SPEAKER MODELS DESIGNED FROM COMPLETE DATA SETS: A NEW APPROACH TO TEXT-INDEPENDENT SPEAKER VERIFICATION

D. R. Dersch[1] *and R. W. King*[2]
[1]Speech Technology Research Group
Department of Electrical Engineering, The University of Sydney, NSW 2006
[2]Faculty of Information Technology, University of South Australia, SA 5095
Email: dersch@speech.su.oz.au, robin.king@UniSA.edu.au

ABSTRACT

In this paper we present a new approach to text independent speaker verification. Speaker models are created from *complete* data sets, derived from a set of sentences. A decision on an identity claim is based on the calculation of the mean next neighbour distance between a speaker model and a test utterance. A Vector quantization technique serves to efficiently extract this frame based similarity measure. It is the purpose of this paper to investigate this new approach and test its performance on a large database as a function of a number of parameters, i.e., the number of data vectors in each model and the length of the test utterance. The best results on a set of 108 speakers are 0.93% false rejection rate and 0.98% false acceptance rate.

1. INTRODUCTION

Speaker recognition and verification are challenging problems in signal processing. A number of different approaches have been proposed to tackle these tasks, for example Hidden Marcov Models, Gaussian Mixture Models or Vector quantization (Vq) procedures, see e.g., [6], [9], [11] and [5]. All these model based approaches perform a data reduction in the first step, as the complexity of the models is much smaller than the complexity of the data set. We argue that the data reduction aiming at a reduction of the complexity also entails a loss of information, which might be relevant for the verification task.

We have reported previously on a new approach [2] which yields promising results for text-independent speaker verification as well as for various other problems in the fields of automatic speech processing, i.e. text-independent speaker recognition and accent classification. It is the purpose of this paper to carry out a series of text-independent speaker verification experiments to further investigate this new approach.

In our approach a speaker model is formed from a complete data set of sequences of 12 dimensional mel-frequency cepstrum coefficients (mfcc) derived from a set of utterances. Each speaker is represented by a 12 dimensional data set in the mfcc space. Speaker recognition is carried out by calculating the mean next neighbour distance in the 12 dimensional mfcc space between data vectors of a transformed test utterance and a speaker model. The identity claim is accepted, if the mean next neighbour distance falls below a global threshold obtained in a training process.

The search for the next neighbour of a test data vector amongst the data vectors of a speaker model involves a large computational effort. In order to reduce this effort we apply a hierachical neural Vq technique [1]. Each codebook vector is associated with a subset of the data set. The restriction of the search for the next neighbour to a limited number of subsets allows a drastic reduction of the computational effort. However, the reduced search might affect the overall performance of the system. It is the purpose of this paper to thoroughly study this approach and answer the following questions: How does the reduced search for the next neighbour affects the performance? How does the number of data vectors in the test utterance and in the model affect the performance?

The model design and the feature extraction are presented in the next section. In Section 3 the method is applied to text-independent speaker verification. The paper concludes with a summary and discussion.

2. MODEL DESIGN

"The best model for a data set is the data set". According to this statement we design a model for a data set $X \subset \mathcal{R}^n$, with $X \equiv \{\mathbf{x}_i \in \mathcal{R}^n \mid i = 1, \ldots, N_x\}$ by itself, where N_x is the number of frames.

Pattern recognition or classification requires the extraction of features from a test data set $Y \equiv \{\mathbf{y}_k \in \mathcal{R}^n \mid k = 1, \ldots, N_y\}$. Here, we consider the special case $N_x \gg N_y$. Feature extraction is performed by calculating the mean next neighbour distance

$$d_{nn} = \langle ||\mathbf{x}'(\mathbf{y}) - \mathbf{y}|| \rangle_y. \quad (1)$$

between each data vector $\mathbf{y}$ and its next neighbour

$\mathbf{x}'(\mathbf{y})$ in X. $\langle\cdots\rangle_y$ denotes an average over all data vectors in Y.

The search for the next neighbour according to Eq. (1) is a very time consuming procedure. However, a large number of data vectors, far away from $\mathbf{y}$, might reasonably be excluded from the searching process. A codebook obtained by a Vector quantization procedure might serve to limit the search for $\mathbf{x}'(\mathbf{y})$ and thus significantly reduce the computational effort.

Vq (see e.g. [4]) addresses the problem of mapping a data space $X \subset \mathcal{R}^n$ onto a finite set of N ($N_x \gg N$) codebook vectors $\mathbf{w}_r \in W \equiv \{\mathbf{w}_r \in \mathcal{R}^n \mid r = 1,\ldots,N\}$. Each data vector $\mathbf{x} \in X$ might be assigned to the next codebook vector $\mathbf{w}_{r\prime} \in W$ by the condition

$$\|\mathbf{x} - \mathbf{w}_{r\prime}\| = \min_r \|\mathbf{x} - \mathbf{w}_r\|. \qquad (2)$$

A codebook W therefore defines a partition of the data space, a so-called *Voronoi tessellation*. All data vectors within a Voronoi cell are associated to one codebook vector.

In order to reduce the computational effort to calculate d_{nn} we might limit the search for the next neighbour to *one* Voronoi cell. First $\mathbf{w}_{r\prime}$ is obtained for each data vector $\mathbf{y}$ similar to Eq. (2). Then the Voronoi cell of $\mathbf{w}_{r\prime}$ is searched for the next neighbour $\tilde{\mathbf{x}}'(\mathbf{y})$. Equation (1) then changes to

$$\tilde{d}_{nn} = \langle\|\tilde{\mathbf{x}}'(\mathbf{y}) - \mathbf{y}\|\rangle_y. \qquad (3)$$

Equation (3) entails a speed up by a factor of approximately N as compared to Eq. (1). A further reduction might be achieved by hierarchically structured codebooks as proposed in [3].

However, d_{nn} and $\tilde{d}_{nn}$ might slightly differ. Whenever data vectors $\mathbf{x}'(\mathbf{y})$ are not in the subsets of $\mathbf{w}_{r\prime}$ $\tilde{d}_{nn}$ is an over estimation of the mean next neighbour distance. To remedy the weakness of an over estimation we recently proposed a framework that allows a extension of the search space to a *certain* number of Voronoi cells surrounding $\mathbf{y}$ (see e.g. [3]) by means of a fuzzy Vq procedure [10, 1]. This approach allows to trade off the accuracy of the calculation of d_{nn} and the computational effort.

3. TEXT INDEPENDENT SPEAKER VERIFICATION EXPERIMENTS

In the following section we apply the described method to text-independent speaker verification. The results are based on a new set of experiments which are more thoroughly discussed than the results previously reported [2]. Here, we focus on the following questions: How does the reduced search for the next neighbour affect the performance? How does the number of data vectors in the test utterance and in the model affect the performance?

3.1. Speech coding

The speech material is a subset of 108 native Australian English speakers taken from the Australian National Database Of Spoken Language (ANDOSL), see e.g. [7]. The database comprises a set 200 phonetically rich, read sentences for each speaker. The sound pressure signal sampled at a rate of 20 kHz is parameterized by 12 mel-frequency cepstrum coefficients by applying a Hamming window of 16 msec duration and 5 msec step size. The mfcc spectrum is pre-emphasised by a filter coefficient of 0.97. A silence detection is performed by cutting of frames below a threshold of 0.1 of the normalised log energy. The experiments are performed on a set of 54 male and 54 female speakers.

3.2. Speaker models and experimental setup

To train and test the system we selected for each speaker three different sets of sentences: one set to create the model, one to adjust the threshold and a third to test the speaker verification system. For each speaker these three different sets are randomly selected from the database of 200 sentences. This setup ensures, that the verification experiments are text independent.

From the first set of sentences we designed speaker models of three different sizes, from five, twenty and forty sentences. The average number of data vectors in each of these models is $3,422$, $13,858$ and $25,671$, corresponding to 17.11, 69.3 and 128.4 seconds recorded speech after cutting off silence. Each model is constructed with a codebook of 10 codebook vectors. The models consisting of the codebook and the data set are stored in a binary data structure. The necessary disc space for each speaker model is 0.18, 0.72 and 1.33 MB, corresponding to five, twenty and forty sentences in the model.

We used the second set of sentences to calculate the global threshold for the similarity measure to decide about the identity claim of a speaker. For each speaker eight imposters where randomly selected. For different sentences of these imposters and of the true speaker the mean next neighbour distance to the true speaker model is obtained. The global threshold is identified with that value of the mean next neighbour distance where the false rejection rate (frr) is equal to the false acceptance rate (far). This error is the so-called equal error rate (eer).

The performance of the system is tested on the third set of sentences using the remaining set of imposters and the true speaker. We tested utterances with a length of one and two sentences. In order to answer the question of how the reduced search for the next neighbour affects the performance, we consider the two limiting cases: a full search for the next neighbour according to Eq. (1) and a reduced search within only

one Voronoi cell according to Eq. (3). The results are summarized in the next section.

3.3. Results

The text-independent speaker verification experiments are carried out for three different speaker groups: 54 male speakers, 54 female speakers and a group containing both genders (108 speakers). The total number of imposter and true speaker attempts in each of these three groups which are used to obtain the global threshold are summarized in Table 1. The number of attempts in the test set are shown in Table 2. In the following we first report on the

speaker set	imp.	true	sum
male/female	2,160	540	2,700
male + female	4,320	1,080	5,400

Table: 1: Total number of imposter and true speaker attempts to obtain the global threshold for the group of male (female) speaker (row 1) and the group consisting of both genders (row 2).

speaker set	imp.	true	sum
male/female	12,150	270	12,420
male + female	53,460	540	54,000

Table: 2: Total number of imposter and true speaker attempts in the test set for the group of male (female) speaker (row 1) and the group consisting of both genders (row 2).

results for the five-sentence-speaker-models. Tables 3 and 4 summarize the results for a complete and a reduced next neighbour search. The decision in each experiment is based on one test sentence. The errors are given in percent.

speaker set	eer	frr	far	mean_{test}
male	1.53	1.48	2.23	1.85
female	5.00	3.70	2.89	3.30
male + female	1.62	0.74	2.35	1.50

Table: 3: Results of text-independent speaker verification for speaker models designed from five sentences: equal error rate (eer) derived from the training set, false rejection rate (frr), false acceptance rate (far) on the test set and the mean value of frr and far (mean_{test}). The decision is based on one sentence. The next neighbour search is performed on the complete data set.

A comparison of the eer shows only a small difference between the full and the reduced search, whereas the mean_{test} is slightly higher for the reduced search. The computational effort for the reduced search is approximately ten times smaller than for the full search.

Table 5 shows the results for a reduced next neighbour search and decisions based on two sentences. Table

speaker set	eer	frr	far	mean_{test}
male	1.44	1.85	2.10	1.98
female	5.05	4.44	3.05	3.75
male + female	1.67	1.48	2.31	1.90

Table: 4: The same setup as above, but the next neighbour search is limited to one Voronoi cell.

speaker set	eer	frr	far	mean_{test}
male	0.79	0.37	1.04	0.71
female	2.96	3.70	1.14	2.42
male + female	0.78	0.93	0.98	0.96

Table: 5: The same setup than in the previous Table, but here a speaker attempt comprises two sentences.

speaker set	eer	frr	far	mean_{test}
male	0.65	0.74	0.74	0.74
female	2.36	1.85	0.91	1.38
male + female	0.76	0.56	1.26	0.91

Table: 6: Results for speaker models designed from twenty sentences, a next neighbour search limited to one Voronoi cell and one test sentence.

5 clearly reveals that the increase in the length of the test utterance also results in a lower eer, frr and far. A similar effect is observed if we increase the number of data vectors per model. Table 6 shows the results for the twenty-sentence-models. A comparison of Table 5 and 6 shows, similar results for the set of the male and the mixed gender speakers, whereas the mean values of far and frr differ about one percent for the female speakers.

In order to determine, if a further increase in the number of data vectors per model entails a further increase in the performance, we trained models for ten randomly selected male speakers from forty sentences. Again, we randomly selected eight imposters for each speaker to obtain a global threshold. The system is tested on the remaining set of male speakers. The total number of imposter and true speaker attempts in the training and the test set is summarized in Table 7. A comparison of the forty-sentence-models with the twenty-sentence-models on the same speaker subset is shown in Table 8. As a result we find that a further increase in the total number of data vectors per model yields a smaller eer and a smaller far. In this example

speaker set	imp.	true	sum
train (male)	400	100	500
test (male)	2,250	50	2,300

Table: 7: Total number of imposter and true speaker attempts in the training set for the group of training (row 1) and the test speaker (row 2) for the set of ten male speaker models.

speaker set	eer	frr	far	mean_{test}
male (40 sen)	0.75	0.00	0.13	0.07
male (20 sen)	1.25	0.00	0.71	0.36

Table: 8: Results of text-independent speaker verification for speaker models designed with forty sentences (row 1) and twenty sentences (row 2) for ten different speakers. The next neighbour search is limited to one Voronoi cell. The decision on the speaker attempt is based on one sentence.

the frr is 0.0% for both experiments.

Generally we find that in each experiment the error rate for the female speaker is much higher than for the male speaker. A similar result has been observed for speaker recognition experiments [3] as well as for different databases [9]. A comparison of the mixed gender experiments with the male speaker shows comparable error rates. A closer investigation reveals that this effect is due to a very low cross gender (male-female, female-male) confusion.

4. SUMMARY AND DISCUSSION

In this paper we presented a new approach to speaker verification. Here, models are designed from *complete* data sets. Pattern matching is performed by calculating similarity measures between data sets. In our approach we use a Vq process to allow an efficient calculation of the similarity measure, rather than to reduce the complexity of the data set.

We showed that the reduced search for the next neighbour causes only a small degradation in the performance and that an increase in the total number of data vectors in a model and in the test utterance reduces the error rates.

The quality of a speaker verification system depends on a number of parameters such as the error rates, the number of speakers, the database and its characteristics and the computational effort for training and testing. Therefore, a comparison of different systems has to be carried out very carefully and would extend the frame of this paper. For a closer discussion about the comparison of different speaker verification and recognition systems see [8]. However, if we only consider the number of speakers and the accuracy of the system then we find that our approach outperforms the approach described in [5]. Reynolds [9] reported a mean_{test} in the range of $0.24-0.5\%$ on a subset of the TIMIT database (112 male and 56 female speakers). For a more thorough comparison of our approach with state of the art systems we are currently repeating our experiments on benchmark databases for these tasks.

The training and testing of our system is very simple and efficient. The calculation of the mean next neighbour distance for one test sentence and a five-sentence-model takes about 3.2 sec on a SUN SPARC 10 with 64 MB RAM. We use a Manhattan Metric rather than an Euclidian distance measure to calculate the distances in Eqs. (1-3). For an optimal choice of the codebook size with respect to the number of data vectors the calculation of the distance to the next neighbour involves only $4n\sqrt{N_x}$ additions and substructions per test frame, where N_x is the number of data vectors and n the dimension. For hierachies of codebooks as proposed in [3] the number of operations is even smaller.

REFERENCES

[1] D. R. Dersch and P. Tavan *"Control of annealing in minimal free energy vector quantization"*, Proc. ICNN-94, pp. 698–703, Orlando, 1994.

[2] D. R. Dersch, *"The acoustic fingerprint: A method for speaker identification, speaker verification and accent identification"*, Proc. SST-96, pp. 307–312, Canberra, 1996.

[3] D. R. Dersch, *"Feature extraction from complete data sets: A new approach to pattern recognition and its application to text-independent speaker identification"*. ACNN-97, forthcoming, Melbourne, 1997.

[4] R. M. Gray, *"Vector quantization"*, IEEE ASSP Magazine, vol. 1, pp. 4–29, 1984.

[5] A. L. Higgins and L. G. Bahler, *"Text-independent speaker verification by discriminator counting"*, Proc. ICASSP-91 pp. 405–408, 1991.

[6] T. Matsui and S. Furui, *"Comparison of Text Independent Speaker Recognition Methods Using VQ-distortion and Discrete/Continuous HMM's"*, Proc. ICASSP-92, pp. 157-160, 1992.

[7] B. Millar, J. Vonwiller J. Harrington, and P. Dermody, *"The Australian National Database of Spoken Language"*, ICASSP-94, pp. 67–100, Adelaide, 1994.

[8] J. Oglesby, "What's in a number? Moving beyond the equal error rate", *Speech Com.*, vol. 17, pp. 193-208, 1995.

[9] D. A. Reynolds, "Speaker identification and verification using Gaussian mixuter speaker models", *Speech Com.*, vol. 17, pp. 91-108, 1995.

[10] K. Rose, E. Gurewitz and G. Fox *"Statistical mechanics and phase transitions in clustering"*, Phys. Rev. Lett., vol. 65, pp. 945–948, 1990.

[11] F. K. S. Soong, A. E. Rosenberg, L. R. Rabiner and B. H. Juang, *"A Vector Quantization Approach to Speaker Recognition"*, Proc. ICASSP-85, pp. 387–390, 1985.