ON THE SEPARATION OF SPEECH SIGNAL VARIANCES FROM TWO SOURCES

Xue YANG, J. Bruce Millar and Iain Macleod

Computer Sciences Laboratory
Research School of Information Sciences and Engineering
Australian National University
Canberra, ACT 0200, Australia
email: {xue,bruce,iain}@cslab.anu.edu.au

ABSTRACT - Variability is an inherent characteristic of speech signals and its management plays an important role in speech processing. This variability arises from many sources. In this study, techniques are developed to explore the separation of speech signal variances from two sources in the low-order cepstral space. Hidden Markov modelling and statistical comparison of multi-variate distributions are used in a novel way to capture local variance in the midst of the normal dynamics of speech signals. Subsequent analysis demonstrates differences between this local variance and the global variance which is often used to characterise signal variance without reference to its significant components, and reveals speaker characteristics which may not be observed by simply looking at the global variance of speech data.

1. INTRODUCTION

The variance of speech signals encodes the information that is intrinsic to speech communication. However the overall variance has contributions from many sources which are mingled together in ways that are not easy to disentangle. Measures of speech variance are important in characterising specific speech sounds or the speech of specific speakers. Accurate assessment of such measures should enable improved performance of speech and speaker recognition systems. In both of these domains the Mahalanobis distance between a testing vector and a reference vector is regularly used to replace a simple Euclidean distance if the elements of such vectors have widely different variances and significant inter-dimensional correlations in the vector space. This has been argued for speaker recognition many years ago by Atal (1976).

Three studies on the extensively used "cepstral" space have calculated cepstral variances based on (a) a wide range of sounds from several speakers (Juang et al., 1987), (b) a single word from a large range of speakers (Tohkura, 1987), and (c) some isolated and connected words repeated several times by individual speakers (Soong et al., 1988). Such measures contain several sources of variance which may be characterised as follows: type (a) includes inter-speaker variance, intra-speaker variance, and phonetic variance; type (b) includes inter-speaker variance and phonetic variance; and type (c) includes intra-speaker variance and phonetic variance. Additional sources of variance are also always present such as that due to the recording environment, but these may be regarded as secondary for our present purpose.

In VQ-based text-independent speaker recognition systems, global variances derived from training data from individual speakers are sometimes used in Mahalanobis distance calculations for each codeword (e.g. Soong et al., 1988). This global variance clearly contains components of both intra-speaker variance and phonetic variance. If (to a first order approximation) a codeword can be regarded as representing a particular phonetic state and variances of the codeword as representing intra-speaker variance for that state, it seems less than optimum to use the global variance in place of intra-speaker variance for that state when computing deviation from the VQ-based speaker model.

In this study, we develop techniques to enable the separation of variance that is essentially of phonetic origin and that which is essentially the variance of speaker performance in attaining phonetic goals. The only form of intra-speaker variance considered is that associated with repetition of the same word on different occasions (assuming normal physiological and emotional states); we refer to this as intra-speaker repetition variance hereafter. Consequently this separation enables us to demonstrate differences between intra-speaker repetition variance and global variance which is derived from speech data comprising repetitions of the same word. This separation can also reveal some phenomena which may not be observed by simply looking at global variance of speech data.

2. SPEECH MATERIAL AND ANALYSIS

Four English monosyllabic words *we*, *you*, *how* and *high* were selected for this study. The phonetic

variance in these words traverses the vowel space in all its dimensions: characterised by tongue body movement in articulatory domain and by the dynamics of acoustic representations in the acoustic domain. This data corpus was designed to be spoken 32 times by 11 Australian male speakers in 4 recording sessions, with a target time interval between sessions of 1 week and with each word being spoken 8 times in each session by each speaker. In the course of data collection, several speakers selected were not able to attend on schedule. As a result, minimum and maximum time intervals between successive sessions were respectively 1 day and 12 days, with an average time interval between successive sessions of 6.5 days. Occasional speaking errors made by two speakers meant that they had less than 32 correct utterances for some words --- speaker IM has 31 for *high*, and speaker JW has 31 for *we*, and 30 for both *you* and *high*. These deviations from the original design are relatively minor and should not affect the overall results.

The data were originally sampled at 20000 samples/s, and then down sampled to 10000 samples/s. The vocalic parts of the recorded utterances were manually identified for this study. Each utterance was reduced to a sequence of vectors, each comprising 12 low-order cepstral coefficients, by LPC analysis of frames of 20ms width with 50% overlap.

3. METHODS

In this study, we want to separate intra-speaker repetition variance and phonetic variance for each given word on a per-speaker basis. For this purpose, we need to model each word in such a way that the word's acoustic variance attributable to its phonetic characteristics and that attributable to intra-speaker repetition are separately manifested. We have explored the use of hidden Markov modelling, which is an example of stochastic techniques for the study of nonstationary processes (Poritz, 1988), for this purpose. In a hidden Markov model, the sequence of states describes temporal properties, while the probability distribution in each state characterises the statistics of the associated time interval.

In this study, the trajectory of LPC based cepstrals for a given word/speaker combination is represented as a sequence of states. Each state is assumed to correspond to one of the sequence of vocal tract configurations (or phonetic states) used in the production of that word, the acoustic properties of which are characterised by the state's internal probability distribution.

Hidden Markov modelling is a procedure to estimate the model parameters based on observation sequences. It comprises three primary parts: (1) a model and its structure; (2) a criterion for estimation; and (3) an algorithm to implement the procedure. These three parts are described in 3.1, 3.2 and 3.3 below.

3.1 The model and its structure

A hidden Markov model can be summarised by its parameters $\lambda = (a_0, A, B)$, where

- $A = \{a_{ij}, i, j = 1, ...N\}$ is the state transition probability matrix,
- a_0 is the initial state probability distribution,
- $B = \{b_i, i = 1, ...N\}$ are output probability distributions in each state i,
- N is the number of states in a model.

We select a left-to-right with no skips structure (Figure 1) as we are interested in modelling the sequential dynamics of each word by the state sequence. This structure has the property that as time increases the state index either increases sequentially or stays the same. When used to model a given word, each state can be assumed to represent the articulatory configuration or phonetic state at the associated time, and to describe the corresponding acoustic properties by its output distribution.

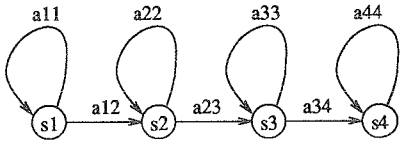

Figure 1: Structure of a 4 state left-right hidden Markov model

The self-transition probability a_{ii} conveys information about the duration of state i although in this architecture it does not model duration accurately. As we are interested only in the statistical properties of each state, all non-zero values in the transition probability matrix were made equal. Due to the

mathematical constraint that the sum of transition probabilities from any state should be unity, we set all values to 0.5 so that (apart from the last state) the constraint is obeyed.

Taken together, the above considerations result in a state transition probability matrix of the following form, with N = 4 for example:

$$A = \begin{pmatrix} 0.5 & 0.5 & 0.0 & 0.0 \\ 0.0 & 0.5 & 0.5 & 0.0 \\ 0.0 & 0.0 & 0.5 & 0.5 \\ 0.0 & 0.0 & 0.0 & 0.5 \end{pmatrix}$$

The initial state probability distribution of 0.5 for state one and state two allows an equal chance of starting in either state.

The output distribution for each state is assumed to follow a multivariate Gaussian distribution. Hence the variance within a state, in the form of the covariance matrix of this multivariate Gaussian, can be regarded as representing intra-speaker repetition variance of the acoustic properties of the corresponding vocal tract configuration or phonetic state.

A novel method for the selection of the number of states required to model a specific word by a specific speaker involves an iterative statistical comparison of the output distribution of adjacent states on trial models. We set a criterion to test statistically inter-state relationships prior to acceptance of an optimal model with a particular number of states. The same criterion is applied to models for all words and speakers. This may result in different numbers of states where different dynamic characteristics exist.

The iterative procedure is as follows:

- Step 1: Start with the number of states N = 2.
- Step 2: Make model estimations of maximum joint likelihood in $P(O, s|\lambda)$ with N states using the segmental K-means algorithm.
- Step 3: Test the hypothesis that there is a pair of adjacent states whose mean vectors are no longer unequal at a significance level α, using Hotelling's T-square test. If the hypothesis is accepted, go to Step 5.
- Step 4: Increase N by 1, go to Step 2.
- Step 5: The final number of states for the model is N - 1.

This procedure increases the number of states until at the significance level α we can still say that any two adjacent states have output distributions representing separate distributions. If this level of significance is weak (large value of α) then the procedure generates a large value of N. The selection of α determines to a first approximation the acoustic difference which we adopt as our quantum of acoustic phonetic change. If this quantum is large (small N) then acoustic dynamics will more strongly influence measures of intra-speaker variance internal to each of the N states. In theory, when the number of states in a model increases, the values of T-square in Hotelling's T-square test for all pairs of adjacent states decrease. The T-square value statistically represents the distance between two adjacent states. In practice, the values of T-square for all pairs of adjacent states do not at the same time fall below the threshold determined by the degrees of freedom (number of datapoints, dimension of variables) and significance level (α) when the number of states increases. We stop increasing the number of states as soon as we find (as in step 3 above) that there is one pair of adjacent states having no difference at level of significance α. In the initial study reported here a value of α = 0.01 was chosen.

3.2 The criterion for estimation

In general hidden Markov modelling, given the observation sequence O, the model parameters λ are adjusted to maximize $P(O|\lambda) = \sum_{s \in S} P(O, s|\lambda)$. This means that λ is chosen in order to maximize the sum of probabilities for all possible state sequences S that O might have.

An alternative criterion, which better suits our application, is to maximize the joint likelihood $P(O, s_m|\lambda) = max_{s \in S} P(O, s|\lambda)$. This means that λ is chosen to maximize the likelihood of the state sequence s_m for O, where s_m is the most likely among all possible state sequences S.

In our study, given an observation sequence O of a word, we are interested in a particular state sequence s of O, which can correspond to the sequence of phonetic states assumed to characterise that word. Therefore, we select maximum joint likelihood $max P(O, s|\lambda)$ as the criterion for estimation of model parameters λ.

3.3 The algorithm

The development of a particular algorithm for the estimation of model parameters arises from the interaction of the selection of the estimation criterion and the model type.

The Baum-Welch algorithm (Rabiner, 1989) is appropriate for estimation of hidden Markov model parameters λ in the sense of maximum likelihood $P(O|\lambda)$. In our study, because the criterion function is $max P(O, s|\lambda)$, we use the segmental K-means algorithm (Rabiner et al., 1985) to estimate λ.

Since the initial probability distribution a_0 and the state transition probability matrix A have been fixed, the segmental K-means algorithm is used to estimate parameters of the output distribution, in terms of the mean vector and covariance matrix of the multivariate Gaussian distribution, for each state.

The segmental k-means algorithm repeatedly uses the Viterbi algorithm (Forney, 1973) to segment each observation sequence of a word into states and to make a maximum likelihood estimation of the mean vector and covariance matrix for each state based on the observations segmented, from all observation sequences, into that state, until the criterion function converges.

As the segmental K-means algorithm is itself an iterative estimation procedure, the initial model is significant for the convergence to the final estimated model. The initial model is usually (Rabiner, 1989; Hidden Markov Model Toolkit V1.5, 1993) created from a uniform segmentation in time of each observation sequence into states.

In this study, we developed a novel method for forming the initial model based on segmentation of observation sequences according to their dynamic properties. It involved three steps:

- Step 1: the sum of Euclidean distances between each pair of adjacent observations (frames) from an observation sequence is computed;
- Step 2: given the number of states N, a threshold is chosen by dividing this sum by N; and
- Step 3: starting with the first observation of the sequence, we accumulate Euclidean distances for each successive pair of adjacent observations until the sum reaches the chosen threshold. All observations before this point (when the sum of Euclidean distances exceeds the threshold), are allocated to state number one. This procedure is restarted and continued until all observations have been examined and allocated to a corresponding state.

If the characteristics of the observation sequence change in a linear fashion (i.e. the dynamics are linear), the method described above results in a uniform segmentation in time. However, in reality, the dynamics are normally non-linear. Initial models derived using this method should be closer to the final converged models used for this study than those obtained via uniform segmentation in time.

4. RESULTS

A total of 44 left-to-right hidden Markov models were trained using the methods described above. As can be seen in Table 1, the number of states varies from speaker to speaker and from word to word. Together with the mean vector for each state, this reflects the differing degrees of acoustic realisation of phonetic variance in all these speaker/word pairs. Relatively greater spectral variation of the speech signal of a word results in bigger number of states in the model.

The pooled covariance over all states in a model was computed. This measure of covariance for each speaker/word pair provides a better measure of the intra-speaker repetition variance than the global variance calculated across all frames of the raw cepstral analysis vectors. In the pooled intra-state variance measure the majority of the phonetic variance expressed in the dynamics of the acoustic pattern over time has been accounted for by the variance between states and the number of states required.

Table 2 shows the proportion of the global variance derived by summing the diagonal elements of the pooled covariance matrix, expressed as a percentage and labelled as true intra-speaker repetition variance. For the speech corpus used in our study and using a quantum of acoustic phonetic dynamics defined by $\alpha = 0.01$ as described above, it can be seen that the true intra-speaker repetition variance is in nearly every case less that half of the global variance and in quite a number of cases is of the order of a quarter of the global variance.

Figure 2 shows the global variance and intra-speaker repetition variance per cepstral dimension for each word spoken by speaker AH. The upper plot in each case is the global variance. There are distinct differences in curve shapes of global variance across 4 words and distinct differences in curve shape between intra-speaker repetition variance and global variance of each word. This indicates that in our

speaker	number of states			
	we	*you*	*how*	*high*
AH	20	15	15	14
BM	16	15	15	15
DD	22	20	17	20
GC	17	17	15	16
IM	20	21	15	16
JW	16	14	12	14
KR	19	17	20	18
NF	18	18	13	15
SB	15	15	18	12
TB	14	15	16	19
WB	16	13	13	16

Table 1

speaker	Percentage of intra-speaker repetition variance present in the global variance (%)			
	we	*you*	*how*	*high*
AH	30.7	39.3	51.7	51.4
BM	21.5	33.3	44.8	39.3
DD	26.5	28.0	41.8	39.4
GC	25.4	49.2	37.5	51.2
IM	22.4	37.2	39.4	25.0
JW	26.9	38.7	44.0	38.7
KR	21.8	36.0	27.0	23.5
NF	17.8	26.5	49.2	33.5
SB	27.7	39.6	45.4	47.1
TB	29.8	39.2	45.1	32.5
WB	22.9	27.8	48.5	42.5

Table 2

study, phonetic variance, to certain extent, dominates the global variance in qualitative aspects as well. The limited data corpus of this study did not permit comparison with the shape of global variance over phonetically balanced material which is commonly used to represent intra-speaker variance in speech technology systems.

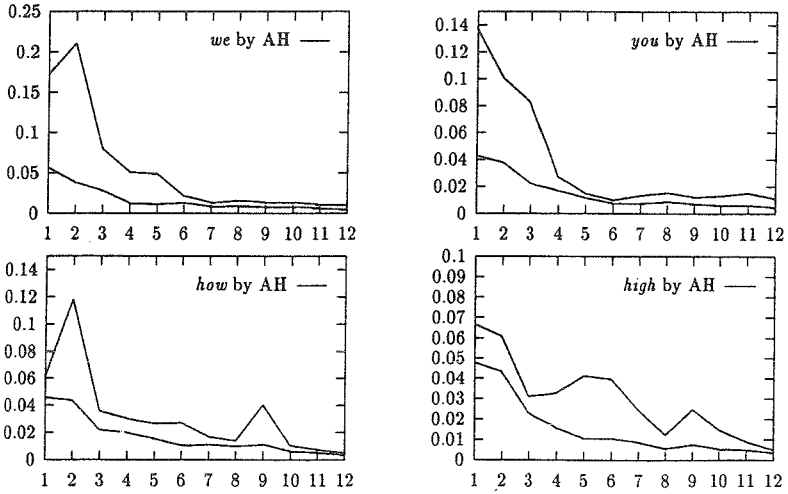

Figure 2: The global variance and intra-speaker repetition variance for each LPC cepstral coefficient for the 4 words by speaker AH.

Our modelling work also reveals a fact that some speakers have relatively greater dynamics but less intra-speaker variance, while others have relatively less dynamics but greater intra-speaker variance. Figure 3 gives an example of this. Because speaker KR has greater dynamic range for *how* than speaker AH, his model has a larger number of states (20) than AH's (15). This suggests that KR pronounces *how* more consistently than AH even though his global variance is higher. This phenomenon can not be observed if we simply look at the global variance or models with the same and very small number of states.

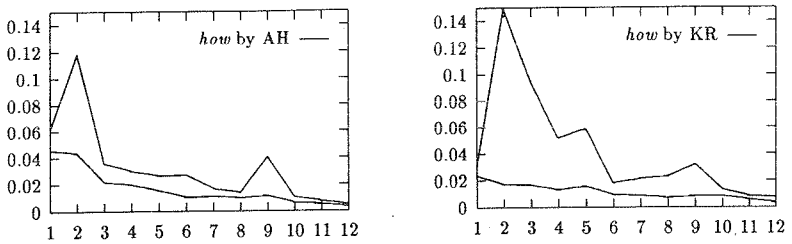

Figure 3: The global variance and intra-speaker repetition variance for each of LPC cepstral coefficient for word *how* by speaker AH and KR.

5. CONCLUSIONS

We have developed techniques, based on hidden Markov modelling using an iteratively designed architecture, to model given word/speaker combinations. Both intra-speaker repetition variance and the characteristics of the acoustic realisation of phonetic variance have been demonstrated in the output of the model.

The techniques described in this paper can be applied to assess quantitatively the relative contributions to global variance from phonetic sources that are essential in creating a phonemic sequence, and from other speaker specific sources such as repetition variance but also physiological, psychological, and environmental influences. It should therefore be possible to provide an analysis of a repeated performance of a given phonemic sequence in terms of the phonetic dynamism applied and the other intra-speaker variants involved.

Our study also suggests that the incorporation of variance in VQ-based text-independent speaker recognition systems may benefit from the incorporation of codeword-dependent measures of variance rather than global variance which includes the phonetic variance across the whole analysis space.

6. REFERENCES

Atal, B.S. (1976) *Automatic recognition of speakers from their voices* Proc. IEEE, Vol. 64, 460--475.

Forney, G.D. (1973) *The Viterbi algorithm* Proc. IEEE, Vol. 61, 268--278.

Hidden Markov Model Toolkit V1.5 (1993), Cambridge University Engineering Department Speech Group & Entropic Research Laboratories Inc.

Juang, B.H, Rabiner, L.R. and Wilpon, J.G. (1987) *On the use of bandpass liftering in speech recognition* IEEE Trans ASSP, Vol. 35, 947--953.

Poritz, A.B. (1988) *Hidden Markov Models: a guided tour* Proc. ICASSP-88, Vol. 1, 7--13.

Rabiner, L.R., Juang, B.H., Levinson, S.E. and Sondhi, M.M. (1985) *Recognition of isolated digits using hidden Markov models with continuous mixture densities* AT&T Technical Journal, Vol. 64, No. 6, 1211--1234.

Rabiner, L.R. (1989) *A tutorial on hidden Markov models and selected applications in speech recognition* Proc. IEEE, Vol. 77, 257--286.

Soong, F.K. and Rosenberg, A.E. (1988) *On the use of instantaneous and transitional spectral information in speaker recognition* IEEE Trans ASSP, Vol. 36, 871--879.

Tohkura, Y. (1987) *A weighted cepstral distance measure for speech recognition*, IEEE Trans ASSP, Vol. 35, 1414--1422.