THE ACOUSTIC FINGERPRINT: A METHOD FOR SPEAKER IDENTIFICATION, SPEAKER VERIFICATION AND ACCENT IDENTIFICATION

D. R. Dersch
Speech Technology Research Group
Department of Electrical Engineering
Building J03, The University of Sydney
NSW, 2006, Australia
Phone +61-2-93514509
email: dersch@speech.su.oz.au

ABSTRACT – We present a new parameter free framework based on similarity measures between data spaces and demonstrate its performance for different tasks in the field of automatic speech processing, e.g., for text-independent speaker identification, speaker verification and accent identification. The speech data are taken from the ANDOSL database. The speech signal is coded by 12 mel-frequency cepstrum coefficients (mfcc). The identification and verification task is performed by calculating similarity measures between data spaces occupied by utterances in the 12-dimensional mfcc-space. In order to reduce the computation effort, we apply a Vector Quantization technique. On a set of 108 speaker we achieve an accuracy of 100% for text-independent speaker identification, 100% successful acceptances and 99.81% successful rejections for text-independent speaker verification, respectively. First results for accent identification yields an accuracy of 72.3 - 74.5% for the discrimination of two different Australian accents, Lebanese Arabic and South Vietnamese.

1. INTRODUCTION

Speaker identification and verification is an challenging problem to automatic speech processing. Many years of research in this area is reflected in a great number of publications. A list of more than three hundred fifty publications related to speaker verification and identification has been compiled by Chollet (1995). A number of approaches make use of Vector Quantization (VQ) techniques (Gray, 1984) as a first step to reduce the complexity of the preprocessed speech signal, see e.g., Soong et al. (1985), Gong and Haton, (1990) and He et al. (1995).
We argue that Vector Quantization applied in order to reduce the complexity of a data set also entails a loss of information, which is relevant for classification. It is the purpose of this paper to present a new pattern matching approach, which performs a feature extraction on a *complete* data set. Feature extraction is carried out by extracting similarity measures between data sets. In our approach a fuzzy VQ procedure (Rose et al. 1990 and Dersch and Tavan 1994) has a key function to reduce the computational effort, rather than the complexity of a data set.
Figure 1 shows the overall design of the system. First, the speech signal is transformed into a stream of vectors comprising 12 mel-frequency cepstral coefficients.

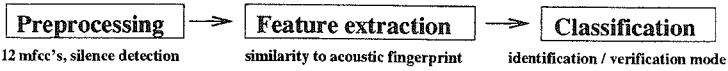

Figure 1: Speaker identification and verification system

The speech material and the relevant preprocessing steps are presented in the next Section. The feature extraction based on similarity measures between an utterance and a given data set is performed in the second step. An efficient implementation of the feature extraction makes use of a combination of a codebook *and* the underlying data distribution. In the following we call the combination of the codebook and the data set the *acoustic fingerprint*. In Section 3 we show, how such a acoustic

fingerprint is created. Classification is carried out in the third step. In the speaker identification mode similarities are calculated, comparing the acoustic fingerprints of a set of speakers with a subspace occupied by an utterance. The utterance is then associated with the speaker with the highest similarity measure. In the verification experiments a global threshold for the similarity measure is obtained by a set of training utterances. The identity claim is accepted if the similarity measure falls under the threshold. Results on text-independent speaker identification and verification on a set of 108 speakers (54 male and 54 female) are summarized in Section 4 and 5. First results on accent identification are presented in Section 6. Conclusions are summarized in Section 7.

2. SPEECH CODING

The speech material is taken from the Australian National Database Of Spoken Language (ANDOSL) see e.g. Vonwiller et al., 1995. The database comprises speech samples from native and non-native Australian English speakers. From each speaker 200 phonetically rich sentences are recorded. 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 stepsize. 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 speaker identification experiments are performed on a set of 54 male and 54 female native Australian English speakers. A training set of another three male native Australian English speakers is used in the speaker verification experiments. Accent identification is carried out for Lebanese Arabic and South Vietnamese Australian English accents. Here, we take a training and test set of 8 and 11 speakers from each accent group.

3. FEATURE EXTRACTION AND THE ACOUSTIC FINGERPRINT

"The best model for a data set is the data set". According to that phrase we occupy a subspace $X \subset \mathcal{R}^{12}$, with $X \equiv \{\mathbf{x}_i \in \mathcal{R}^{12} \mid i = 1, \ldots, N_x\}$ within the 12-dimensional mfcc space by a set of sentences. The data set is considered to be large enough that characteristic features of a speaker or an accent are coded in the local distribution of data vectors. The dark shaded area in Figure 2 show a two dimensional caricature of such a characteristic subspace X. The light shaded area illustrates a subspace $Y \equiv \{\mathbf{y}_k \in \mathcal{R}^{12} \mid k = 1, \ldots, N_y\}$ occupied by an utterance, for example by a sentence, with $N_x \gg N_y$.

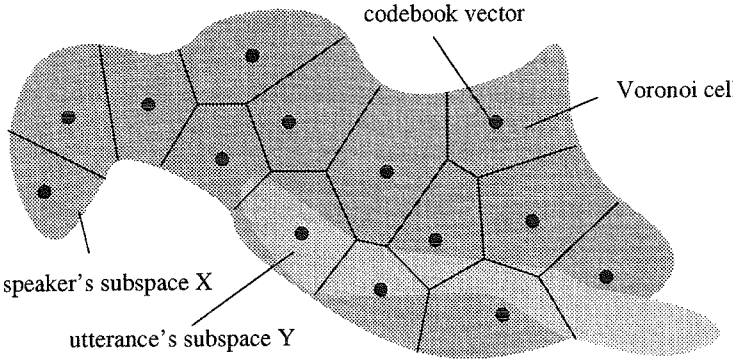

Figure 2: The acoustic fingerprint

Feature extraction is performed by calculation a similarity measure between the subspaces X and Y. A promising similarity measure is the mean next neighbour distance between each data vector $\mathbf{y}$ of an utterance and data vectors $\mathbf{x}$ in X.

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

Here, $\mathbf{y}$ is a data vector of an utterance and $\mathbf{x}'(\mathbf{y})$ is the next neighbour in the characteristic subspace X. $\langle\cdots\rangle_y$ denotes an average over all data vectors of the utterance. The search for the next neighbour of $\mathbf{y}$ is a very time consuming procedure. However, a large number of data vectors, far away from $\mathbf{y}$, might be excluded from the searching process. In the following we want to show how a codebook obtained by a fuzzy Vector Quantization (VQ) procedure serves to limit the search space for $\mathbf{x}'(\mathbf{y})$ and thus significantly reduces the computational effort.

Vector Quantization addresses the problem of mapping a data space $X \subset \mathcal{R}^n$ onto a finite set of N codebook vectors $\mathbf{w}_r \in W \equiv \{\mathbf{w}_r \in \mathcal{R}^n \mid r = 1, \ldots, N\}$. Each data vectors $\mathbf{x} \in X$ might be assigned to the next codebook vector $\mathbf{w}_{r\prime} \in W$ by the condition

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

Thus, a codebook W defines a partition of the data space – a so-called *Voronoi tessellation*. All data vectors within a Voronoi cell are associated to just one codebook vector, which is the next. Figure 2 shows a codebook and the corresponding Voronoi tessellation for the two dimensional characteristic subspace X.

In order to speed up the calculation of d_{nn} we limit the search for the next neighbour to *one* Voronoi cell. We first obtain $\mathbf{w}_{r\prime}$ for each data vector $\mathbf{y}$ similar to Equation (2). We then search within the Voronoi cell of $\mathbf{w}_{r\prime}$ for the next neighbour $\tilde{\mathbf{x}}'(\mathbf{y})$. Equation (1) changes to

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

The use of Equation (3) entails a speed up by a factor of approximately N as compared to Equation (1). For a further speed up we use a Manhattan metric, rather than an Euclidian distance measure in Eqn. (3), to reduce the number of numerical operations.

However, d_{nn} and $\tilde{d}_{nn}$ might slightly differ. Whenever $\mathbf{x}'(\mathbf{y})$ is not in the subset of $\mathbf{w}_{r\prime}$, which might occur for data vectors at the borders of a Voronoi cell, $\tilde{d}_{nn}$ is an over estimation of the mean next neighbour distance $\tilde{d}_{nn} = (1+\epsilon)d_{nn}$, with $0 \leq \epsilon \ll 1$. The over estimation becomes more pronounced for large codebooks and high dimensional data spaces, as it is more likely that a data vector is next a border. To remedy that weakness we might expand the search to a certain number of Voronoi cells surrounding $\mathbf{y}$. The fuzzy VQ procedure actually permits the extension of the search space. Here, a data vector $\mathbf{y}$ is assigned to codebook vectors $\mathbf{w}_r$ according to a *conditional probability* $p(r|\mathbf{y})$. We might expand the search for $\mathbf{x}'(\mathbf{y})$ to Voronoi cells assigned to the set of codebook vectors $[\mathbf{w}_{r(1)}, \cdots, \mathbf{w}_{r(k)}]$ with

$$\sum_{i=1}^{k} p(r[i]|\mathbf{y}) \geq p_{sum}. \tag{4}$$

Here, $p_{sum} \leq 1.0$ is the cumulative probability of the sorted list of conditional probabilities $p(r[i]|\mathbf{y}) \geq p(r[i+1]|\mathbf{y})$, with $i \in [1, \cdots, k]$. Equation (4) allows to expand the search space from one Voronoi cell ($k = 1$ or $p_{sum} = \delta$, with $0 < \delta \ll 1$) to the whole data set ($k = N$ or $p_{sum} = 1.0$), either by the choice of k or p_{sum}. Therefore, the fuzzy VQ procedure allows to optimize the calculation of $\tilde{d}_{nn}$ with respect to accuracy and performance.

The fuzzy VQ scheme entails a tessellation into Voronoi cells with equal number of data vectors. The conditions for this property – we call it *load-balance*– are discussed by Dersch and Tavan (1994). A search within load balanced sets is actually the optimal way with respect to performance, which is another advantage of this scheme.

We call the combination of a data set coding characteristic features and a codebook, which enables an effective feature extraction the *acoustic fingerprint*. In the next Sections we apply the presented scheme to speaker identification, speaker verification and accent identification.

4. SPEAKER IDENTIFICATION

The Speaker identification experiments are performed on a set of 54 male and 54 female speakers. For each speaker a acoustic fingerprint is calculated from 20 sentences. For each speaker we choose a different subset of sentences in the acoustic fingerprint. The verification sentences for each speaker are always different from the sentences in it's own fingerprint, but they might occur in a fingerprint of another speaker. This setup ensures that the experiments are text-independent. It is important to note, that this setup is even "unfair" for the correct speaker, as other speakers might have the advantage of uttering a sentence included in the acoustic fingerprint of the correct speaker.
The mean number of data vectors in each acoustic fingerprint is 13858, corresponding to 69.3 sec recorded speech without silence. A codebook of 10 codebook vectors is used to point at subsets of the dataspace. Only one Voronoi cell is searched for the next neighbour $\mathrm{x}'(\mathrm{y})$.
Decisions about the identity of a speaker are made after a fraction of a sentence, e.g., the first quarter, half, three quarters and the full sentence, corresponding to a mean duration of $0.87, 1.75, 2.62$ and 3.5 seconds, after cutting off silence. The similarity measure Eqn. (3) is calculated between the sentence fraction and the acoustic fingerprint of each speaker. A segment is then associated to the speaker with the lowest value. Five different sentences are tested for each speaker. The total number of decisions is 2160. The identification results are summarized in Table 1.

fraction of sentence	0.25	0.5	0.75	1.0
mean duration [sec]	0.87	1.75	2.62	3.5
accuracy [%]	97.04	100.00	100.00	100.00
$\Delta\tilde{d}_{nn}$[%]	13.80	16.98	18.07	18.48

Table 1: Results of text-independent speaker identification: duration, accuracy and performance as a function of the processed sentence fraction.

In order to evaluate the quality of the results we define a performance measure

$$\Delta\tilde{d}_{nn} = \langle \frac{\tilde{d}_{nn}(J=2)}{\tilde{d}_{nn}(J=1)} - 1 \rangle_s. \tag{5}$$

Here $\tilde{d}_{nn}(J=1)$ is the shortest mean next neighbour distance and $\tilde{d}_{nn}(J=2)$ the second shortest, corresponding to highest and second highest similarity between a fingerprint and a given segment. $\langle \ldots \rangle_s$ denotes a mean value for segments of similar length s. Table 1 shows, that all 540 sentence fractions of length 0.5 and longer are correct identified, whereas for fractions of length 0.25 an accuracy of 97.04% is achieved, corresponding to 16 wrong decisions. The performance measure increases with increasing segment length from 13.80% to 18.48%.

5. SPEAKER VERIFICATION

Speaker verification experiments are carried out with an acoustic fingerprint comprising 20 and 40 sentences. The experiments are similar to the one described above. Only complete sentences are used to reject or accept a identity claim of a speaker. A global threshold for the similarity measure is obtained by a set of training sentences. Here, 10 training sentences are taken from each speaker and 10 sentences from three male speakers different from the set of 108 speakers. The identity claim is accepted, if the similarity measure falls under the threshold.

threshold	3.9	3.95	**4.0**	**4.05**	**4.1**	4.15	4.2	4.3	4.4
correct acceptances [%]	94.44	96.67	**98.06**	**98.89**	**99.17**	99.72	99.81	99.91	100
correct rejections [%]	100	99.97	**99.94**	**99.85**	**99.46**	99.20	98.74	97.41	95.25

Table 2: roc for speaker verification experiments on the training set.

The receiver operator characteristics (roc) for the speaker verification experiments on the training set are summarized in Table 2. According to the demand – high correct acceptances or high correct rejections – one might choose a global threshold in the range of 4.0 - 4.1. In the test phase we process 5 sentences from each speaker through all 108 fingerprints. The 58320 decisions subdivide in 540 acceptances and 57780 rejections. Table 3 shows the roc for the speaker verification experiments on the test set. For a threshold in the range of 4.0 - 4.1 we achieve 98.52 - 99.26% correct acceptances and 99.60 - 98.91% correct rejections.

threshold	3.9	3.95	4.0	4.05	4.1	4.15	4.2	4.3	4.4
correct acceptances [%]	95.00	97.78	**98.52**	**98.89**	**99.26**	99.44	100	100	100
correct rejections [%]	99.89	99.77	**99.60**	**99.34**	**98.91**	98.33	97.60	95.42	92.37

Table 3: roc for speaker verification experiments on the test set as a function of a global threshold for the mean next neighbour distance. For each of the 108 speakers 5 test sentences are processed through each of the 108 acoustic fingerprints comprising 20 sentences.

In order to check whether an increase in the total number of sentences in the acoustic fingerprint and an extended search for the next neighbour improves the classification results of our system, we carried out a second set of speaker verification experiments. Here, the acoustic fingerprint comprises 40 sentences and a codebook of 20 codebook vectors. The search for the next neighbour is performed on four Voronoi cells. We considered a set of 10 different male speakers. The training and testing is performed as described above. Table 4 shows the roc on the training set.

threshold	3.55	3.60	3.65	3.7
correct acceptances [%]	98.00	**100**	**100**	100
correct rejections [%]	100	**100**	**100**	99.67

Table 4: roc for speaker verification experiments on the training set with a acoustic fingerprint of 40 sentences.

According to the demand one might choose a global threshold in the range of 3.60 - 3.65. 5 test sentences from all 108 speakers are processed through the 10 fingerprints, corresponding to 50 acceptances and 5350 rejections. Table 5 summarizes the results.

threshold	3.45	**3.5**	3.55	**3.60**	**3.65**	3.7
correct acceptances [%]	94.00	**100**	100	**100**	**100**	100.0
correct rejections [%]	99.98	**99.98**	99.88	**99.81**	**99.73**	99.55

Table 5: roc for speaker verification experiments on the test set for 10 different male speakers verified on a set of 108 speakers.

The results for a threshold in the range of 3.60 - 3.65 are 100% correct acceptances and 99.81 - 99.73% correct rejections. The best results on the test set is achieved at a threshold slightly below the chosen values, at 3.55 with 100% correct acceptances and 99.98% correct rejections, respectively.

6. ACCENT IDENTIFICATION

In this Section we report first results on accent identification. We consider two different Australian accent groups, Lebanese Arabic and South Vietnamese. The fingerprints are created by 30 different sentences from 8 different speakers and a codebook of 100 codebook vectors. The fingerprint of the Lebanese Arabic accent group comprises 206802 data vectors, the one from the South Vietnamese 197970. The test is performed by single sentences from 11 Lebanese Arabic and 11 South Vietnamese speakers. From each speaker 5 different sentences are tested. For a different number of similarity measures we achieve an overall accuracy of 72.3 - 74.5%.

7. SUMMARY AND CONCLUSIONS

In this paper we presented a new parameter free framework called the acoustic fingerprint. We successfully applied the method to several task in the field of automatic speech processing, e.g., for text-independent speaker identification, speaker verification and accent identification. As an important feature no labelling of the data is necessary. The first results on accent identification are yet not as good as those published by Kumpf and King (1996). We try to further improve the performance by adding other relevant features to the mfcc space. The use of only two acoustic fingerprints, one for each accent, might cause some kind of smearing effect. The use of single speaker fingerprints, similar to the speaker verification and classification experiments might perform better. In the speaker identification results we achieved 100% correct classification on a set of 108 speakers. In the speaker verification results we obtain 98.52 - 99.26% correct acceptances and 99.60 - 98.91% correct rejections for an acoustic fingerprint of 20 sentences and 100% successful acceptances and 99.81% successful rejections for an acoustic fingerprint of 40 sentences. It has been shown, that an increase in the number of data vectors in the acoustic fingerprint and an extended search for the next neighbour entails an increase in the performance. The setup for text-independent experiments was chosen in a way to disadvantage the correct speaker, as other speakers might utter a sentence included in the acoustic fingerprint of the correct speaker which makes the problem harder. In our approach we used a global threshold for all speaker, rather an individual, which might perform better. One can argue, that the proposed scheme is time consuming and needs a lot of disc space. However, the feature extraction for one sentences from a 20 sentence acoustic fingerprint takes about 12.8 sec on a SUN SPARC 10 with 64 MB. The occupied disc space for such an acoustic fingerprint is 0.72 MB. The data structure for the acoustic fingerprint allows both, fast I/O and minimizing the used disc space. A further speed up might be achieved by enlarging the codebook or using a hierarchy of codebooks. Higgins et al. (1994) used another approach to extract similarity measures between data spaces. But they used a method to discard "redundant" data vectors, which is actually some kind of agglomerative clustering procedure (Duda and Hart, 1973). Therefore they compare codebooks rather than data sets. The presented results are hard to compare with other approaches, as the performance of a system, strongly depends on the database.

Acknowledgments: The author is grateful to Prof. Robin King for constructive comments on this paper.

8. REFERENCES

Chollet G. (1995) *Long Index of References on Automatic Speaker Verification*, http://sig.enst.fr/ chollet/ForMehdi/SpRecV1.Lind.html

Dersch D. R.and Tavan P. (1994) *Load balanced vector quantization*, Proceedings of the International Conference on Artificial Neural Networks, ICANN 94, pages 1067-1070, Springer London.

Dersch D. R.and Tavan P. (1994) *Control of annealing in minimal free energy vector quantization* Proceedings of the IEEE International Conference on Neural Networks ICNN 94, pages 698-703.

Duda R. O. and Hart P. E. (1973) *Pattern Classification and Scene Analysis*, Wiley, New York.

Gong Y. and Haton J. (1990) *Text-independent speaker recognition by trajectory space comparison*, Proceedings ICASSP 90, pages 285-288

Gray, R. M. (1984) Vector quantization, IEEE ASSP Magazine 1, No. 2:4-29.

He, J. and Liu, L. and Palm, G. (1995) *On the use of Features from Prediction Residual Signals in Speaker Identification*, Proceedings EUROSPEECH 95, pages 313-316,

Higgins, A. L. and Bahler, L. G. and Porter, J. E. (1994) *Voice Identification Using Nearest-Neighbor Distance Measure*, Proceedings ICASSP 93, pages 375-378

Kumpf, K., King, R.W. (1996) *Automatic Accent Classification of Foreign Accented Australian English Speech*, Proceedings ICSLP 1996, forthcoming.

Rose K., Gurewitz E., and Fox G. (1990) *Statistical mechanics and phase transitions in clustering*, Phys. Rev. Lett., 65:945-948.

Soong, F. K. S., Rosenberg, A. E., Rabiner, L. R., and Juang, B. H. (1985) *A Vector Quantization Approach to Speaker Recognition*, Proceedings ICASSP 85, pages 387-390

Vonwiller J., Rogers I., Cleirigh C., and Lewis W., (1995) *Speaker and Material Selection for the Australian National Database of Spoken Language* J. of Quan. Lin. Vol. 2 pages 177-211