Symmetric Distortion Measure for Speaker Recognition

Evgeny Karpov, Tomi Kinnunen, Pasi Fränti

Department of Computer Science
University of Joensuu, Finland
{ekarpov,tkinnu,franti}@cs.joensuu.fi

Abstract

We consider matching functions in vector quantization (VQ) based speaker recognition systems. In VQ-based systems, a speaker model consists of a small collection of representative vectors, and matching is performed by computing a dissimilarity value between the unknown speaker's feature vectors and the speaker models. Typically, the average/total quantization error is used as the dissimilarity measure. However, this measure lack the symmetricity requirement of a proper distance measure. This is counterintuitive because match score between speakers X and Y is different from the match score between Y and X. Furthermore, the distortion measure can yield a zero value (perfect match) for non-identical vector sets, which is undesirable. In this study, we study ways of making the quantization distortion functions proper distance measures. The study includes discussion of the theoretical properties of different measures, as well as an evaluation on a subset of the NIST99 speaker recognition evaluation corpus.

1. Introduction

Classically, *speaker recognition* systems are divided into two classes [5, 12]: *identification* and *verification* systems. In the identification problem (or 1:N *matching*), the task is to indicate the best-matching speaker with a given unknown speaker from a database of N known speakers. In the verification problem (or 1:1 *matching*), the task is to make a decision whether the unknown speaker is who he/she claims to be.

Both speaker recognition tasks include the same basic components: (1) *feature extraction*, (2) *speaker modeling*, and (3) *speaker matching*. Only the last phase, (4) *decision logic*, depends on the task. In this study, our focus is on the matching part, and therefore it does not matter which one of the two tasks we use in our experiments. We have selected to use the identification task since its performance is easier to quantify.

The feature extraction module converts the raw speech waveform to a sequence of *acoustic feature vectors*, denoted here as $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$. In the speaker enrollment phase, a model of the speaker's vocal space is formed by using the training vectors. In the recognition phase, same type of feature vectors are extracted and they are compared with the stored speaker model(s), giving a *measure of similarity* between the vector sequence and the model(s).

Various features as well as speaker models have been proposed for speaker recognition. Classical features include the *cepstrum* with many variants [1, 11, 5], and *line spectral frequencies* [5]. Recently, *subband processing* has also become a popular technique [3, 2, 15, 28, 7, 24]. Some of the various modeling techniques include *vector quantization* (VQ) [29, 18, 16, 9], *Gaussian mixture models (GMM)* [27, 26], *covariances models* [4], and *neural networks* [10]. An overview of several modeling techniques is given in [25].

In this paper, we will focus on the VQ approach because it is simple to implement and computationally efficient. In the VQ approach, match score computation is based on dissimilarity measure between the unknown speaker's feature vectors and the model. However, the baseline quantization distortion lacks some properties of a proper distance measure. In this paper, we discuss these shortcomings, and ways to attack them.

2. VQ-Based Speaker Identification

In the VQ-based approach to speaker identification [29, 13, 18, 9], the speaker models are formed by clustering the speaker's feature vectors into K disjoint clusters. Each cluster is represented by a *code vector* $\boldsymbol{c}_i$, which is the centroid (average vector) of the cluster. The resulting set of code vectors $\{\boldsymbol{c}_1, \cdots, \boldsymbol{c}_K\}$ is called a *codebook*, and it serves as the model for the speaker.

Good clustering should produce a compact codebook whose vectors have the same underlying probability distribution as the training vectors. Thus, the codebook effectively reduces the amount of data by preserving essential information. We have previously addressed the issue of codebook generation for speaker recognition [18]. Our main conclusion was that the choice of the clustering algorithm is not crucial, but the codebook size must be selected carefully. Larger codebook models more accurately the original distribution, but might start to *overfit* if set too large. Typical speaker codebook size is around 64 to 512 vectors, depending on the selected features.

The matching function in the baseline VQ-based speaker recognition [29] is the *quantization distortion* between the two vector sets to be compared. Given a feature vector $\boldsymbol{x}_i$, generated by the unknown speaker, and a codebook $C = \{\boldsymbol{c}_1, \cdots, \boldsymbol{c}_K\}$, the quantization error e of the vector $\boldsymbol{x}_i$ with respect to C is given by

$$e(\boldsymbol{x}_i, C) = \min_{\boldsymbol{c}_j \in C} d(\boldsymbol{x}_i, \boldsymbol{c}_j), \tag{1}$$

where $d(\cdot, \cdot)$ is a distance metric defined over the feature space. In other words, the quantization distortion for a single vector is computed as the distance to the nearest vector in the codebook. Typically Euclidean metric is used as the distance measure. *Total quantization distortion* D_Q is defined as the sum of the individual distortions:

$$D_Q(X, C) = \sum_{\boldsymbol{x}_i \in X} e(\boldsymbol{x}_i, C). \tag{2}$$

Sometimes (2) is normalized by the number of test vectors. However, this normalization factor $1/|X|$ is the same for all speakers, and therefore it does not change the order of the speakers in the matching result.

Several modifications have been proposed to the baseline VQ distortion matching [30, 22, 14, 16, 19, 9]. For instance, in [16, 19], we assign to each VQ code vector a *discriminative weight* so that a code vector that is close to some other speaker's code vector is given a small contribution to the overall distance. Some alternative VQ methods are compared in [9]. Despite the existence of these possibly more sophisticated methods, the baseline matching function (2) is typically used due to its simplicity. However, it has a few shortcomings, which will be discussed next.

3. Shortcomings of the Baseline Measure

It is easy to see that (2) is not *symmetric*, i.e. $D_Q(X, C) \neq D_Q(C, X)$ in general. For instance, by choosing $X = \{(0, 0)\}$ and $C = \{(0, 1), (1, 0)\}$, we get $D_Q(X, C) = 1$ but $D_Q(C, X) = 1 + 1 = 2$. The lack of symmetricity is counter-intuitive. Mathematically, a a *distance measure* $D(A, B)$ between two objects A and B of interest satisfies the following three properties [23]:

(i)	$D(A, B) \geq 0$	for all A, B
(ii)	$D(A, B) = 0$	if and only if $A = B$
(iii)	$D(A, B) = D(B, A)$	for all A, B

In fact, (2) satisfies only the condition (i). By choosing $X = \{(0, 0)\}$ and $C = \{(0, 0), (0, 1)\}$, we get $D_Q(X, C) = 0$ but $X \neq C$. In fact, we can show that $D_Q(X, C) = 0$ if and only if $X \subseteq C$:

Theorem 1. $D_Q(X, C) = 0$ *if and only if* $X \subseteq C$.

Proof. First, assume $X \subseteq C$. For each $\boldsymbol{x}_i \in X$ there is identical vector in C, and therefore $e(\boldsymbol{x}_i, C) = 0$. It follows that $D_Q(X, C) = 0$.

Conversely, let $D_Q(X, C) = 0$. Since all the sum terms in (2) are nonnegative, $e(\boldsymbol{x}_i, C) = 0$ for all $x_i \in C$, i.e. $\min_{\boldsymbol{c}_j} d(\boldsymbol{x}_i, \boldsymbol{c}_j) = 0$. Since $d(\cdot, \cdot)$ is a distance measure, the nearest neighbor of $\boldsymbol{x}_i$ in C must be $\boldsymbol{x}_i$ itself, i.e. $\boldsymbol{x}_i \in C$. Therefore, $X \subseteq C$. □

The result means that if the target speaker happens to produce a "subvoice" of some other speaker, the target speaker will be incorrectly assigned to this speaker. The quantization distortion (2) is not originally designed for pattern matching, but to give a quality index of a vector quantizer.

4. Symmetric Measures

Given $D_Q(X, C), D_Q(C, X)$, and a function $\mathcal{F} : \mathbb{R}^2 \to \mathbb{R}$ that satisfies $\mathcal{F}(a, b) = \mathcal{F}(b, a)$ for all $a, b \in \mathbb{R}$ (i.e., $\mathcal{F}$ is symmetric), we can construct a symmetric measure $D_{\mathcal{F}}(X, C) = \mathcal{F}(D_Q(X, C), D_Q(C, X))$. Commonly used symmetrization functions include minimum, maximum, sum and product. These induce the following symmetric measures:

$$\begin{aligned}
D_{min}(X, C) &= \min(D_Q(X, C), D_Q(C, X)) \\
D_{max}(X, C) &= \max(D_Q(X, C), D_Q(C, X)) \\
D_{sum}(X, C) &= D_Q(X, C) + D_Q(C, X) \\
D_{prod}(X, C) &= D_Q(X, C) \cdot D_Q(C, X)
\end{aligned}$$

All of these are symmetric, as they satisfy requirement (iii). The requirement (ii), however, is satisfied only by D_{max} and D_{sum}, as shown by the following Theorem:

Theorem 2. *(a)* $D_{max}(X, C) = 0 \iff X = C$
(b) $D_{sum}(X, C) = 0 \iff X = C$

Proof. Theorem (a) can be proven as follows:

$$\begin{aligned}
&D_{max}(X, C) = 0 \\
&\qquad \iff \max(D_Q(X, C), D_Q(C, X)) = 0 \\
&\qquad \iff D_Q(X, C) = 0 \wedge D_Q(C, X) = 0 \\
&\qquad \iff X \subseteq C \wedge C \subseteq X \\
&\qquad \iff X = C
\end{aligned}$$

Case (b) can be proven in similar way:

$$\begin{aligned}
&D_{sum}(X, C) = 0 \\
&\qquad \iff D_Q(X, C) + D_Q(C, X) = 0 \\
&\qquad \quad D_Q(X, C) \geq 0 \wedge D_Q(C, X) \geq 0 \\
&\qquad \iff D_Q(X, C) = (C, X) = 0 \\
&\qquad \iff X \subseteq C \wedge C \subseteq X \\
&\qquad \iff X = C
\end{aligned}$$

□

Neither D_{min} nor D_{prod} satisfy requirement (iii). It can be easily proven by counterexample: when $D_Q(X, C) = 0$ and $D_Q(C, X) = 1$ both D_{min} and D_{prod} will be zero.

The distance properties of the symmetric measures along with the baseline measure D_Q are summarized in Table 1. Based on these properties, we expect D_{max} and D_{sum} to perform the best in practice since these are real distance measures. However, none of these measures is a *metric*, which would be another natural requirement for a proximity measure.

Table 1: Summary of the distance properties of the measures.

Measure	Distance property		
	(i)	(ii)	(iii)
D_Q	✓		
D_{min}	✓		✓
D_{prod}	✓		✓
D_{sum}	✓	✓	✓
D_{max}	✓	✓	✓

5. Experiments

5.1. Speech Material and Parameter Setup

For the experiments, we used a subset of the *NIST 1999 speaker recognition evaluation corpus* [21] (see Table 2). We selected to use the data from the male speakers only. For training, we used both the "a" and "b" files for each speaker. For identification, we used the one speaker test segments from the same telephone line. In general it can be assumed that if two calls are from different lines, the handsets are different, and if they are from the same line, the handsets are the same [21]. In other words, the training and matching conditions have very likely the same handset type (electret/carbon button) for each speaker, but different speakers can have different handsets. The total number of test segments for this condition is 692.

The parameters for different feature sets and training algorithm were based on our previous experiments with the NIST corpus [17]. The frame length and shift were set to 30 ms and 20 ms, respectively, and the window function was Hamming. We use the standard MFCCs as the features [8]. A pre-emphasiz filter $H(z) = 1 - 0.97z^{-1}$ is used before framing. Each frame is multiplied with a 30 ms Hamming window, shifted by 20 ms. From the windowed frame, FFT is computed, and the magnitude spectrum is filtered with a bank of 27 triangular filters spaced linearly on the mel-scale. The log-compressed filter outputs are converted into cepstral coefficients by DCT, and the 0^{th} cepstral coefficient is ignored. Speaker models are generated by the LBG clustering algorithm [20].

Table 2: Summary of the NIST-1999 subset

Language	English
Speakers	230
Speech type	Conversational
Quality	Telephone
Sampling rate	8.0 kHz
Quantization	8-bit μ-law
Training speech (avg.)	119.0 sec.
Evaluation speech (avg.)	30.4 sec.

5.2. Results

First, we experimented with different vector metrics and distortion measures by fixing the codebook size to 64. The results are shown in Table 3. For a fixed distortion measure, the error rates increase in most cases when moving from Euclidean to Maximum metric. However, the differences between metrics are relatively small, as it was expected.

The smallest error rate 16.8 % is reached with the baseline measure D_Q and D_{max} using the Euclidean distance. The order of the four symmetric measures is as expected, D_{min} and D_{prod} perform the worst while D_{sum} and D_{max} the best. However, the baseline measure D_Q is not the worst as it was hypothesized.

Next, we fixed the metric to the Euclidean and varied the codebook size (see Table 4). Increasing codebook size improves performance in most cases, as expected from our previous experience [18, 17]. However, for D_{sum} and D_{max} the performance in fact degrades with the codebook size, for which the reason is unclear. The best performance (13.6 %) is obtained with the baseline measure D_Q with the codebook size 1024.

5.3. Discussion

In general, we conclude that the performance of the proposed symmetric measures is not as good as expected. In particular, the experiments do not support the hypothesis regarding the order of the measures based on their distance properties. Therefore, further theoretical development is needed. It is possible that the *metric* properties would be important, i.e. the distance functions satisfying the triangular inequality. However, a "reasonable" metric between two vector sets is not easy to define. One candidate would be the Hausdorff metric [6] which have also tested with poor success. The reason for this is that the Hausdorff metric links only one vector from each of the sets, and assigns the distance between the distance of the sets as the distance between these vectors. As a consequence, information about the distribution *shape* is lost.

One possible future direction could be towards information theoretic distance measures. For instance, the *Kullback-Leibler (KL) divergence* satisfies all other requirements of a distance measure except the symmetricity property. The symmetrizations proposed in this study could be applied in the same way to the KL divergence, as well.

6. Conclusions

We have studied the distance properties of symmetric distortions for text-independent speaker recognition using the vector quantization approach. We proposed to use four commonly used symmetrizations. For D_{sum} and D_{max}, the measures were proven to be properly defined distance measures.

Table 3: Error rates for different distance measures.

	Euclidean	Manhattan	Maximum
Baseline	16.8	17.2	19.0
D_{min}	37.0	34.7	36.7
D_{prod}	22.5	20.9	23.3
D_{sum}	16.9	17.0	19.0
D_{max}	16.8	17.3	19.0

Table 4: Error rates for varying codebook size.

Distance	CB=64	CB=256	CB=512	CB=1024
Baseline	16.8	14.9	14.2	13.6
D_{min}	37.0	32.4	29.6	24.4
D_{prod}	22.5	18.9	18.6	18.1
D_{sum}	16.9	17.9	18.2	18.9
D_{max}	16.8	16.5	18.4	23.0

However, the experimental results indicated that the symmetric measures were no better than the baseline quantization distortion. Therefore, the future work should consists of deeper analysis as well as totally new directions, e.g. in an information-theoretic framework.

7. Acknowledgements

The work of E. Karpov was supported by the National Technology Agency of Finland (TEKES) (project “Puhetekniikan uudet menetelmät ja sovellukset”, TEKES dnro 8713103).

8. References

[1] B. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustic Society of America*, 55(6):1304–1312, 1974.

[2] L. Besacier, J.F. Bonastre, and C. Fredouille. Localization and selection of speaker-specific information with statistical modeling. *Speech Communications*, 31:89–106, 2000.

[3] Laurent Besacier and Jean-François Bonastre. Subband architecture for automatic speaker recognition. *Signal Processing*, 80:1245–1259, 2000.

[4] F. Bimbot and L. Mathan. Text-free speaker recognition using an arithmetic-harmonic sphericity measure. In *Proc. 3th European Conference on Speech Communication and Technology (Eurospeech 1993)*, pages 169–172, Berlin, Germany, 1993.

[5] J. Campbell. Speaker recognition: a tutorial. *Proceedings of the IEEE*, 85(9):1437–1462, 1997.

[6] E. Chavez, G. Navarro, R. Baeza-Yates, and J.L. Marroquin. Searching in metric spaces. *ACM Computing Surveys*, 33(3):273–321, 2001.

[7] R.I. Damper and J.E. Higgins. Improving speaker identification in noise by subband processing and decision fusion. *Pattern Recognition Letters*, 24:2167–2173, 2003.

[8] J.R. Jr. Deller, J.H.L. Hansen, and J.G. Proakis. *Discrete-Time Processing of Speech Signals*. IEEE Press, New York, second edition, 2000.

[9] N. Fan and J. Rosca. Enhanced VQ-based algorithms for speech independent speaker identification. In *Proc. Audio- and Video-Based Biometric Authentication (AVBPA 2003)*, pages 470–477, Guildford, UK, 2003.

[10] K.R. Farrell, R.J. Mammone, and K.T. Assaleh. Speaker recognition using neural networks and conventional classifiers. *IEEE Trans. on Speech and Audio Processing*, 2(1):194–205, 1994.

[11] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(2):254–272, 1981.

[12] S. Furui. Recent advances in speaker recognition. *Pattern Recognition Letters*, 18(9):859–872, 1997.

[13] J. He, L. Liu, and G. Palm. A discriminative training algorithm for VQ-based speaker identification. *IEEE Trans. on Speech and Audio Processing*, 7(3):353–356, 1999.

[14] A.L. Higgins, L.G. Bahler, and J.E. Porter. Voice identification using nearest-neighbor distance measure. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 1993)*, pages 375–378, Minneapolis, Minnesota, USA, 1993.

[15] T. Kinnunen. Designing a speaker-discriminative filter bank for speaker recognition. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP 2002)*, pages 2325–2328, Denver, Colorado, USA, 2002.

[16] T. Kinnunen and P. Fränti. Speaker discriminative weighting method for VQ-based speaker identification. In *Proc. Audio- and Video-Based Biometric Authentication (AVBPA 2001)*, pages 150–156, Halmstad, Sweden, 2001.

[17] T. Kinnunen, E. Karpov, and P. Fränti. Real-time speaker identification. In *Int. Conf. on Spoken Language 2004 (ICSLP 2004)*, Jeju Island, Korea (Accepted for publication), 2004.

[18] T. Kinnunen, T. Kilpeläinen, and P. Fränti. Comparison of clustering algorithms in speaker identification. In *Proc. IASTED Int. Conf. Signal Processing and Communications (SPC 2000)*, pages 222–227, Marbella, Spain, 2000.

[19] T. Kinnunen and I. Kärkkäinen. Class-discriminative weighted distortion measure for VQ-based speaker identification. In *Proc. Joint IAPR International Workshop on Statistical Pattern Recognition (S+SPR2002)*, pages 681–688, Windsor, Canada, 2002.

[20] Y. Linde, A. Buzo, and R.M. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 28(1):84–95, 1980.

[21] A. Martin and M. Przybocki. The NIST 1999 speaker recognition evaluation - an overview. *Digital Signal Processing*, 10(1-18):1–18, 2000.

[22] T. Matsui and S. Furui. A text-independent speaker recognition method robust against utterance variations. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 1991)*, pages 377–380, Toronto, Canada, 1991.

[23] R.S. Millman and G.D.Parker. *Geometry - a Metric Approach with Models*. Springer Verlag, New York, second edition, 1991.

[24] J. Ming, D. Stewart, P. Hanna, P. Corr, J. Smith, and S. Vaseghi. Robust speaker identification using posterior union models. In *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, pages 2645–2648, Geneva, Switzerland, 2003.

[25] R.P. Ramachandran, K.R. Farrell, R. Ramachandran, and R.J. Mammone. Speaker recognition - general classifier approaches and data fusion methods. *Pattern Recognition*, 35:2801–2821, 2002.

[26] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1):19–41, 2000.

[27] D.A. Reynolds and R.C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Trans. on Speech and Audio Processing*, 3:72–83, 1995.

[28] P. Sivakumaran, A.M. Ariyaeeinia, and M.J. Loomes. Sub-band based text-dependent speaker verification. *Speech Communications*, 41:485–509, 2003.

[29] F.K. Soong, A.E. Rosenberg A.E., B.-H. Juang, and L.R. Rabiner. A vector quantization approach to speaker recognition. *AT & T Technical Journal*, 66:14–26, 1987.

[30] R.-H. Wang, L.-S. He, and H. Fujisaki. A weighted distance measure based on the fine structure of feature space: application to speaker recognition. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 1990)*, pages 273–276, Albuquerque, New Mexico, USA, 1990.