SPEAKER ADAPTATION USING A PREDICTIVE MODEL

Stephen Cox

School of Information Systems,
University of East Anglia, Norwich NR4 7TJ, UK.
Tel: +44 603 592582
e-mail: a083@uk.ac.uea

ABSTRACT

A new technique of speaker adaptation for use in speaker-independent speech recognition systems is presented. The training-data is used to build models (based on linear regression) of sounds. At recognition time, the models are used together with an incomplete set of sounds from a new speaker to estimate values for unheard sounds, which are then used to adapt the speaker-independent models. The technique reduced the error-rate from 17% to 5.3% when applied to a database of 104 speakers speaking the English alphabet.

1 Introduction

A central problem in building speech recognition systems which are designed to work on previously unheard voices is how to model the variation from voice to voice in the acoustical signal representing a given speech unit. This variation is caused by such effects as different vocal tract sizes and shapes, different accents, speaking styles etc. Given some data from a previously unheard speaker, it seems natural to attempt to 'tune in' the system to work better on the new voice, a technique which has become known as 'speaker adaptation'. Speaker adaptation techniques have usually concentrated on adapting estimates of the system's speech model parameters after it has been exposed to vocabulary examples from a new speaker e.g. [4]. In most cases, when the system hears an example of a sound X from the new speaker, it updates only the parameters of the model of X. In this work, our premise is that an example of the sound X from the new speaker contains potentially useful information about *many* speech sounds that that speaker is likely to produce, so that when the system hears an example of sound X from the new speaker, it updates parameters of several other models as well as the model for X. The prediction of new sounds is made by building regression models of vocabulary sounds produced by the speakers available for training the system. Then, given an incomplete set of vocabulary sounds from a new speaker, the models are used to predict values of unheard vocabulary sounds from the speaker, and these predictions are used to adapt the system model parameters. In this work, we have assumed that the sounds provided by the new speaker are labelled (i.e. the adaptation is supervised) but the technique can easily be extended to unsupervised adaptation.

In [1], we reported on a preliminary investigation of this idea in which we used 11 different vowel sounds from 30 speakers. The advantage of using this data was that each example was a single vector which dispenses with the need for time alignment procedures. In this paper, we report on the extension of the technique to a real speech recognition problem.

2 Preliminary investigation using static vectors

To aid understanding of the technique used in these experiments, a brief description of the work on isolated vowel sounds is included here. For a fuller account, see [1]. The data consisted of a single example of each of 11 different vowels from each of 30 speakers. Each example was represented by an 8-dimensional mel-frequency cepstral coefficient (MFCC) vector [3] obtained by averaging several vectors representing a steady-state vowel. The data was divided into a training-set of speakers 1-16 and a test-set of speakers 17-30. Each vowel class was modelled as a multivariate Gaussian probability density (based on the training-set examples) and an example from the test-set was classified by computing the likelihood of each of the 11 probability densities producing the example.

Throughout this work (and the new work presented here), it was assumed that the vector dimensions were uncorrelated. Firstly, 'scattergrams' of the training-set speakers' data were made for all possible pairs of vowel classes and in each vector dimension. The correlation coefficients of the data in each plot were computed, the average value (over all vowel-pairs and all 8 dimensions) being 0.48, with no significant negative correlations. In the first experiment, the linear regression coefficients corresponding

to these 'scattergrams' were computed and stored. The test-set speakers' data was divided into an adaptation-set of 5 vowels and a test-set consisting of the other 6. Each 'adaptation' vowel was used with the appropriate regression coefficients to obtain a prediction of each of the 6 test-set vowels and these predicted values were used to modify the estimate of the means of the corresponding 6 vowel classes. The modification was of the form
$\texttt{new\_mean} = p * \texttt{present\_mean} + (1-p) * \texttt{predicted\_value}$
where $0 < p < 1$ (p is chosen empirically). The 6 unseen vowels (only) were then classified using the set of modified means. The result was a 40% decrease in the 'no-adaptation' error-rate. It was also a small improvement on another speaker-adaptation technique which was described in [2] and which we may term 'bias vector' adaptation. Briefly, the technique consists of computing the average distance $\boldsymbol{\delta}$ (the 'bias vector') between each of the speaker's sounds given for adaptation and their corresponding class means, and modifying the means of the distributions of the *unseen* vowels by $\texttt{new\_mean} = \texttt{present\_mean} + p * \boldsymbol{\delta}$ where $0 < p < 1$.
In other experiments described in [1], multiple linear regression was used to decrease the error-rate by a further 12%. However, the new work described here is based on the simple linear regression technique described above.

3 The data and models

The speech database for these experiments was provided by British Telecom [5] and consisted of 3 utterances of the alphabet from each of 104 speakers recorded in a soundproof room with a high-quality microphone at a bandwidth of approximately 8 kHz. Each utterance was manually endpointed and processed into frames of duration 16 ms, each frame consisting of a 17-d vector containing 8 MFCCs, 8 differential coefficients and a log-energy coefficient. The training-set (52 speakers) was used to construct a 10 state continuous density hidden Markov model (HMM) of each alphabetic class, the state PDFs being unimodal Gaussian with a diagonal covariance matrix. The topology of the HMM was a simple one in which state i was connected only to itself and state $i+1$.
The test-speakers' data was divided into a set of 'adaptation classes' and a set of 'test classes':

Adaptation classes:	ADEFGIKMOPQRX
Test classes:	BCHJLNSTUVWYZ

The division was made in such a way as to keep the data as phonetically balanced between the two sets as possible. During testing, the HMMs of all classes that could be adapted were adapted, but only the utterances of the test classes were tested. The rationale for this procedure was that (a) less data would have been available for adaptation if we had tested on all classes and (b) the object of the experiment was to test the ability of the model to predict the test classes and so it was felt that it was less important to test the adaptation classes.

4 Experimental procedure

4.1 Overview

Firstly, the Viterbi algorithm was used to align the training set utterances with their corresponding models so that each frame in an utterance was associated with a state. Taking each training-set speaker in turn, all the frames from his/her utterances associated with a given state of a given model were averaged, so that each speaker had a mean vector associated with each state of each HMM. Throughout this paper, we term this vector the *speaker state vector*. Hence the $13 \times 10 = 130$ HMM states play the rôle of the vowel classes in the previous work and the adaptation method is:

(a) make a prediction of the speaker state vector for each state of each test class

(b) use the prediction to adapt the corresponding HMM state mean

Each state of each adaptation class was paired with each state of each test class and a correlation coefficient and two linear regression coefficients (for each vector dimension) were computed and stored. At testing time, the adaptation utterances provided by a new speaker are aligned with their corresponding models. The speaker state vector of a given adaptation class model can then be used to obtain a prediction of the speaker state vector of *any* state of *any* test class model by transforming the adaptation data using the appropriate pairs of stored regression coefficients. This raises the question of how to make the best use of the adaptation data for predictive purposes which proves to be a key issue in the technique.

4.2 Making best use of the available adaptation data

Suppose we have available from a new speaker some example utterances of a subset of the adaptation classes speaker. It is clear that the suitability of an adaptation class for predicting a given test class depends on the phonetic similarity of the two classes. For instance, we might expect that data from the adaptation class 'M' would be good at predicting the test class 'N' but not necessarily at predicting class 'W'. The correlations of the 'scattergram' data between the states provide us with a quantitative way of assessing the suitability of states of the adaptation classes to predict states of the test classes. There are clearly many sensible ways of using the adaptation data to make predictions of the test class values, but initially, we used the following simple method:

The prediction of the speaker state vector of state k of test class l ($S_t^l(k)$) is made using state i of adaptation class j ($S_a^j(i)$), where $S_a^j(i)$ has the highest *average* correlation coefficient (averaged over all vector dimensions) with $S_t^l(k)$.

Note that this criterion allows each speaker state vector of an individual test class model to be predicted from a state of a different adaptation class model, but ensures that components within a speaker state vector are predicted from the same adaptation speaker state vector.

4.3 Bayesian adaptation of the state means

Suppose the optimum predictor (according to the above criterion) of $S_t^l(k)$ is $S_a^j(i)$. We use the speaker state vector of $S_a^j(i)$ together with the appropriate set of regression coefficients (2 for each vector dimension) to make a prediction of the speaker state vector of $S_t^l(k)$. Let ρ be the correlation coefficient (in a given dimension) between the two states. If $|\rho|$ is high, we would be confident of a good prediction of the speaker state vector. However, if $|\rho|$ is low, we would not want to place much confidence in this prediction and would prefer to use the existing 'speaker-independent' mean in the HMM. A Bayesian approach is clearly appropriate here. The *a priori* distribution of the state mean *for a given speaker* can be taken as the sample distribution associated with that state (estimated from the training-data). This is assumed Gaussian with mean μ and variance σ^2 in a given vector dimension. Taking the predicted value (say y) of the speaker state vector in this dimension to also be an estimate of the mean, the distribution of the predictions is also assumed Gaussian with mean y and variance $(1-\rho^2)\sigma^2$, where ρ is the correlation coefficient between the states in this dimension. We may thus write the likelihood of observing a mean value of z in this vector dimension as:

$$L(z) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[\frac{-(z-\mu)^2}{2\sigma^2}\right] + \frac{1}{\sqrt{2\pi}\,\sqrt{1-\rho^2}\,\sigma}\exp\left[\frac{-(z-y)^2}{2(1-\rho^2)\,\sigma^2}\right] \quad (1)$$

The optimum value of z maximises $L(z)$ in equation 1. Notice that when $|\rho|$ is close to 1, the second term becomes very large as $z \to y$, i.e. the expression is maximised by choosing z to be close to the value predicted by the regression. When $\rho = 0$, the terms are both Gaussians with equal variances, the first with mean μ and the second with mean y. Hence it appears that equal weight is given to the *a priori* mean and the prediction, the latter being worthless when $\rho = 0$. However, when $\rho = 0$, the regression line is horizontal and the prediction $y = \mu$, always, so that equation 1 is maximised by $z = \mu$. In practice, the maximisation of equation 1 cannot be done in closed form and rather than implement a numerical maximisation, some simulations showed that using:

$$z = |\rho|^2 y + (1 - |\rho|^2)\mu \quad (2)$$

gave a good approximation to the observed maximum. It is intended to experiment with rigorous maximisation of equation 1 at a later date.

5 Results

We investigated the effect on the recognition error-rate of making data available for adaptation from an increasing number of classes, and compared the performance of this new technique with that obtained by using the 'bias-vector' technique described in [2]. In Fig 1, the abscissa is the number of classes for adaptation available from each speaker. The identities of these classes are given in Table 1 below:

No of classes	Class identities
1	E
2	EA
3	EAX
4	EAXI
5	EAXIQ
6	EAXIQD
7	EAXIQDK
8	EAXIQDKM
9	EAXIQDKMP
10	EAXIQDKMPF
11	EAXIQDKMPFG
12	EAXIQDKMPFGO
13	EAXIQDKMPFGOR

Table 1: Identities of the adaptation classes in Fig 1.

It should be noted that the results in Fig 1 are for testing on the test classes only, although all the models were active for each recognition. The models of the test classes were adapted according to the new technique described in section 4; the models of the *adaptation* classes available from the speaker were adapted by replacing the mean of the appropriate state PDF by the speaker state vector. Any adaptation classes not available from the speaker were not adapted. In practice, we found that adapting the adaptation class means made very little difference to the result when testing only the test classes. The results show that when only a single class ('E') is available for adaptation, the error-rate is reduced from 17% to 9.6%. A system which did not use a predictive technique (i.e. which adapted only the model of 'E') could not hope to obtain such a large improvement given data from only a single class in the vocabulary. Addition of adaptation data increases performance monotonically

until a plateau (5.3%) is reached after nine classes are given. No statistical analysis of the results has been performed but given the size of the test set (about 2000 utterances), an improvement of about 1% is statistically significant and the improvement of about 12% shown here is of real practical significance.

6 Conclusions and Future Work

We believe the method of speaker adaptation proposed here has good potential and should be of particular relevance to large-vocabulary speech recognition systems in which rapid adaptation to a new speaker's voice is essential. In this first study, we have used a very simple method of using the available adaptation data for predictive purposes and a key issue in the further development of the technique is how to make better use of this data. The preliminary studies on static vowel vectors suggest that multiple linear regression is a promising technique for improving estimates.

References

[1] S.J. Cox. Speaker adaptation in speech recognition using linear regression techniques. *Electronics Letters*, 28(2):2093–2094, October 1992.

[2] S.J. Cox and J.S. Bridle. Unsupervised speaker adaptation by probabilistic spectrum fitting. In *Proc. IEEE Conf. on Acoustics, Speech and Signal-processing*, pages 294–297, April 1989.

[3] S.B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28:357–366, 1980.

[4] C.H. Lee, C.H. Lin, and B.H. Juang. A study on speaker adaptation of continuous density HMM parameters. In *Proc. IEEE Conf. on Acoustics, Speech and Signal-processing*, April 1990.

[5] J.A.S. Salter. The RT5233 alphabetic database for the connex project. Technical Report RT52/G231, BT Technology Executive, April 1989.