INVESTIGATION OF SPEECH AND SPEAKER RECOGNITION BASED ON TRAJECTORY MODELING OF UTTERANCES

W. J. Tey, N. P. Jong, and R. Togneri

Centre for Intelligent Information Processing Systems (CIIPS)
Department of Electrical and Electronic Engineering
University of Western Australia

ABSTRACT - We present in this paper a modelling technique used to capture the dynamic and temporal behaviour of transitions between phonemes. This model relies on the trajectory instead of the geometrical position of the observations in the parameter space. Transition based models provide an alternative method for acoustic-phonetic modelling of the speech signal. In our modelling technique, the trajectory is modelled by regression analysis of low-order polynomials followed by statistical clustering of these coefficients. This technique is used for both speech recognition as well as speaker recognition. Results on a small trial set of isolated alphabet sounds and speakers for both speech and speaker recognition are presented. The speech recognition rate using the trajectory model is found comparable to traditional HMM modelling. However, the poor results for the speaker identification suggest that the current trajectory model is not suitable for this recognition task.

INTRODUCTION

Represented in an N-dimensional parametric space, a running speech signal is a moving point. The trace of this moving point is termed the trajectory of a speech signal (Gong et al., 1991 & 1994). A sound is made up of clusters and transient trajectories in the speech space. Clusters are mostly formed by sustained sounds such as phonemes. Transient trajectories are the transition of the utterances from one cluster to another cluster. In continuous speech, sustained sounds are either very short or completely omitted. This is the reason for the breakdown of the state based HMM modelling (Rabiner, 1989) in continuous speech recognition. This leads to the motivation for building a piece-wise (segmental) transient trajectory model for the continuous speech trajectory (Ostendorf et al., 1989). In this paper, a regression model is applied to the modelling of the trajectory. A low order polynomial is fitted directly to each dimension of the transient trajectory.

TRAJECTORY ANALYSIS

The trajectory of a speech signal can be viewed using *fview* (Lee et al., 1993), a speech research tool developed within CIIPS for visualisation of higher-dimensional data. It provides a geometric and graphical method for inspecting the trajectory formed by a sequence of feature vectors. The trajectory in the speech space is projected onto 2D or 3D for viewing which can be regarded as shadow of the trajectory. Shown below are the trajectories of some utterances.

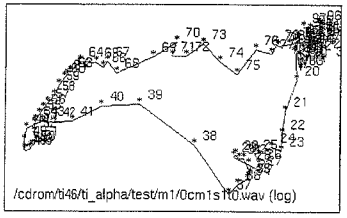

FIGURE 1(a)

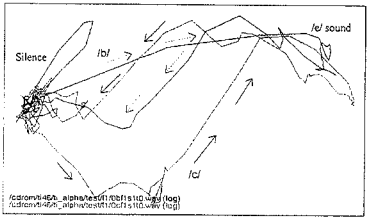

FIGURE 1(b)

Figure 1(a) shows the multi-dimensional feature vectors of the word 'c' projected onto a two-dimensional space. Three clusters can be clearly located for this word. The transient trajectories are the transitions of the feature vector from one cluster to another cluster. Figure 1(b) shows the superimposed trajectory of the two words, 'b' and 'c'. The arrows indicate the direction of the travelling signal. The two words can be distinguished by looking at their transitions from silence to the /e/ sound.

The transitions from the /e/ sound back to the silence, however, had a similar shape. So, the first transition contains the most information that distinguishes the two words. This is the reason that the transition from the silence to the /e/ sound is segmented for this analysis.

TRAJECTORY MODELLING

Since the trajectory is formed by a sequence of vectors of dimension N, it is sensible to model each component dimension of the trajectory separately. The trajectory model is a simple model which attempts to fit each dimension of the trajectories with a polynomial. The coefficients for the polynomials are obtained through a least-square estimation technique [Press et al., 1995]. Coefficients from a trajectory are grouped together to form a feature vector. Two parameterisation techniques are considered, namely, the rate-independent parameterisation and the rate-dependent parameterisation.

The polynomial used to model each dimension of the trajectory is,

$$y(x) = b_1 + b_2 x + b_3 x^2 + \cdots + b_M x^{M-1}$$

Where,

b is the polynomial coefficients
y(x) is a function of the index parameter x

A trajectory formed by the M vector sequence $\mathbf{a}_i \in \mathbf{R}^N$, i = 1, 2, ..., M, are arranged as a design matrix shown below:

$$\begin{bmatrix} \mathbf{a}_{11} & \mathbf{a}_{12} & \cdots & \mathbf{a}_{1N} \\ \mathbf{a}_{21} & \mathbf{a}_{22} & \cdots & \mathbf{a}_{2N} \\ \vdots & \vdots & & \vdots \\ \mathbf{a}_{M1} & \mathbf{a}_{M2} & \cdots & \mathbf{a}_{MN} \end{bmatrix}$$

Each column of the design matrix is modelled separately by a low order polynomial. In this case, the model of a trajectory consists of N polynomials. The recognition is therefore based on the coefficients of the polynomials.

For rate-independent parameterisation, the index parameter, x, is the distance along the trajectory, d_i. Consider the case when two speakers or even the same speaker are speaking at a different rate, in this case the sampling of the trajectory will be different, although the path traced by the trajectory (ie. distance along the trajectory) may be the same. Therefore, rate-independent parameterisation normalises the difference in the speaking rate for different speakers so that the difference in speaking rate does not affect the recognition of an utterance. More formally, given a trajectory specified by the M-vector sequence $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \ldots, \mathbf{a}_M$. The distance of the i^{th} vector along the trajectory, d_i, is defined as

$$d_1 = 0$$
$$d_i = \|\mathbf{a}_i - \mathbf{a}_{i-1}\| + d_{n-1} \qquad i = 2, \ldots, M$$

where $\|\mathbf{a}_i - \mathbf{a}_{i-1}\|$, denote the Euclidean distance between vectors $\mathbf{a}_i$ and $\mathbf{a}_{i-1}$.

For rate dependent parametrisation, the index parameter, x, is simply the frame number of the feature vector, i. The aim is to preserve the difference between speakers, in particular, the rate of the utterance for different speakers, which may be more important for speaker recognition.

The variation in speaking tone or pitch can be a problem in speech recognition. The geometric location of the trajectory is related to the pitch information and one of the objectives of the proposed trajectory model is to model the transient behaviour of the speech signal instead of the geometrical position. A proposed method to overcome the problem is to shift the origin to the first point of the trajectory. This, again normalises the variations of the different speakers.

EXPERIMENTS AND RESULTS

The b, c, d, e, g, p and t (a subset of /e/ set) from the TI46 was chosen for the isolated-alphabet speech recognition and text-dependent speaker identification experiments.

For speech recognition, 3 males and 3 females were chosen to eliminate the gender difference in the speech. The training and testing set both consisted of 336 samples (8 samples per alphabet for each speaker).

For speaker identification, since males and females are seldom confused in speaker recognition, all the 8 male speakers were chosen for the task. Furthermore, only b, c and d are used. The identification experiment is carried out on each alphabet separately. Therefore, for each alphabet, 10 sample utterances from each speaker were used for training and 16 utterances from each speaker were used for testing.

The transient trajectory is manually segmented from the trajectory. With the help of *fview*, approximate start and end, in terms of the frame number, of a transient trajectory of interest are identified. The portion of utterance which corresponds to the transient region is then segmented according to the given start and end.

Short-time analysis is performed on the segmented speech by calculating the spectrum magnitude with an FFT followed by a 16-order Mel-scale Gaussian filter bank averaging of the FFT co-efficients. A 2^{nd} order polynomial is then used to model each dimension of the filter bank coefficients. The resulting coefficients from each transient trajectory are concatenated to form a feature vector. Therefore, a polynomial coefficient vector is formed for each sample of a speaker. A PNN, Probabilistic Neural Network, (Specht, 1988) is used as a classifier for both speech utterance and speaker.

Shown in table 1 are the isolated word recognition results.

Rate independent modelling	Rate dependent modelling
68.56%	79.04%

TABLE 1 Recognition result on the /e/ set alphabets using rate independent and rate dependent trajectory modelling

One conclusion that can be made is that speech recognition is dependent highly on the rate of speaking. By using the rate independent approach, the information in the rate of the word is lost. Therefore the recognition rate with rate independent modelling is worse than that of rate dependent modelling. A typical confusion matrix is shown below,

	b	c	d	e	g	p	t	%correct
b	26		19	1	2			54.17%
c		45	1		2			93.75%
d	6	1	36			1	2	78.26%
e	1			45	1	1		93.75%
g	1	2	2		29	4	10	60.42%
p	2			2		42	2	87.75%
t	1				1	5	41	85.42%
Overall %correct								79.02%

TABLE 2 Typical confusion matrix for speech recognition.

A comparison of using the whole speech utterance and the segmented transient trajectory using the standard three states HMM model was also performed. The result is shown in table 4.

	Whole word utterance	Segmented utterance
Testing on training set	95.22%	99.10%
Testing on testing set	81.19%	93.11%

TABLE 3 Recognition rate of HMM modelling for whole word utterance and segmented utterance.

Increasing the number of states in the HMM whole word modelling did not show any significant improvement in recognition rate. The implication of the results is that the transient trajectory contains more information than the whole utterance. This can be explained by two reasons. Firstly, since there was silence included in the whole word utterance, this might cause confusion to the HMM. But this argument was disproved in the experiment by increasing the number of states where the recognition rate remained quite constant. The second reason is that the trajectory part was not modelled properly and missed entirely in favour of the stationary state regions namely, the silence and vowel regions. This second reason is the more likely explanation given the stochastic state-based modelling approach of the HMM.

The speaker recognition rate, using rate dependent or rate-independent parameterisation, for each word examined is shown below,

	b	c	d	average
Rate independent modelling	32.54%	24.60%	28.5%	28.54%
Rate dependent modelling	34.92%	38.89%	34.1%	35.97%

TABLE 4 Speaker recognition rate for different utterance using two different approaches.

The results are clearly unacceptable for any practical speaker recognition tasks. One of the possible explanations for this is that the vowel stationary state information is important for speaker recognition. However, with trajectory modelling, the state information has been deliberately removed during the transient trajectory segmentation process. This indicates that speaker recognition relies on the state information of the utterance rather than the trajectory information of the utterance, at least within the constraints of the present work. Notice that the speaker recognition rate using rate dependent modelling is better than that achieved using rate independent modelling. Although the difference is not really significant, it does suggest that the rate dependent modelling is possibly more suitable for speaker recognition, which is expected.

For comparison, speaker recognition using HMM on whole word utterance and the segmented trajectory is carried out. The results are shown in the following table.

	Whole word utterance	Segmented utterance
b	93.65%	56.35%
c	81.75%	77.78%
d	84.25%	65.87%
average	86.55%	66.67%

TABLE 5 Comparison of speaker recognition rate using HMM.

The speaker recognition based on the transient information is about 20% worse than that of the whole word utterance. Again, this seems to imply that the transient trajectory holds little information about the speaker and it is the state information rather than the trajectory information that is more important for speaker identification.

CONCLUSIONS

This paper presents a simple model for a trajectory. The model consists of a set of polynomials. Current model suggests that the transient trajectory model is more suitable for speech recognition tasks rather than speaker recognition. The better recognition achieved based on only the transient trajectory using the HMM showed that the transient trajectory contains most of the information for speech recognition. The recognition rate using the trajectory model presented is still worse than that achieved by the HMM. This indicates that the current modelling technique for the trajectory is less precise than the HMM, possibly due to the simplicity of the current modelling approach. However, it does show that the transient trajectory needs special attention during speech recognition. The poor results for speaker identification using transient trajectory suggest that state information rather than trajectory information is more important for speaker recognition.

Another important aspect which needs further investigation is the true interpretation of rate dependent and rate independent parametrisation. Although the current experiment carried out shows that the rate dependent approach is better than the rate independent approach for both speech and speaker recognition, it is possible that our definition of rate independent modelling eliminates some valuable information for speech and speaker recognition.

Some future work under investigation includes the use of a more robust and flexible modelling technique for trajectory modelling rather than a simple polynomial fit. Advanced techniques like dynamic modelling of non-linear time series (Judd, 1995) and incorporation with the HMM or similar statistical framework for robust training of the models (thereby avoiding manual segmentation of trajectories) (Deng, 1994) are currently being investigated.

REFERENCES

Deng, L., Aksmanovic, M., Sun, X. & Wu, C.F.J. (1994) *Speech recognition using hidden Markov models with polynomial regression functions as nonstationary states*, IEEE Trans. Acoust. Speech and Signal Processing, vol. 2, no. 4, 507-520.

Gong, Y. & Haton, J.P. (1990) *Text-independent speaker recognition by trajectory space comparison*, IEEE International Conference on Acoustics, Speech and Signal Processing, April 1990, 285-288.

Gong, Y. & Haton, J.P. (1994) Stochastic trajectory modelling for speech recognition, IEEE International Conference on Acoustics, Speech and Signal Processing, April 1994, 57-60.

Krishnan, N & Rao, P.V.S. (1994) *Segmental phoneme recognition using piecewise linear regression*, Proc ICASSP, I49-52.

Lee, G. & Alder, M.D. (1993) *A geometric interpretation of speech features*, Proc. of the First Australian and New Zealand conference on Intelligent Information Processing Systems.

Ostendorf, M & Roukus, S (1989) *A stochastic segment model for phoneme-based continuous speech recognition*, IEEE Trans. Acoust. Speech and Signal Processing, vol. 37, no. 12, 1857-1869.

Ostendorf, M. & Diagalakis, V. (1991) *The stochastic segment model for continuous speech recognition*, Proc. ICASSP, 964-968.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. (1995) *Numerical recipes in C*, 2nd ed., 671-681.

Rabiner, L.R. (1989) *A tutorial on hidden Markov models and selected application in speech recognition*, Proc. IEEE, vol. 77, no. 2, 257-285.

Reynolds, D.A. & Rose, R.C (1995) *Robust text-independent speaker identification using Gaussian mixture speaker models*, IEEE Trans. Speech and Audio Processing, vol. 3, no.1, 72-83.

Specht, D.F. (1988), *Probabilistic neural Network for classification, mapping or associative memory*, IEEE Conference on Neural Networks, vol. 1, 525-532.

Thomson, M.M. (1995) *Statistical modelling of speech feature vector trajectories based on a piecewise continuous mean path*, IEEE International Conference on Acoustics, Speech and Signal Processing, 361-364.

Judd, K. and Mees A. (1995) *On selecting models for nonlinear time series*, Physica D, vol. 82, 426-444.