Conformity, a New Method for Text-Independent Speaker Recognition

Philippe Thévenaz, Heinz Hügli

Abstract— **This article presents Conformity as a new method for text-independent speaker recognition. Its principle is to compare statistical distributions of samples. We estimate these distributions by vector quantization using a single, universal codebook. One computes the frequency of apparition of each codeword for test and reference speech, and the frequencies are compared.**

We perform several experiments with a database consisting of many speakers and a high number of tests, and conduct experiments in an opened test methodology. We compare the results to those obtained by other well-known methods.

Our new method yields good results when used alone. It may also be combined advantageously with the classical vector quantization method to which it is complementary. The joint use of these two methods permits an enhanced recognition rate.

Keywords— **Speaker verification, text independence, vector quantization, opened test.**

1. Introduction

The problem addressed is that of text-independent speaker recognition. Text independence makes it harder than text dependence, because one suffers from the high variability of speech related to text freedom. The method known as Average Vector Quantization Error (VQ) is one of the most efficient methods currently available to solve this problem; its principle is to vector-quantize speech with the aid of some reference codebook. The comparison of unquantized with quantized speech gives the degree of fitness of the speech with respect to the codebook, which in turn is speaker dependent. However, a weakness of this method is that only the distribution of features with respect to their nearest codewords is considered, the global distribution being ignored so far.

This paper presents Conformity (Conf) as a new method for text-independent speaker recognition. The underlying principle of the new method is advantageous because it recovers some of the information ignored by VQ; this principle is to directly compare statistical distributions of samples. We estimate these distributions by vector quantization using a single, universal codebook. One computes the frequency of apparition of each codeword for test and reference speech, and the frequencies are compared. One finds then that the classical VQ method doesn't process at all the information available in the codeword selections, while our new method considers them exclusively. It follows that the two methods may be complementary.

Ph. Thévenaz was with the Institute of Microtechnology of the University of Neuchâtel while working for this project. He is now with the National Institutes of Health, Bethesda (MD). E-mail: Philippe.Thevenaz@helix.nih.gov

H. Hügli is with the Institute of Microtechnology of the University of Neuchâtel. E-mail: Heinz.Hugli@imt.unine.ch

To prove this statement, we performed some experiments with a database consisting of many speakers, with a balanced number of male and female people. The high number of tests done assures that the computed error rates are statistically significant. Furthermore, experiments are conducted in an opened test methodology. In order to compare the Conf intrinsic efficiency to that of other methods, we performed four experiments: VQ, Conf, and two other methods making also use of LPC-cepstra. We then combined the results pairwise in order to prove that Conf and VQ are complementary.

2. Outline

The organization of this paper is as follows: we have presented above the problem addressed and we have sketched our contribution to its solution. Following the present outline, we will briefly discuss in paragraph 3 the general features of the VQ approach. Then we will present in paragraph 4 the Conformity method from a formal point of view, followed by some considerations motivating its use in speaker recognition. We will confirm experimentally in paragraph 5 the validity of the proposed new method and, finally, we will conclude in paragraph 6.

3. VQ approach

Most often, text-independent speaker recognition methods compare features based on long term speech statistics. Statistical estimations are made independently on test and on reference speech, and then compared [1]. Globally, the VQ approach follows such a scheme.

3.1 Learning

In the learning phase, a representative is created by applying a classification algorithm to a speaker-dependent speech corpus. The result is a set of coding vectors named codebook; these codebook entries define unequivocal classes in the representation space. A code is also associated to each entry. The next learning step relates to the kind of recognition task (verification or identification) and provide a threshold, for example.

3.2 Distance Computation

The distance computation between test speech and a reference representative proceeds by classifying incoming speech with respect to the personalized codebook we just described. After classification, one has to provide a distortion measure between the test part associated to a given reference class and the representative for this very same class. Doing so, each class contributes now to an accumulated global distance measure between a test speech sample and a reference representative. An example of a VQ distance computation is given in figure 1.

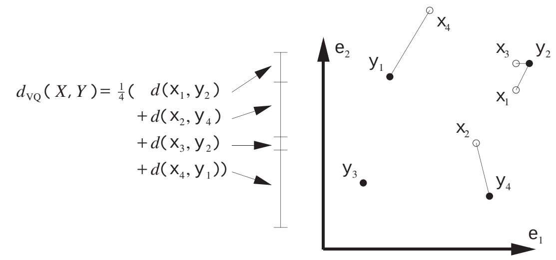

Fig. 1. Average vector quantization error method. A first distance computation (only the result is shown here) selects the nearest neighbor **y** of some incoming speech represented by a set of test vectors **x**. A second distance computation produces a contribution of the considered sample to the global distance. Usually, this second distance is computed in the same way as the first one; in this example, both are euclidean.

3.3 Justification

The two hypothesis underlying this approach are first that normal speech is made of the concatenation of a limited number of steady states. Second, the state collection for one person has to differ from the state collection for another person. If these hypothesis hold true, then a personalized codebook is efficient when reconstructing speech coming from the speaker it is intended for, and inefficient for any other people.

This method of comparison between test and reference speech can be traced back to [2], although its main interest appears with promising results in a twofold article of Soong and Rosenberg in 1986 [3], [4]. Recently, some authors tried to enhance its efficiency by considering delayed VQ [5] or matrix quantization [6].

4. Conformity approach

We present here a new method called Conformity, designed to specifically solve some problems linked with the VQ method. The main idea is to work directly in the statistical space; the statistics computation is made by vector-quantizing any incoming speech by a speaker-independent codebook and by estimating the load of each of its entries. The hypothesis that has to be satisfied in order to allow the distinction between speakers is that some given speaker will ignore some entries that another one will use often.

4.1 Formalism

Let Y be a universal, speaker-independent codebook made of a set of K coding vectors $\mathbf{y}_k$ in a working space U, and let X be a set of P pre-processed speech samples

$$Y = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_K \mid \mathbf{y}_k \in U \wedge \forall k \in [1, K]\}$$

$$X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_P \mid \mathbf{x}_p \in U \wedge \forall p \in [1, P]\}$$

Let d be a measure of distance between a sample vector $\mathbf{x}$ and a coding vector $\mathbf{y}$. The nearest neighbor rule defines then a cell around each coding vector. The union of these cells creates a partition of U. Let q be the assignment function of $\mathbf{x}$ to its nearest neighbor $\mathbf{y}$

$$\mathbf{y} = q(\mathbf{x}) = \underset{\forall \mathbf{y}_k \in Y}{\text{ArgMin}}\ d(\mathbf{x}, \mathbf{y}_k)$$

Let $\delta(x)$ be the unit sample scalar function and let $\delta(\mathbf{x})$ be the unit sample vector function

$$\delta(x) = \begin{cases} 1 & x = 0 \\ 0 & x \neq 0 \end{cases} \quad \forall x \qquad \delta(\mathbf{x}) = \prod_{n=1}^{N} \delta(x_n) \quad \forall \mathbf{x} \in U$$

The probability of use of any codebook entry can be estimated by a histogram computation, yielding a vector $\mathbf{h}$

$$\mathbf{h} = \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_K \end{pmatrix} \qquad h_k = \frac{1}{P} \cdot \sum_{\forall \mathbf{x} \in X} \delta(q(\mathbf{x}) - \mathbf{y}_k) \quad \forall k \in [1, K]$$

The universal codebook Y defines a partition of the working space U. The component j of the reference frequency histogram $\mathbf{h}^{(k)}$ may then be viewed in this space as a (hopefully good) approximation of the expectation of the fraction of samples having to fall within the cell j. One may interpret then $\mathbf{h}^{(k)}$ as a representation of a reference probability density function (pdf). If a test sample behaves according to exactly the same pdf, then its frequency histogram $\mathbf{h}^{(i)}$ should be about the same than $\mathbf{h}^{(k)}$ for a finite number of samples $P^{(i)}$. The comparison between test speech of speaker (i) and a reference speaker (k) is then simply made by computing some distance between the two frequency histograms $\mathbf{h}^{(i)}$ and $\mathbf{h}^{(k)}$

$$d_{\text{Conf}}((i), (k)) = d(\mathbf{h}^{(i)}, \mathbf{h}^{(k)})$$

4.2 Discussion

As it appears while reading paragraph 3.2, the VQ method is not based on the direct comparison of statistics. It would be however the case if the classification algorithm was applied to both reference and test speech, and if the resulting codebooks were compared. Since this does not hold true, the traditional VQ method fails under certain conditions. For example, it may happen that test speech is made of many occurrences of a single vector, casually centered on one of the reference codebook entries and yielding a null distance, although the distribution of reference speech may be much more dilute than the distribution of test speech. In this case, the advantage of the comparison by the Conf method over the VQ method is that it allows to discover the discrepancy while the latter may not. The figure 2 illustrates another case where the VQ method is weak at discerning two distributions of samples while the Conf method does a good job.

Another point is that a distance computed using the classical VQ method is speaker-dependent. Hence, it generally does not fit nicely an identification task because two distances produced using two different codebooks may not

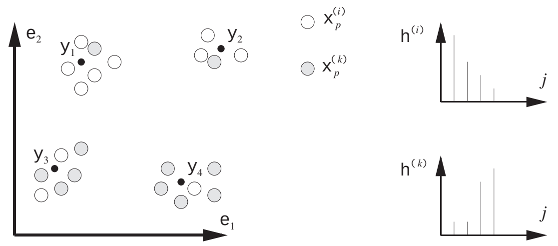

Fig. 2. Conformity method. As for the VQ method, two kinds of distance have to be computed. The first one serves the nearest neighbor rule, and the second the final distance computation. The example shows two sets of incoming speech which would produce about the same VQ distance with respect to the given codebook Y. One can see however that the frequency histograms $\mathbf{h}$ are here very different from one another and allow an easy discrimination between the two sets.

be compared without a speaker-dependent normalization step. For example, some given codebook may usually produce smaller distances than another one, although both offer exactly the same discriminating power with respect to a verification task. In this case, an identification task would associate test speech more often to the codebook producing small distances than to the codebook producing large ones. Hence, in an identification task, the advantage of the Conf method over the VQ method is that the frequency histograms are computed in a way that is not speaker-dependent.

Finally, figure 3 schematically shows how the information available within speech may be split into two parts by the vector quantization process. The classical VQ method profits by one part, while the other part is left to our Conf method. If both parts are to some extent independent with respect to their speaker identity content, then the combination of the two recognition methods may be beneficial to the global success rate.

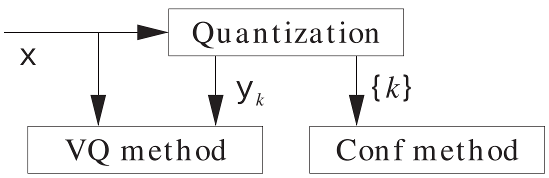

Fig. 3. The vector quantization splits incoming speech in two parts: quantized vectors $\mathbf{y}_k$ and code entries $\{k\}$. The first ones are used by the VQ method and the second ones by the Conf method.

5. Experiments

In this paragraph, we experiment the Conf method in a text-independent speaker verification task, the text independence being a natural choice because our new method is based on long-term statistics. The choice of a verification task instead of an identification one allows us to compare more easily the Conf efficiency to that of other methods, because we have seen that some of these (e. g. VQ) are not well-adapted to an identification task.

We will use $\mathbf{c}$ the complex cepstrum of the linear prediction synthesis filter as main feature and we will try four different methods, namely its mean value ($\langle \mathbf{c} \rangle$), its average vector quantization error ($\mathrm{VQ}_{\mathbf{c}}$), the average vector quantization error of its slope ($\mathrm{VQ}_{\boldsymbol{\Delta}\mathbf{c}}$) and its Conformity ($\mathrm{Conf}_{\mathbf{c}}$). We will compare three different distance computations, namely the euclidean distance (d_e), the weighted euclidean distance (d_w) and the Mahalanobis distance (d_M).

5.1 Database

Our database consists of conversational french speech obtained from radio broadcasting over three consecutive days, each one being named session till the end of this paper. The number of male and female speakers is balanced.

5.2 Methodology

In the training phase of our methodology, session I is used for building 9 representatives for each out of 10 speakers and session II is used for estimating thresholds yielding equi-error rates in a closed test verification task. In the opened test phase, session III is used for independently testing the classifiers with the aid of the thresholds previously estimated; the inter-speaker tests are done here with 12 speakers, different from those encountered in the training phase. Table I summarizes this procedure.

TABLE I
Methodology.

	I	II	III
References	Representatives	Thresholds	
Tests			Tests

This opened test methodology results in a final false-accept error rate ρ_a generally different from the false-reject error rate ρ_r; the corresponding overall quality is measured as the arithmetic mean of these two values ($\rho = \frac{1}{2}(\rho_a + \rho_r)$). We use our methodology under the very same conditions for all the methods at hand in order to allow their fair comparison. The number of tests done in assessing the error rates is quite high: about twice those found in similar studies, e. g. [7] where 3456 verification comparisons pro method are conducted, against 6120 in our case, or 12240 if one counts the computations needed by the threshold estimation step.

5.2.1 Pre-processing

Speech is cut into contiguous non-overlapping snatches of 8 s duration, without any respect to text and without pause removal. We build each representative with a pair of snatches, which corresponds to 16 s of speech. A test sample consists of a single snatch, that is, 8 s.

Speech is low-pass filtered with $f_c = 3.4$ KHz; it is then sampled with $f_s = 8.0$ KHz and quantized with $q = 16$ bit resolution. It is cut in overlapping frames of 0.030 s duration stepped each 0.010 s. After pre-emphasis with $\mu = 0.95$, each frame is multiplied by a Bartlett window and fed to linear prediction analysis (LPC) with $p = 14$ as analysis order. The resulting LPC coefficients are transformed into p cepstral coefficients which possess alleged good properties for speaker recognition when used in conjunction with (sometimes weighted) euclidean distance or with Mahalanobis distance.

5.3 Individual results

We present here the four recognition methods we mentioned above, the first three being classic and the last one new. Their results are given in table II.

TABLE II
INDIVIDUAL RESULTS.

Method	Dist.	ρ_a %	ρ_r %	$\frac{1}{2}(\rho_a+\rho_r)$ %
$\langle \mathbf{c} \rangle$	d_e	39.8	55.9	47.9
	d_w	10.4	21.9	16.2
	d_M	6.2	25.7	15.9
$\mathrm{VQ}_{\mathbf{c}}$	d_e	4.6	10.3	7.5
	d_w	3.3	8.2	5.7
$\mathrm{VQ}_{\Delta\mathbf{c}}$	d_e	42.5	25.2	33.8
$\mathrm{Conf}_{\mathbf{c}}$	d_e	9.0	10.7	9.9
	d_w	7.3	11.6	9.4
	d_M	0.3	60.9	30.6

The first method is $\langle \mathbf{c} \rangle$. It is a well known text-independent recognition method, see e. g. [7]. We compared three distances (euclidean d_e, weighted euclidean d_w and Mahalanobis d_M). The second method is $\mathrm{VQ}_{\mathbf{c}}$, another well known text-independent recognition method [4]. Here, the codebook size is $K = 32$. The third method is $\mathrm{VQ}_{\Delta\mathbf{c}}$ [4]. Here, the codebook size is again $K = 32$. We observe that our results obtained using this method are very bad; preliminary experiments discouraged us to attain a better success with the two other distances d_w and d_M. It appears then that, on our data, the $\mathrm{VQ}_{\Delta\mathbf{c}}$ method performs worse than in the case reported in [4].

Finally, we also tried on the same data our new $\mathrm{Conf}_{\mathbf{c}}$ method. As the speaker-independent codebook has to be universal, we made it bigger than for the $\mathrm{VQ}_{\mathbf{c}}$ or $\mathrm{VQ}_{\Delta\mathbf{c}}$ methods by selecting $K = 128$. We made the distance computations using d_e, d_w or d_M distances.

We see that the $\mathrm{Conf}_{\mathbf{c}}$ method scores second, after $\mathrm{VQ}_{\mathbf{c}}$ and before the classical $\langle \mathbf{c} \rangle$ method; $\mathrm{VQ}_{\Delta\mathbf{c}}$ comes last. We see also that for $\mathrm{Conf}_{\mathbf{c}}$, switching from d_e to d_w offers a small efficiency enhancement, but that the d_M distance produces bad results. In this last case, it is interesting to note the great difference between ρ_a and ρ_r; the explanation can be found in an over-learned weight matrix, due to the big dimensionality of the $\mathbf{h}$ vector. In fact, this method produces no errors at all when tested on the training data only.

5.4 Combined results

We combine here pairwise the best results of the four methods observed so far, namely $\langle \mathbf{c} \rangle$ (d_w), $\mathrm{VQ}_{\mathbf{c}}$ (d_w), $\mathrm{VQ}_{\Delta\mathbf{c}}$ (d_e) and $\mathrm{Conf}_{\mathbf{c}}$ (d_w). The joint use of all these methods is obtained through a weighted sum of the distances observed individually; the weights are chosen so that the variance contributions are equalized within each pair. We give in table III the observed results. We underlined the entries denoting the existence of an improvement over the best individual method of the considered pair ($\rho_{12} \le \min(\rho_1, \rho_2)$).

TABLE III
Average error rates of combined methods.

$\frac{1}{2}(\rho_a+\rho_r)$	$\mathrm{VQ}_{\Delta\mathbf{c}}$	$\langle \mathbf{c} \rangle$	$\mathrm{Conf}_{\mathbf{c}}$	$\mathrm{VQ}_{\mathbf{c}}$
$\mathrm{VQ}_{\Delta\mathbf{c}}$	33.8			
$\langle \mathbf{c} \rangle$	<u>13.8</u>	16.2		
$\mathrm{Conf}_{\mathbf{c}}$	11.6	10.5	9.4	
$\mathrm{VQ}_{\mathbf{c}}$	11.6	8.2	<u>5.3</u>	5.7

We observe in table III that 2/6 cases only produce an efficient combination. These two cases are the pair $(\langle \mathbf{c} \rangle, \mathrm{VQ}_{\Delta\mathbf{c}})$, which respectively emphasizes long- and short-term speech behavior, and the pair $(\mathrm{VQ}_{\mathbf{c}}, \mathrm{Conf}_{\mathbf{c}})$, where we benefit from the fact that these two last methods address different kinds of information, as expected. From all pairs, this last one shows the best results, with an error rate $\rho = 5.3\%$.

6. Conclusions

With respect to four text-independent speaker recognition methods, our new $\mathrm{Conf}_{\mathbf{c}}$ method scores second with an error rate $\rho = 9.4\%$, $\mathrm{VQ}_{\mathbf{c}}$ being first. The pairwise combinations show that the joint method $(\mathrm{Conf}_{\mathbf{c}}, \mathrm{VQ}_{\mathbf{c}})$ performs better than $\mathrm{Conf}_{\mathbf{c}}$ or $\mathrm{VQ}_{\mathbf{c}}$ alone, the associated error rate being $\rho = 5.3\%$. Finally, our results show that the Conf method is not only efficient by itself, but also that it combines favorably with VQ.

References

[1] S. Furui, *Comparison of Speaker Recognition Methods Using Statistical Features and Dynamic Features*, IEEE Trans. ASSP, Vol. 29, No. 3, 1981, pp. 197–200.

[2] K. P. Li, E. H. Wrench Jr., *An Approach to Text-Independent Speaker Recognition with Short Utterances*, Proc. ICASSP, Boston, 1983, pp. 555–558.

[3] A. E. Rosenberg, F. K. Soong, *Evaluation of a Vector Quantization Talker Recognition System in Text-Independent and Text-Dependent Modes*, Proc. ICASSP, Tokyo, 1986, pp. 873–876.

[4] F. K. Soong, A. E. Rosenberg, *On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition*, Proc. ICASSP, Tokyo, 1986, pp. 877–880.

[5] B.-H. Juang, F. K. Soong, *On the Speaker Recognition Based on Source Coding Approaches*, Proc. ICASSP, Albuquerque, 1990, pp. 613–616.

[6] M.-S. Chen, P.-H. Lin, H.-S. Wang, *Speaker Identification Based on a Matrix Quantization Method*, IEEE Trans. ASSP, Vol. 41, No. 1, 1993, pp. 398–403.

[7] M. Shridar, N. Mohankrishnan *Text-Independent Speaker Recognition: A Review and Some New Results*, Speech Comm., Vol. 1, No. 3-4, 1982, pp. 257–267.