TRACE-SEGMENTATION OF ISOLATED UTTERANCES FOR SPEECH RECOGNITION

Euvaldo F. Cabral Jr.
University of São Paulo -DEE/EPUSP
Laboratory of Communications and Signals
Caixa Postal 8174
São Paulo, SP, 01065-970, Brazil

Graham D. Tattersall
University of East Anglia
Department of Electronic Engineering
School of Information Systems
Norwich, Norfolk, NR4 7TJ, England, UK

ABSTRACT

Trace-segmentation (TS) is a method for non-linear time-normalization of a sequence of speech representation frames prior to recognition of the sequence. Numerous attempts to perform speech recognition using trace-segmentation have been made in the past but these attempts have failed to provide the same performance as DTW of HMM recognition. The reason for this failure may be due to the use of inappropriate distance metrics to perform the segmentation or the use of an inappropriate spatial sampling interval along the trace. This paper describes an investigation into these problems, in which the appropriate Nyquist sample rate of the spatial trace is determined by analyzing the frequency of the temporal variation of the speech frames. It is also shown that separate segmentation of the trajectory described by each individual coefficient in the speech frame leads to much improved recognition which exceeds the performance provided by DTW recognition of the same database.

1. INTRODUCTION

One of the outstanding problems of speech recognition is how to deal with the temporal variability of speech. Conventionally, temporal variability is handled by the use of dynamic time warping (DTW) or hidden Markov models (HMM). These techniques work surprisingly well even though the implicit speech production model used in DTW and HMM may have little connection with real speech.

However, it is often suggested that the speech production models used in HMMs are not representative of real speech production and artificial neural net (ANN) techniques have been widely trailed as an alternative. Thus far, the performance of ANN based recognition compares unfavorably with HMM based recognition. One problem is that feedforward ANNs such as the multi-layer perceptron (MLP) have a fixed dimensionality input field to which the variable number of speech frames in an utterance must somehow be sensibly applied.

A number of approaches have been adopted to solve this problem. The first is to introduce time delays between layers in feed forward networks so that the frames in an utterance can be applied separately in sequence to the input of the network [1]. The presence of the time delays ensures that information about previous frames is retained in the network so that the current outputs of units within the network are fed back to the inputs of units in previous layers. This recurrent architecture retains some information about all previous inputs rather than just a few as in the feedforward TDNN. However, the choice of recurrent architectures is enormous and learning often proves unreliable [2].

Another approach which can be used when recognizing isolated utterances is to perform linear time normalization of the utterance. This means that a small constant number of frames are selected from the original frame sequence representing the utterance. The selected frames are chosen to be at approximately uniform time intervals regardless of the time length of the utterance. The selected frames are then concatenated into a single vector of fixed dimensionality which can be applied as an input to a conventional feedforward neural net or conventional pattern classifier [3]. In spite of its crudity, this approach works well on isolated words but is inferior to HMM or DTW classification because no attempt is made to model the time variability in the utterance.

Finally, time variability can be accommodated by the technique of trace-segmentation (TS) which is sometimes called *variable frame rate coding*. The idea behind TS is that the sequence of frames representing an utterance describe a trace or trajectory through a space whose dimensions are the dimensions of the frame vector. It is proposed that the shape of the trace is characteristic of the utterance and that the shape should therefore be encoded as a vector of fixed dimensionality which can be applied to a feedforward ANN or conventional pattern classifier. A possible encoding of the trace shape is to concatenate the series of vectors at a number of uniformly spaced points along the trace and the process of selecting the spatially equidistant points along the trace is called *trace-segmentation*.

This approach has been explored by many workers in the field [4], [5], but has not led to better performance than obtained by *linear time normalization*. It is believed that this may be due the use of an unsuitable distance metric to segment the trace or because the Nyquist sample rate criterion is being violated in choosing the spatially equidistant samples along the trace. This paper examines both these possibilities and determines the appropriate sample rate for each coefficient in the speech frame. It is also shows that the trace-segmentation process is much more successful when applied to each frame coefficient individually, rather that the frame as a whole. The

0-7803-2431-5/95 $4.00

latter approach is called *Individual Trace-Segmentation* (ITS) and it will be shown that used in conjunction with a k-nearest neighbor classifier, it provides significantly better performance than simple trace-segmentation or DTW on a highly confusable vocabulary.

Although the rationale for investigating trace-segmentation is to generate a fixed dimensionality representation of an utterance which can be classified by a feedforward ANN, this paper demonstrates the properties of the trace-segmentation algorithms in conjunction with a k-nearest neighbor classifier as it is believed that the use of an ANN would add an unnecessary uncertainty to the results.

2. SIMPLE TRACE-SEGMENTATION (TS) AND THE ITS ALGORITHM

Consider an utterance u represented by a sequence of m N-dimensional speech frames $(v_1^u, v_2^u, \ldots, v_m^u)$ taken at uniform intervals of time. These vectors are non uniformly spaced points on a curved trace through the N-dimensional space. The process of simple trace-segmentation (TS) involves division of the total spatial length, L_t, of the trace into n_p equal length segments and definition of the beginning and end of each segment by an appropriate N-dimensional vector. The concatenation of these vectors is then used as an encoding of the shape of the trace. The beginning and end segments vectors are usually made equal to the spatially nearest vector, v_j^u in the original speech sequence. If greater accuracy is required, the segment vectors can be estimated by linearly interpolating between the two nearest speech vectors, v_j^u and v_{j+1}^u, which lie on either side of the segment boundary. The TS algorithm can be summarized as follows:

i) *Estimate the total spatial length, L_t, of the trace. This is done by summing the distances between the m successive vectors in the original speech sequence.*

$$L_t = \sum_{i=1}^{m-1} d(\vec{v}_i, \vec{v}_{i+1}) \tag{1}$$

where

$$d(\vec{v}_i, \vec{v}_{i+1}) = \left(\sum_{j=1}^{N} (v_{i,j}^u - v_{i+1,j}^u)^2\right)^{1/2} \tag{2}$$

The length of each spatial segment is then:

$$L_s = L_t / n_p \tag{3}$$

ii) Find the two vectors in the original speech sequence which lie on either side of the i^{th} segment boundary. This is done by finding the value of j for which the following inequalities are true:

$$\sum_{k=1}^{j} d(v_k^u, v_{k+1}^u) > i.L_s \tag{4a}$$

$$\sum_{k=1}^{j-1} d(v_k^u, v_{k+1}^u) < i.L_s \tag{4b}$$

iii) Linearly interpolate between v_j^u and v_{j+1}^u to find the segment boundary vector t_i^u

$$t_i^u = v_j^u + (v_{j+1}^u - v_j^u).\alpha \tag{5}$$

where α is given by:

$$\alpha = \frac{iL_s - \sum_{k=1}^{j-1} d(v_k^u, v_{k+1}^u)}{d(v_j^u, v_{j+1}^u)} \tag{6}$$

The ITS algorithm applies the original trace segmentation strategy of linear interpolation in a N-dimensional space to a one-dimensional space. The ITS algorithm is the same as the TS algorithm except that it must be applied N times, i.e. one for each component of N-dimensional vector of features, and (2) becomes

$$d(\vec{v}_i, \vec{v}_{i+1}) = \left| v_i^u - v_{i+1}^u \right| \tag{7}$$

3. THE SPEECH DATABASE

All the analysis and experimental work described in this paper was based on a subset of the Connex Alphabet Database (binary version 0.1) from British Telecom Labs. This speech consists of three examples of each of the letters of the alphabet uttered by a total of 104 speakers. The speech was recorded in a silence cabinet through a high quality handset, digitally sampled at 20 KHz using a 16 bits A/D. The subset chosen for the trace-segmentation work consisted of 6 letters - B, P, M, R, S and T. Fifty two speakers have been designated training talkers and other fifty two for testing.

The time domains samples were converted to eight dimensional MFCC frames at a one millisecond rate. This unusually high frame sample rate was chosen to ensure that none of the coefficients in the frame were undersampled. This is important since one of the objectives of the work was to find the appropriate Nyquist rate for each coefficient and this could not be done if aliasing had already been produced by producing the frames too low rate.

4. FINDING THE FREQUENCY VARIATION OF THE MFCCs

An initial investigation into the frequency variation of each of the coefficients in the speech frame was made to ensure that any failure in the performance of the trace-segmentation algorithm could not be attributed to undersampling of the speech representation. The investigation was done by evaluating the power spectral density function for the waveform described by each coefficient in the sequence of speech frames for each utterance in the database. These functions were averaged over the entire set of utterances to provide a picture of the magnitude of the different frequencies of variation of each of the coefficients in the speech frame representation.

The average functions for each of the eight MFCCs in the speech frames are presented in Table 1. It can be seen that the maximum significant frequency of variation in any coefficient is lesser than 100 Hz. This coincides with the common assumption that articulation frequencies are limited to 100 Hz, and thus that the sample rate for the speech frames should no be less than 5ms - 10 ms.

Coefficient	3 dB frequency (±2 Hz)	Maximum Significant Frequency (±5 Hz)
C0	28	95
C1	17	53
C2	16	78
C3	7	46
C4	7	52
C5	12	58
C6	10	95
C7	13	90
C8	10	87

Table 1: 3 dB and maximum significant frequency (Hz) of each MFC coefficient

5. RECOGNITION EXPERIMENTS USING THE TS AND ITS ALGORITHMS

The usefulness of the TS and ITS algorithms were accessed by conducting recognition experiments with a k-nearest neighbor classifier in conjunction with a number of pre-computed class template patterns. For both the TS and ITS experiments, ten templates per class were generated using the Modified k-Means Clustering Algorithm [6]. Classification was then attempted using the k-nearest classifier with a number of different k values.

The first experiment was applied to the TS algorithm to determine the appropriate number of spatial segments which should be used. Recognition was attempted using 20, 40, 100 and 200 segments per utterance in conjunction with a k-nearest neighbor classifier using k values of 1 and 2. The results presented in Table 2 show that the best number of segments is 20, with performance deteriorating as the number of segments is increased. The performance obtainable using DTW recognition of the same data at 1 ms frame rate is shown for comparison.

Having determined that 20 segments provided best performance with the TS algorithm, a comparison was made between the TS and ITS algorithms using an extended range of k-nearest neighbor classifiers. Table 3 presents the recognition results of the speech recognition experiment for TS and ITS for ten values of the parameter k of the k-nearest neighbor classifier. It is evident that ITS is greatly superior to the simple trace segmentation algorithm, and is significantly better than DTW for all the tested k values.

K	average	mode
1	71.1	TS 20
1	68.5	TS 40
1	68.9	TS 100
1	67.2	TS 200
1	73.3	DTW
2	72.1	TS 20
2	66.2	TS 40
2	66.6	TS 100
2	67.2	TS 200
2	72.5	DTW

Table 2: Recognition performance using the TS algorithm with various numbers of segments per utterance

Type of Recognition	TS	ITS	DTW
k=1	71.1	76.7	73.2
k=2	72.1	77.3	73.2
k=3	68.2	80.3	77.1
k=4	68.4	79.7	76.8
k=5	60.9	79.1	77.1
k=6	58.6	78.7	78.7
k=7	58.2	78.1	78.4
k=8	56.6	77.5	76.4
k=9	53.0	76.8	74.8
k=10	53.7	74.5	72.5
average	62.1	77.9	75.8

Table 3: Recognition performance using the ITS and TS algorithms

6. DISCUSSION

The work described in this paper has sought to provide understanding of the relatively poor performance provided by conventional trace-segmentation applied to speech recognition, and to test improved versions of the algorithm.

It was initially believed that one reason for the poor performance of the trace segmentation might be sub Nyquist sampling of the coefficients in the speech frames. However, an

analysis of the frequency of variation of the coefficients during the course of an utterance shows that the energy in frequencies above 100 Hz is very small. This supports the conventional belief that a frame sample interval of 5 ms to 10 ms is adequate. However, it should be noted that it is still possible that short lived, class discriminatory variations in the coefficients may take place, but which contribute little high frequency energy because they last for such a short time. The frequency analysis presented in this paper does not therefore prove that a frame sample rate of 5 ms is adequate for all speech events.

An extension of the notion that the performance of trace-segmentation based recognition is poor because of sub-Nyquist sampling, is that insufficient spatial sample points are defined along the utterance trajectory. This hypothesis has been tested by comparing the performance obtained on the same problem using various numbers of spatial sample points. The results are counter intuitive: increasing the number of sample points causes performance to deteriorate, and the optimum number of points appears to be about 20 per utterance.

In view of these results, it was suspected that the poor performance of the trace-segmentation based recognizer was more probably caused by the use of an inappropriate distance metric for segmentation. An obvious defect in conventional trace-segmentation, (TS), is that the coefficients exhibiting the most variation during the utterance will tend to dominate the distance measurements upon which segmentation is based. High variation may be due to noise rather that phonetically significant events in the speech and so the ITS algorithm was proposed in which each coefficient's trajectory is segmented separately. This leads to a very marked improvement in performance which even exceeds that obtainable using DTW.

7 ACKNOWLEDGMENTS

The authors are indebted to British Telecom PLC for the access to the Conex Database and to the Department of Electronics in the UEA for the use of its computational facilities. This work has been supported by a grant (733/89-8) from CAPES, the Brazilian Federal Research Sponsorship Agency.

8. REFERENCES

[1] A. Waibel et al., *Phoneme Recognition Using Time Delay Neural Networks*, IEEE Trans. on ASSP, Vol. 37, No. 3, March 1989.

[2] L. B. Almeida, *Backpropagation in Non-feedforward Networks*. Chapter 5 of *Neural Computing Architectures*, I. Alexander, North Oxford Academic Press, 1989.

[3] P. W. Linford and G. D. Tattersall, *Non-linear Time Normalization of Utterances For Speech Recognition Using MLP's*, Proc. of the Inst. of Acoust., Vol. 12, Part 10, 1990.

[4] M. H. Kuhn, H. Tomaschewski, and H. Ney, *Fast Non-linear Time Alignment for Isolated Word Recognition*, in Proc. ICASSP, Atlanta, GA, pp. 736-740, March 1991.

[5] R. Pieraccini, *Pattern Compression in Isolated Word Recognition*, Signal Processing, Vol. 7, pp. 1-15, 1984.

[6] J. Wilpon and L. Rabiner, A Modified K-Means Clustering Algorithm, IEEE Trans. on ASSP, Vol. 33, No. 3, June 1985.