshold is user choice. It is depended upon the environment from where the utterance is collected.

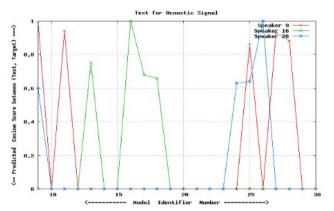


Figure 7. Voice test score for speaker 9, 16, 26

According to the figure 5, 6 and 7 if we consider the threshold is 0.9, then out of 30; only four speakers are falsely accepted i.e. 13%. If we consider a threshold is 0.8, then out of 30 speakers, only six speakers are falsely accepted i.e. 20%. If we consider the threshold is 1.0, then two speakers are falsely accepted i.e. 7%.

6. CONCLUDING REMARKS

This work presents the impact of dimension compactness in total variability space. The proposed methodology sufficiently reduces the computation time and works for small size of test utterance. The cosine scoring provides fast predicts about the matching. One of the most achievements is that if the predicted score is 1.0, it is the highly perfect match with little false acceptance by considering the noise full medium. The scoring particularly depends in which noise full environment the utterance being collected. The false accept and false reject depends on the decision threshold, but that could be controlled according to noise level of the environment. This methodology sufficiently reduces the false reject. For the specific suspicious target speaker, the proposed methodology enhanced the ASR system.

7. ACKNOWLEDGMENTS

Author would like to thank Niranjan Kumar Mandal for his continuous encouragement and motivation.

8. REFERENCES

- Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P. 2001. Front-end factor analysis for speaker verification. *IEEE Trans. Audio Speech Lang. Process.* 19, 788–798.
- [2] Kenny, P., Gupta, V., Stafylakis, T., Ouellet, P and J. Alam. 2014. Deep Neural Networks for extracting Baum-Welch statistics for Speaker. *In Proc. of Odyssey-2014*.
- [3] Kobayashi, T., Ye, J. 2014. Acoustic feature extraction by statistics based local binary pattern for environmental sound classification. *IEEE International Conference on Acoustics*, *Speech and Signal Processing (ICASSP-2014)*, 3052-3056, doi: 10.1109/ICASSP.2014.6854161.
- [4] Kumar, K., Kim, C., Stern, R. 2011. Delta-spectral cepstral coefficients for robust speech recognition. *IEEE*

- International Conference on Acoustics, Speech and Signal Processing (ICASSP- 2011). 4784-4787, doi: 10.1109/ICASSP.2011.5947425, 2011.
- [5] Kenney, P., Boulianne, G., Dumouchel, P. 2005. Eigen voice modeling with sparse training data. *IEEE transaction on* speech and audio processing. Vol 13, No 3, 345 – 354, doi:10.1109/TSA.2004.840940.
- [6] Bahari, M., Saeid, R., Hamme, H., Leeuwen, D. 2013. Accent recognition using i-vector, Gaussian Mean Supervector and Gaussian posterior probability supervector for spontaneous telephone speech. *IEEE International Conference on Acoustics, Speech and Signal Processing*. 7344 – 7348, doi: 10.1109/ICASSP.2013.6639089.
- [7] Kanrar, S., et al. 2015. Text and Language Independent Speaker Identification by GMM based i Vector. *Proceedings* of the Sixth International Conference on Computer and Communication Technology ICCCT '15. 95-100.
- [8] Kanrar, S. 2015.Impact of Threshold to Identify Vocal Tract. Advances in Intelligent Systems and Computing .Vol 404, 97-105, doi: 10.1007/978-81-322-2695-6 9, 2015.
- [9] Reynolds, D., Quatieri, T., Dunn, R. 2000. Speaker verification using adapted Gaussian mixture models. *Digital Signal processing*, Vol. 10, 19-41.
- [10] Xiang, B., et.al. 2002. Short-time gaussianization for robust speaker verification. *In Proc. ICASSP*, Vol. 1, 9681-684.
- [11] Reynolds, D. 1995. Automatic speaker recognition using Gaussian mixture speaker models. *Lincoln Laboratory J.* 8(2), 173-191.
- [12] Reynolds, D. 2003. Channel robust speaker verification via feature mapping. Proc, Internet. Conf. Acoustics Speech Signal Process, 53-56.
- [13] Martinez, D., Plchot, O., Burget, L., Glembek, O., Matejka, P. 2011. Language recognition in i vectors space. In: Annual Conference of the International Speech Communication Association (Inter speech), 861–864.
- [14] Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P. 2007. Joint factor analysis versus eigen channels in speaker recognition. *IEEE Trans. Audio, Speech, Lang Process.* vol. 15, no. 4, 1435–1447.
- [15] Kanrar S., et al. 2015. Detect Mimicry by speaker Recognition system. *Adv. Intell. Syst. Comput.* 339, 21–31. doi:10.1007/978-81-322-2250-7 3.
- [16] Kanrar S., et al. 2016. E-health monitoring system enhancement with Gaussian mixture model. *Multimedia Tools and Applications, Springer US*. 1-23, doi:10.1007/s11042-016-3509-9.
- [17] Kanrar S., et al. 2016. Video traffic analytics for large scale surveillance. *Multimedia Tools and Applications, Springer* US. 1-28, doi: 10.1007/s11042-016-3752-0.