The use of broad phonetic class models in speaker recognition

J.W. Koolwaaij & J. de Veth

A2RT, Dept. of Language and Speech
University of Nijmegen
{koolwaaij|deveth}@let.kun.nl

Abstract

In this paper we investigate the use of broad phonetic class (BPC) models in a text independent speaker recognition task. These models can be used to bring down the variability due to the intrinsic differences between mutual phonetic classes in the speech material used for training of the speaker models. Combining BPC recognition with text independent speaker recognition moves a bit in the direction of text dependent speaker recognition: a task which is known to reach better performance.

The performance of BPC modelling is compared to our baseline system using ergodic 5-state HMMs. The question which BPC contains most speaker specific information is addressed. Also, it is investigated if and how the BPC alignment is correlated with the state alignment from the baseline system to check the assumption that states of an ergodic HMM can model broad phonetic classes [3].

1. Introduction

For text-independent speaker recognition HMM-based methods have been popular for several years now. Especially ergodic HMMs are suited to model a speaker's characteristics. Because these speaker models are trained on speech of which no transcription is available, they model both the speaker and his/her speech. One way to separate the speech information and the speaker information in the HMMs is to use a set of HMMs per speaker in which each HMM models a particular linguistic or phonetic unit. Different levels of segmentation are possible, e.g., word level, subword level, phone level or phonetic class level. Because the speech we are dealing with is conversational, without any restriction on the vocabulary, we have chosen for modelling on phonetic class level. Broad phonetic class models seem to be the best compromise between insufficient data to train more, and more specific models on the one hand and a reduction of linguistically induced acoustic variation in a single model on the other. In this paper we intend to investigate whether this approach to separating speech and speaker characteristics does improve speaker verification performance.

2. Database

Experiments were carried out on the NIST 1998 evaluation data (taken from the Switchboard-2 phase 2 corpus) [4]. The task is text independent automatic speaker verification (TI-ASV). Not all train and test conditions are evaluated in this paper. Specifically, results are presented for 250 female targets. This population is large enough to obtain representative results, and the performance of our baseline system was proven to be comparable for male and female speakers. Training was done on 120 seconds of speech per target speaker taken from only one session, and the 25000 test samples contained only 3 seconds of speech each. All speech is conversational telephone speech.

3. The baseline system

Acoustic features are 12 LPC-based zero-mean cepstral coefficients, plus log energy and their first and second time derivatives, yielding 39-dimensional feature vectors. We applied a preemphasis with factor 0.97 and used a Hamming window (length 25.6 ms, step 10.0 ms).

Target speakers are modelled by ergodic HMMs; one model per target is trained on the total of 2 minutes of training speech. The HMM topology used has 5 states, 32 mixtures per state and a diagonal covariance matrix. We did not train silence or non-speech models. In addition to the target models there is a gender dependent world model with the same topology. The world model is trained on 40 minutes of speech from a pool of female speakers taken from the NIST 1997 evaluation data. The world model data contains recordings from several handset types.

The target model is created by adapting the world model to the target's speech characteristics using the training data. This provides more robust and reliable model estimations with limited training data, and the loglikelihood ratio (LLR) measures only those acoustic events for which the target and the world model really differ.

We wanted to use target independent accept/reject thresholds; thus, we were obliged to apply target dependent normalisation of the LLR scores before the threshold test. For this purpose we used the z-norm technique: For each existing target model a corresponding impostor LLR distribution is estimated, using a development set of 30 same sex impostor speakers. This distribution is described by a mean μ_S and a standard deviation σ_S for each individual target model S. During testing of utterance X the z-norm parameters are applied as follows:

$$LLR_{\mathbf{Z}}(X|S) = \frac{LLR(X|S) - \mu_S}{\sigma_S} \tag{1}$$

So in fact applying z-norm means rescaling the impostor LLR distribution to a standard normal distribution. The target normalisation technique z-norm can easily be transformed to a handset normalisation technique (h-norm [5]) by choosing the mean and standard deviation to be handset dependent. In this paper h-norm is not applied yet.

4. The BPC system

For the BPC system the feature extraction and the algorithms for training and testing are exactly the same as in the baseline system. The only difference between the two systems is that for the BPC system train and test data were initially segmented into 5 broad phonetic classes and silence (or non-speech). The classes we used are vowels, fricatives, plosives, nasals and liquids. The speech material used for training of the BPC classifier was selected from a Dutch database named VIOS, which contains a large number of telephone calls recorded with the on-line version of a spoken dialogue system called OVIS [6]. Training material consisted of 25,104 utterances (319,849 BPCs) from these telephone calls. On a Dutch test database the best sentence BPC error rate was found to be equal to 38.6%. The confusion matrix shows that especially vowels and liquids are mixed up. The segmentation of the NIST train and test data is done as a free BPC recognition task. The alignment of the test data is used as input for the Viterbi algorithm.

To keep the model complexity per target about the same as for the baseline system, each target speaker is modelled by a set of 5 ergodic HMMs (one per BPC) each with 4 states and 8 mixtures per state; the additional non-speech model had the same topology. The world is also modelled by a set of 6 ergodic HMMs (one for each phonetic class and one for silence) with the same topology and trained on the same training data as the world model in the baseline system. The silence model is chosen to be target dependent, to be able to measure channel and environment effects on the LLR, but the silence segments are not contributing to the LLR during testing.

5. Results and discussion

5.1. Comparison of BPC and baseline system

For both systems results are reported in terms of Half Total Error Rate (HTER) [2]. From equations 2 and 3 it can be seen that HTER is equal to the posterior error probability in the condition where the prior probabilities for claimant and impostor attempt are equal.

$$\mathrm{P}(\mathrm{Error}) = \mathrm{P}(\mathrm{Accept}|\mathrm{Impostor})\mathrm{P}(\mathrm{Impostor}) + \mathrm{P}(\mathrm{Reject}|\mathrm{Claimant})\mathrm{P}(\mathrm{Claimant}) \tag{2}$$

$$\mathrm{HTER} = \frac{1}{2}\mathrm{P}(\mathrm{Accept}|\mathrm{Impostor}) + \frac{1}{2}\mathrm{P}(\mathrm{Reject}|\mathrm{Claimant}) \tag{3}$$

Since the degree of matching between the conditions in testing data and the conditions in training data is very important for performance, we determined ASV performance depending on two important factors. The first factor is whether or not the same telephone number is used in training and testing. The second factor is whether or not the same handset type is used in training and testing. This results in three[1] experimental conditions which are, graded according to increasing difficulty:

- SNST: Same telephone number and same handset type
- DNST: Different telephone number and same handset type
- DNDT: Different telephone number and different handset type

HTER results are reported in Table 1. The second and third column show the performance for, respectively, the baseline system and the BPC system.

Since the BPC system yields individual LLR scores for each phone class, we need to fuse these scores. Of the fusion techniques investigated the time-normalised sum of z-normalised LLRs of each particular BPC (Eq. 4) gave the best results. So the score used for the ASV decision is defined as

$$LLR_{\mathbf{Z}}(X|S) = \sum_{i=1}^{5} LLR_{\mathbf{Z}}(X = BPC_i|S) \cdot PF(BPC_i) \tag{4}$$

where $PF(BPC_i)$ is the percentage of frames in the test utterance assigned to BPC_i. As can be seen from Table 1, the baseline system consistently outperforms even the best BPC system.

One of the possible explanations for the different performance of the two systems may be the fact that the silence segments of a test utterance do not contribute to the final LLR in the BPC system, while they do in the baseline system. So in the fourth column of Table 1 the results are displayed for a BPC system for which silence is treated as one of the BPCs and so does contribute to the final LLR. This system does not outperform the baseline system either, so contribution of the silence segments is not the main reason. However, this example confirms our expectation that the silence segments contain information about the recording environment, because performance improves for the same telephone number condition and degrades for the two different telephone number conditions.

Table 1: Performance results for the baseline system, the BPC system and two combined systems (in terms of HTER)

	Baseline	BPC	BPC+Silence	BPC+Baseline
SNST	14.2%	15.9%	15.3%	12.6%
DNST	24.3%	25.2%	26.2%	22.9%
DNDT	33.2%	34.1%	35.0%	32.5%

Since the BPC system was not able to outperform the baseline system we investigated if the BPC system

[1] The fourth condition is almost certain an impostor attempt.

adds something to the baseline system by combining the two. The fusion of the LLRs is done by a linear combination of the LLRs where the weights are estimated using a least-squares method. Estimation is done by jackknifing the test set with the number of subsets N=5 (5000 utterances are used as development set and 20000 utterances as actual test set, and this rotated 5 times. Error rates are averaged over the 5 tests). This improves the performance of the baseline system with a relative 11% for SNST, 6% for DNST and 2% for DNDT. These results suggest that complementary information is available in the BPC and the baseline approach.

5.2. Speaker specificity of BPCs

It may be interesting to know if there is a particular BPC which outperforms the other BPCs for speaker verification. To investigate this issue we computed the HTER per BPC: Only those frames which are assigned to one particular BPC contribute to the score used for decision (in this case the time-normalised and z-normalised LLR per BPC). For each BPC the HTER is reported in Table 2. The vowels perform

Table 2: Performance results for the 5 broad phonetic classes (in terms of HTER)

	Vowel	Fricative	Plosive	Nasal	Liquid
SNST	21.6%	31.2%	21.7%	29.5%	24.1%
DNST	28.6%	39.1%	32.2%	35.0%	28.8%
DNDT	35.6%	43.9%	39.5%	44.1%	37.2%

best, having the lowest error rate. The question arises if the BPC with the lowest error rate is also the most speaker specific BPC? The answer can be positive if and only if each BPC is trained and tested in exactly the same conditions. However, in our case the distribution of BPCs over the frames in the test material is not uniform. (The same is true for the training material but we found that it is less important.) In Table 3 the frequency of occurrence of each BPC is given and it appears that the vowel, which had the

Table 3: Frequency counts of the 5 broad phonetic classes plus silence in the test data.

Vowel	Fricative	Plosive	Nasal	Liquid	Silence	Total
26.4%	7.1%	14.6%	7.4%	18.8%	25.7%	100.0%
0.79s	0.21s	0.44s	0.22s	0.56s	0.77s	3.00s

lowest error rate, is also the most frequent BPC in the test data. So before stating which BPC is the most speaker specific one, we need to compensate for the amount of the test data per BPC, because more test data leads to more reliable estimates of the decision score, which in turn leads to better performance. Figure 1 shows the HTER per BPC as a function of the amount of available test data (see also Table 3) for that particular BPC. The dotted lines are functions of the form $a \cdot t^b$, where t is the amount of test data per utterance in seconds. For the SNST condition the best matching function is

$$\text{HTER}(t) = 19.6 \cdot t^{-0.27}. \tag{5}$$

That this function is adequate (even for extrapolation) is shown by the fact that the BPC system which uses all frames except silence (so 74.3% of 3 seconds is 2.23 seconds) has an error rate of 15.9% (see Table 1) and according to equation (5) the corresponding HTER(2.23) is equal to 15.8%. We can also conclude that to halve the error rate 13 times more test material is needed! If each two speech frames were really independent, we would expect a $1/\sqrt{t}$-relation here, but due to the fact that speech frames simply are not independent of each other we need more than the expected factor 4 of test material to halve the error rate.

The results in Figure 1 suggest that the BPC models we trained cannot be rank-ordered according to speaker specificity. One explanation could be that our BPC classifier is far from perfect, but we do not expect a perfect classifier within the next few years. Phonetic studies on speaker specificity for individual

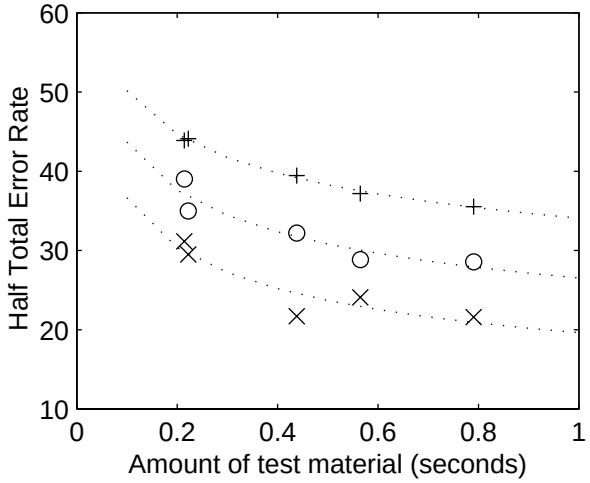

Figure 1: HTER as a function of the amount of test data for each individual BPC [SNST ($\times$); DNST ($\circ$); DNDT ($+$)].

speech sounds lead to the conclusion, however, that different phonetic speech segments actually do carry different amounts of speaker specific information [1]. From that study it appears that it is specific members of a phonetic class that are more speaker specific than other members, rather than classes which as a whole are more speaker specific. Moreover, the results in [1] are based on hand segmented phones in carefully read speech.

5.3. Time alignment

Ergodic HMMs have usually been assumed to be effective for text-independent speaker recognition because they automatically form broad phonetic classes corresponding to each state [3]. To check this assumption we compared the time alignment of the BPCs (in the top-down BPC system) with the time alignment of the states (in the bottom-up baseline system). There are no minimum duration constraints for neither the states in the baseline system, nor the BPCs in the BPC system.

In Table 4 it is shown how often states from the baseline system coincide with BPCs from the BPC system. For example, 9.0% of all frames in the test set are assigned to state number 2 by the baseline system, while these frames are assigned to silence by the BPC system. Since in total 19.1% of all frames in the test set are assigned to state 2, it follows that almost 50% (9.0 of 19.1) of the frames assigned to state 2 seem to come from silence intervals.

Table 4: Frequency of coincidence of states from the baseline system and BPCs from the BPC system. The figures represent proportions of the total number of frames in the test set.

State #	1	2	3	4	5	Total
Silence	5.5	9.0	2.0	7.3	1.9	25.7
Vowel	7.3	4.0	8.1	6.1	0.9	26.4
Fricative	2.2	0.6	1.6	2.7	0.1	7.1
Plosive	3.2	2.0	5.2	4.1	0.1	14.6
Nasal	2.4	0.9	2.8	1.1	0.3	7.4
Liquid	6.2	2.6	5.4	4.2	0.4	18.8
Total	26.8	19.1	25.1	25.5	3.7	100.0

There were no striking similarities, except for the fact that two states (2&5) seem to model silence (or non-speech): Almost 50% of the time of these states are assigned to silence in the BPC approach. Ac-

cording to the data in Table 4, states in an ergodic HMM do not correspond in a straightforward way to humanly defined BPCs.

6. Conclusion

Broad phonetic class modelling is applied in text independent speaker recognition to be able to build more specific speaker models in which less variation due to different phonetic classes is present. This kind of modelling is not able to outperform our baseline system on its own. However, combination of the BPC approach with the baseline system gives a significant improvement in performance. Apparently, the two different approaches describe complementary information.

Further, the speaker specificity of each BPC is investigated. Speaker recognition performance is different for each BPC, but the main reason for those differences appears to be the fact that some BPCs occur more frequently or have a longer duration. After compensation for the number of seconds of test material, performance is more or less the same for all BPCs.

References

[1] H. van den Heuvel, Speaker variability in acoustic properties of Dutch phoneme realisations, Ph.D.Thesis, Nijmegen, 1996

[2] J. Lindberg, J.W. Koolwaaij, H.-P. Hutter, D. Genoud, M. Blomberg, J.-B. Pierrot & F. Bimbot, Techniques for a priori decision threshold estimation in speaker verification, Proc. RLA2C, pp. 89-92, 1998

[3] T. Matsui, S. Furui, Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs, Proc. ICASSP, pp. 157-160, 1992

[4] NIST, Speaker recognition workshop notebook, 31 March - 1 April 1998

[5] D. Reynolds, Comparison of background normalization methods for text-independent speaker verification, Proc. Eurospeech, pp. 963-966, 1997

[6] H. Strik, A. Russel, H. van den Heuvel, C. Cucchiarini & L. Boves, A spoken dialogue system for the Dutch public transport information service. Int. Journal of Speech Technology, Vol.2, No.2, pp. 119-129, 1997