

Enhanced VQ-based Algorithms for Speech Independent Speaker Identification

Ningping Fan and Justinian Rosca

Siemens Corporate Research Inc., 755 College Road East,
Princeton, New Jersey 08540
{NingPing.Fan, Justinian.Rosca}@scr.siemens.com

Abstract. Weighted distance measure and discriminative training are two different approaches to enhance VQ-based solutions for speaker identification. To account for varying importance of the LPC coefficients in SV, the so-called *partition normalized distance measure* successfully used normalized feature components. This paper introduces an alternative, called heuristic weighted distance, to lift up higher order MFCC feature vector components using a linear formula. Then it proposes two new algorithms combining the heuristic weighting and the partition normalized distance measure with group vector quantization discriminative training to take advantage of both approaches. Experiments using the TIMIT corpus suggest that the new combined approach is superior to current VQ-based solutions (50% error reduction). It also outperforms the Gaussian Mixture Model using the Wavelet features tested in a similar setting.

1 Introduction

Vector quantization (VQ) based classification algorithms play an important role in speech independent speaker identification (SI) systems. Although in baseline form, the VQ-based solution is less accurate than the Gaussian Mixture Model (GMM) [1], it offers simplicity in computation. For a large database of over hundreds or thousands of speakers, both accuracy and speed are important issues. Here we discuss VQ enhancements aimed at accuracy and fast computation.

1.1 VQ Based Speaker Identification System

Fig. 1 shows the VQ based speaker identification system. It contains an offline training sub-system to produce VQ codebooks and an online testing sub-system to generate identification decision. Both sub-systems contain a preprocessing or feature extraction module to convert an audio utterance into a set of feature vectors. Features of interest in the recent literatures include the Mel-frequency cepstral coefficients (MFCC) [2], the Line spectra pairs (LSP) [3], the Wavelet packet parameter (WPP) [4], or PCA and ICA features [5], [6]. Although the WPP and ICA have been shown to offer advan-

tages, we used MFCC in this paper to focus our attention on other modules of the system.

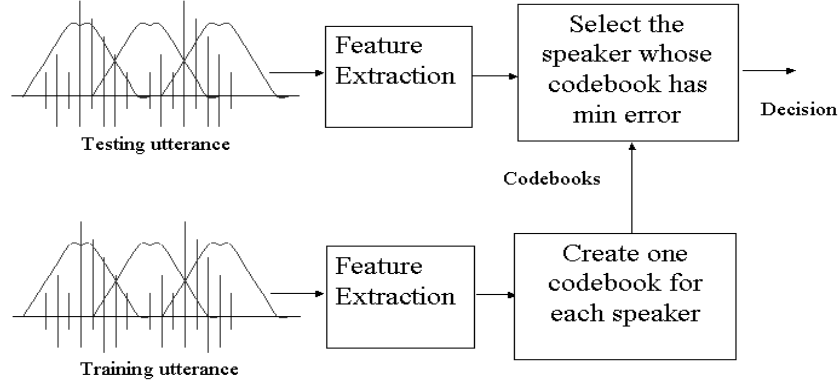


Fig. 1. A VQ-based speaker identification system features an online sub-system for identifying testing audio utterance, and an offline training sub-system, which uses training audio utterance to generate a codebook for each speaker in the database.

A VQ codebook normally consists of centroids of partitions over speaker's feature vector space. The effects to SI by different partition clustering algorithms, such as the LBG and the RLS, have been studied [7]. The average error or distortion of the feature vectors $\{X_t, 1 \leq t \leq T\}$ of length T with a speaker k codebook is given by

$$e_k = \frac{1}{T} \sum_{t=1}^T \min_{1 \leq j \leq S} [d(X_t, C_{k,j})] \quad 1 \leq k \leq L \quad (1)$$

$d(\cdot, \cdot)$ is a distance function between two vectors. $C_{k,j} = (c_{k,j,1}, \dots, c_{k,j,D})^T$ is the j code of dimension D . S is the codebook size. L is the total number of speakers in the database. The *baseline VQ algorithm* of SI simply uses the LBG to generate codebooks and the square of the Euclidean distance as the $d(\cdot, \cdot)$.

Many improvements to the baseline VQ algorithm have been published. Among them, there are two independent approaches: (1) choose a weighted distance function, such as the F-ratio and IHM weights [3], the Partition Normalized Distance Measure (PNDM) [8], and the Bhattacharyya Distance [9]; (2) explore discrimination power of inter-speaker characteristics using the entire set of speakers, such as the Group Vector Quantization (GVQ) discriminative training [2], and the Speaker Discriminative Weighting [10]. Experimentally we have found that PNDM and GVQ are two very effective methods in each of the groups respectively.

1.2 Review of Partition Normalized Distance Measure

The Partition Normalized Distance Measure is defined as the square of the weighted Euclidean distance.

$$d_p(X, C_{k,j}) = \sum_{i=1}^D w_{k,j,i} (x_i - c_{k,j,i})^2 \quad (2)$$

The weighting coefficients are determined by minimizing the average error of training utterances of all the speakers, subject to the constraint that the geometric mean of the weights for each partition is equal to 1.

Let $X_{k,j} = (x_{k,j,1}, \dots, x_{k,j,D})^T$ be a random training feature vector of speaker k , which is assigned to partition j via minimization process in Equation (1). It has mean and variance vectors:

$$\begin{aligned} C_{k,j} &= E[X_{k,j}] \\ V_{k,j} &= E[(X_{k,j} - C_{k,j})^T (X_{k,j} - C_{k,j})] \end{aligned} \quad (3)$$

The constrained optimization criterion to be minimized in order to derive the weights is

$$\begin{aligned} \xi &= \frac{1}{L \cdot S} \sum_{k=1}^L \sum_{j=1}^S \{E[d_p(X_{k,j}, C_{k,j})] + \lambda_{k,j} (\prod_{i=1}^D w_{k,j,i} - 1)\} \\ &= \frac{1}{L \cdot S} \sum_{k=1}^L \sum_{j=1}^S \left\{ \sum_{i=1}^D w_{k,j,i} E[(x_{k,j,i} - c_{k,j,i})^2] + \lambda_{k,j} (\prod_{i=1}^D w_{k,j,i} - 1) \right\} \\ &= \frac{1}{L \cdot S} \sum_{k=1}^L \sum_{j=1}^S \left\{ \sum_{i=1}^D w_{k,j,i} v_{k,j,i} + \lambda_{k,j} (\prod_{i=1}^D w_{k,j,i} - 1) \right\} \end{aligned} \quad (4)$$

Where L is the number of speakers, and S is the codebook size. Letting

$$\frac{\partial \xi}{\partial w_{k,j,i}} = 0, \quad \text{and} \quad \frac{\partial \xi}{\partial \lambda_{k,j}} = 0 \quad (5)$$

We have

$$\lambda_{k,j} = \left(\prod_{i=1}^D v_{k,j,i} \right)^{\frac{1}{D}}, \quad \text{and} \quad w_{k,j,i} = \frac{\lambda_{k,j}}{v_{k,j,i}} \quad (6)$$

Where sub-script i is the feature vector component index, k and j are speaker and partition indices respectively. Because k and j are in both sides of the equations, the weights are only dependent on the data from one partition of one speaker.

1.3 Review of Group Vector Quantization

Discriminative training is to use the data of all the speakers to train the codebook, so that it can achieve more accurate identification results by exploring the inter-speaker differences. The GVQ training algorithm is described as follows.

Group Vector Quantization Algorithm:

- 1) Randomly choose a speaker j .
- 2) Select N vectors $\{X_{j,t}, 1 \leq t \leq N\}$
- 3) calculate error for all the codebooks.
If following conditions are satisfied go to 4)
 - (a) $e_i = \min_{\forall k} \{e_k\}$, but $i \neq j$;
 - (b) $\frac{e_j - e_i}{e_j} < W$, where W is a window size;
 Else go to 5)
- 4) for each $\{X_{j,t}, 1 \leq t \leq N\}$;

$$C_{j,m} \Leftarrow (1-\alpha) \cdot C_{j,m} + \alpha \cdot X_{j,t} \quad \text{where} \quad C_{j,m} = \arg \min_{C_{j,l}} \{d(X_{j,t}, C_{j,l})\}$$

$$C_{i,n} \Leftarrow (1+\alpha) \cdot C_{i,n} - \alpha \cdot X_{j,t} \quad C_{i,n} = \arg \min_{C_{i,l}} \{d(X_{j,t}, C_{i,l})\}$$
- 5) for each $\{X_{j,t}, 1 \leq t \leq N\}$;

$$C_{j,m} \Leftarrow (1-\alpha) \cdot C_{j,m} + \varepsilon \alpha \cdot X_{j,t}, \quad \text{where} \quad C_{j,m} = \arg \min_{C_{j,l}} \{d(X_{j,t}, C_{j,l})\}$$

2 Enhancements

We propose the following steps to further enhance the VQ based solution: (1) a Heuristic Weighted Distance (HWD), (2) combination of HWD and GVQ, and (3) combination of PNDM and GVQ.

2.1 Heuristic Weighted Distance

The PNDM weights are inversely proportional to partition variances of the feature components, as shown in Equation (6). It has been shown that variances of cepstral coefficient of order i is proportional to $1/i^2$ [11]. Clearly $v_i > v_{i+1}, 1 \leq i \leq D-1$,

where i is the vector element index, which reflects frequency band. The higher the index, the less feature value and its variance.

We considered a Heuristic Weighted Distance as

$$d_h(X, C_{k,i}) = \sum_{i=1}^D w_i(S, D) \cdot (x_i - c_{k,j,i})^2 \quad (7)$$

The weights are calculated by

$$w_i(S, D) = 1 + c(S, D) \cdot (i - 1) \quad 1 \leq i \leq D \quad (8)$$

Where $c(S, D)$ is a function of both the codebook size S and the feature vector dimension D . For a given codebook, S and D are fixed, and thus $c(S, D)$ is a constant. The value of $c(S, D)$ is estimated experimentally by performing an exhaustive search to achieve the maximum identification rate in a given sample test dataset.

2.2 Combination of HWD and GVQ

Combination of the HWD and the GVQ is achieved by simply replacing the original square of the Euclidean distance with the HWD Equation (7), and to adjust the GVQ updating parameter α whenever needed.

2.3 Combination of PNDM and GVQ

To combine PNDM with the GVQ requires a slight more work, because the GVQ alters the partition and thus its component variance. We have used the following algorithm to overcome this problem.

Algorithm to Combine PNDM with the GVQ Discriminative Training:

- 1) Use LBG algorithm to generate initial LBG codebooks;
- 2) Calculate PNDM weights using the LBG codebooks, and produce PNDM weighted LBG codebooks, which are LBG codebooks appended with the PNDM weights;
- 3) Perform GVQ training with PNDM distance function, and generate the initial PNDM+GVQ codebooks by replacing the LBG codes with the GVQ codes;

- 4) Recalculate PNDM weights using the PNDM+GVQ codebooks, and produce the final PNDM+GVQ codebooks by replacing the old PNDM weights with the new ones;

3 Experimental Comparison of VQ-based Algorithms

3.1 Testing Data and Procedures

168 speakers in TEST section of the TIMIT corpus are used for SI experiment, and 190 speakers from DR1, DR2, DR3 of TRAIN section are used for estimating the $c(S, D)$ parameter. Each speaker has 10 good quality recordings of 16 KHz, 16bits/sample, and stored as WAVE files in NIST format. Two of them, SA1.WAV and SA2.WAV, are used for testing, and the rest for training codebooks. We did not perform silence removal on WAVE files, so that others could reproduce the environment with no additional complication of VAD algorithms and their parameters.

A MFCC program converts all the WAVE files in a directory into one feature vector file, in which all the feature vectors are indexed with its speaker and recording. For each value of feature vector dimension, $D=30, 40, 50, 60, 70, 80, 90$, one training file and one testing file are created. They are used by all the algorithms to train codebooks of size $S=16, 32, 64$, and to perform identification test, respectively.

The MFCC feature vectors are calculated as follows: 1) divide the entire utterance into blocks of size 512 samples with 256 overlapping; 2) perform pre-emphasize filtering with coefficient 0.97; 3) multiply with Hamming window, and perform short-time FFT; 4) apply the standard mel-frequency triangular filter banks to the square of magnitude of FFT; 5) apply the logarithm to the sum of all the outputs of each individual filter; 6) apply DCT on the entire set of data resulted from all filters; 7) drop the zero coefficient, to produce the cepstral coefficients; 8) after all the blocks being processed, calculate the mean over the entire time duration and subtract it from the cepstral coefficients; 9) calculate the 1st order time derivatives of cepstral coefficients, and concatenate them after the cepstral coefficients, to form a feature vector. For example, a filter-bank of size 16 will produce 30 dimensional feature vectors.

Due to project time constraint, the HWD parameter $c(S, D)$ was estimated at $S=16, 32, 64, D=40, 80$, so that it achieves the highest identification rate using the 190 speakers dataset of TRAIN section. For other values of S and D , it was interpolated or extrapolated from optimized samples. The results are shown in the bottom section of Table 1. The identification experiment was then performed using the 168 speakers dataset from TEST section. We have used different datasets for $c(S, D)$ estimation, codebooks training, and identification rate testing, to produce objective results.

3.2 Testing Results

Table 1 shows identification rates for various algorithms. The value of the learning parameter α is displayed after the GVQ title, and the parameter $c(S, D)$ is displayed at bottom section. Combination of the algorithms are indicated by a “+” sign between their name abbreviations.

Table 1. Identification rates (%) and parameters for various VQ-based algorithms tested, where the 1st row is the feature vector dimension D , and the 1st column is the codebook size S .

S/D	30	40	50	60	70	80	90
Baseline VQ Algorithm							
16	70.5	79.8	83.6	83.9	87.2	85.1	84.5
32	86.0	91.4	91.4	92.9	90.8	92.3	91.4
64	89.3	93.8	94.9	96.1	93.2	96.4	95.8
HWD							
16	78.3	85.7	88.4	87.8	87.5	87.2	88.7
32	89.3	92.6	93.2	95.2	92.3	94.3	93.2
64	95.5	97.6	97.0	96.7	96.1	96.7	98.5
PNDM							
16	93.8	96.1	97.6	97.6	96.4	96.1	93.5
32	94.9	96.7	98.5	98.8	98.2	97.6	96.4
64	96.1	97.9	97.3	98.2	98.2	97.6	96.1
GVQ ($\alpha=0.2$)							
16	81.0	88.7	94.3	95.2	95.5	95.8	96.4
32	92.6	95.2	97	97.3	96.7	97	98.2
64	91.7	97	97	97.6	98.2	98.8	98.5
HWD+GVQ ($\alpha=0.2$)							
16	82.7	92.9	97.0	96.7	97.6	97.0	97.6
32	95.8	98.2	98.2	97.6	99.1	99.7	98.8
64	95.8	98.5	98.8	99.4	99.4	98.2	97.6
PNDM+GVQ ($\alpha=0.08$)							
16	96.4	97.9	99.1	99.1	99.1	99.4	99.4
32	97.3	97.9	98.5	98.8	99.4	98.8	98.2
64	96.7	99.1	97.9	98.5	98.5	97.6	97.6
$c(S, D)$							
16	1.240	0.903	0.566	0.382	0.199	0.134	0.068
32	1.103	0.803	0.503	0.203	0.153	0.103	0.053
64	1.103	0.803	0.503	0.351	0.199	0.134	0.068

The baseline algorithm performs poorest as expected. The plain HWD, PNDM, and GVQ all show enhancements over the baseline. Combination methods further enhanced the plain methods. The PNDM+GVQ performs best when codebook size is 16 or 32, while the HWD+GVQ is better at codebook size 64. The highest score of the test is 99.7%, and corresponds to a single miss in 336 utterances of 168 speakers. It outperforms the reported rate 98.4% by using the GMM with WPP features [4].

4 Conclusion

A new approach combining the weighted distance measure and the discriminative training is proposed to enhance VQ-based solutions for speech independent speaker identification. An alternative heuristic weighted distance measure was explored, which lifts up higher order MFCC feature vector components using a linear formula. Two new algorithms combining the heuristic weighted distance and the partition normalize distance with the group vector quantization discriminative training were developed, which gathers the power of both the weighted distance measure and the discriminative training. Experiments showed that the proposed methods outperform the corresponding single approach VQ-based algorithms, and even more powerful GMM based solutions. Further research on heuristic weighted distance is being conducted particularly for small codebook size.

References

1. Reynolds, A.D., and Rose, C.R.: "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models". *IEEE Transactions on Speech and Audio Processing*, 3(1): 72-83, 1995.
2. He, J., Liu, L., Palm, G.: "A Discriminative Training Algorithm for VQ-based Speaker Identification", *IEEE Transactions on Speech and Audio Processing*, 7(3): 353-356, 1999.
3. Zilca, R.D., and Bistriz, Y.: "Text Independent Speaker Identification Using Lsp Codebook Speaker Models and Linear Discriminate Functions", *Proc. 6th European Conference on Speech Communication and Technology*, (2): 799-802, Budapest, Hungary, 1999.
4. Sarikaya, R., Pellom, B.L., and Hansen, H.L.: "Wavelet Packet Transform Features with Application to Speaker Identification", *IEEE Nordic Signal Processing Symposium*, 81-84, Vigso, Denmark, 1998.
5. Jang, G.J., Yun, S.J., Oh, Y.H.: "Feature Vector Transforming Using Independent Component Analysis and its Application to Speaker Identification", *Proc. 6th European Conference on Speech Communication and Technology*, (2): 767-770, Budapest, Hungary, 1999.
6. Rosca, J., Kofmehl, A.: "Cepstrum-like ICA Representations for Text Independent Speaker Recognition", to be presented at *4th Int. Independent Component Analysis and Blind Signal Separation*, Nara, Japan, April 1-4, 2003.
7. Kinnunen, T., Kilpeläinen, T., Fränti, P.: "Comparison of Clustering Algorithms in Speaker Identification", *Proc. IASTED Int. Conf. Signal Processing and Communications*, 222-227, Marbella, Spain, 2000.
8. Wang, R.H., He, L.S., and Fujisaki, H.: "A Weighted Distance Measure Based on the Fine Structure of Feature Space: Application to Speaker Recognition", *Int. Conference of Acoustic Speech & Signal Processing*, 273-276, 1990.
9. Petry, A., Zanuz, A., Barone, D.A.C.: "Bhattacharyya Distance Applied to Speaker Identification", *Int. Conference on Signal Processing Applications and Technology*, Dallas, Orlando, (1): 2000.
10. Kinnunen, T., Fränti, P.: "Speaker Discriminative Weighting Method for VQ-based Speaker Identification", *Proc. 3rd Int. Conf. on Audio and Video-based Biometric Person Authentication*, 150-156, Halmstad, Sweden, 2001.
11. Rabiner, L. and Juang, B.: *Fundamentals of Speech Recognition*, Prentice Hall, 1993.