PROSODY AS A BASIS FOR DETERMINING SPEAKER CHARACTERISTICS

Michael G. Barlow and Michael Wagner
Department of Computer Science
University College
University of New South Wales

ABSTRACT - A speaker identification experiment based on prosodic features is described. Five speakers recorded a set of four sentences in five separate sessions over a period of one week. For each of these utterances, the energy, fundamental frequency, voicing and linear prediction error contours were extracted. For each sentence (four) and each type of contour (four) distance measures based on dynamic time warping were calculated between all twenty five (five speakers by five repetitions) contours. These distances were compared on an inter-speaker versus intra-speaker basis and the ratio was generally found to be large. Parameters within the distance measuring process, namely warping window size and contour smoothing, were altered and the effects on speaker distances are discussed.

INTRODUCTION

Speaker-characteristic information is encoded in both the dynamic and static features of a speech waveform. Previous research has shown that much speaker-characteristic information is encoded in dynamic features of the speech waveform at a prosodic level (Atal, 1972; Williams and Stevens, 1972; Barlow and Wagner, 1986); yet much investigation is still required.

The approach adopted in this paper is to compare contours of acoustic parameters extracted over the duration of an utterance. These comparisons are performed on an inter-speaker versus intra-speaker basis in an attempt to show a correlation between speaker identity and contour similarity.

The following sections describe the speech data analysed, the contours extracted, the means of contour comparison, the conduct of experiments, and the results obtained.

SPEECH DATA

Five male native speakers of Australian English between the ages of twenty four and forty were asked to record a series of fourteen different sentences in five separate sessions over the period of one week. From this database of utterances four sentences were selected for examination:

1) "Cool shirts please me."
2) "I cannot remember it."
3) "Pay the man first please."
4) "Papa needs two singers."

DATA ANALYSIS

All utterances of the four sentences from all five speakers were digitised at 16kHz with 12-bit quantisation, leading to a test-bed of one hundred distinct utterances. Sentence boundaries were determined automatically, based on energy and zero crossing thresholds (Rabiner and Sambur 1975). Linear predictive analysis was used to determine the source parameters and vocal tract transfer function. A 32 ms analysis window was moved across the data in steps of 16 ms. The input signal was pre-emphasised and Hamming windowed, and the inverse filter A(z) of order 20 and the excitation energy were calculated (Wagner & Fulcher, 1986).

CONTOUR EXTRACTION

The parameters (1) energy, (2) voicing, (3) fundamental frequency and (4) linear prediction error were selected for analysis. All values were extracted from frames of length 32 ms.

Energy was measured in dB and normalised for the frame size. Voicing, F0, and LPC error were

extracted via the auto-correlation method of linear prediction analysis (Markel and Gray, 1976). For unvoiced frames F0 was assigned the value of the minimum F0 value over the entire utterance.

A contour of a given parameter is considered as the sequence of individual values of that parameter over the duration of the utterance .

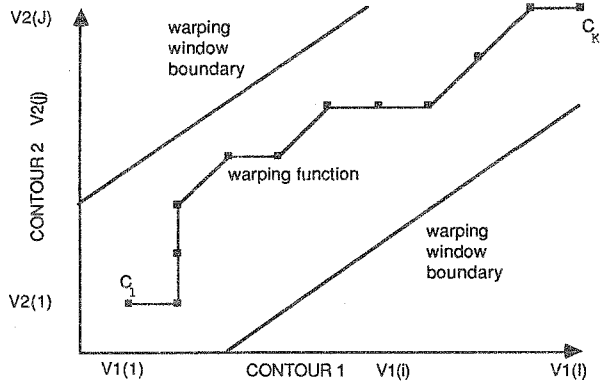

Figure 1. Dynamic time warping: Warping function $C_K = (i_K,j_K)$, k = 1,...,K. Boundaries $C_1 = (1,1)$ and $C_K = (I,J)$. Distance measure $E = \Sigma \mid V1_{ik} - V2_{jk} \mid$, k = 1,...,K.

CONTOUR COMPARISON

The distance between contour values was determined after normalising the contours into the range 0 to 1.

Dynamic time warping (DTW) (Sakoe and Chiba, 1978) was then used to compute the warping function C_K corresponding to the best path match between the normalised contours. Total distance E was defined as the normalised sum of the individual distances between points on the contours along the path of best match (see figure 1).

EXPERIMENTAL METHOD

The goals of the experimentation were two-fold. Firstly to investigate the viability of using a DTW scheme to compare contours, and if applicable to select the 'optimal' parameters of the DTW scheme. Secondly, if the scheme appeared viable, to investigate the correlation of speaker identity to contour shape.

Viability Experiment

Various parameters of the DTW process of contour comparison are alterable and were investigated to determine their influence upon contour distance measures. The selection criterion was the optimisation of intra-speaker versus inter-speaker contour distance. Parameters investigated were pre-processing smoothing and warping window size.

Smoothings investigated were none, mean 3, median 5, and median 5 mean 3 (Hess, 1983). With C being the original contour, N being the new contour, and L the contour length, these smoothings are formulated as follows:-

mean 3: $N_1 = C_1, N_L = C_L$
$N_i = 0.25C_{i-1} + 0.5\ C_i + 0.25C_{i+1},\ i = 2,...,L-1;$

median 5: $N_1 = C_1, N_2 = C_2, N_{L-1} = C_{L-1}, N_L = C_L$
$N_i = \text{median}\ [\ C_j,\ j = i-2,...i+2\ ],\ i = 3,...,L-2;$

median 5 mean 3: Linear application of median 5 followed by mean 3 smoothing.

Warping window size (see figure 1) is the allowable variance in time, as a fraction of contour size, that best path matching may occur over. Values of 5%, 10%, 20% and 50% of contour size were investigated.

Speaker Identity Experiments

Sentence number 1 was selected to use in a straight speaker identification experiment in which each of the twenty-five utterances was declared as being from a given speaker if the distance was minimal. Sentence numbers 2,3, and 4 were selected to determine the amount of variability between intra-speaker and inter-speaker distance scores; with a ratio of mean inter-speaker to mean intra-speaker distance being calculated.

For each sentence (4) and each of the contour types (4), a twenty five (5 speakers by 5 repetitions each) by twenty five distance table was constructed. Each table cell corresponding to the calculated distance between the corresponding contours that 'indexed' that cell. This table was then processed to give the speaker identification and contour variability results.

For the speaker identification system each contour was compared against the four other contours from the same speaker. A mean distance was calculated. Then the same contour was compared against each of the five contours from the other four speakers. Again, a mean distance was derived for each speaker. The utterance was then identified as being from the speaker whose mean distance was smallest. These results were then examined with an eye to the number of correct identifications.

For the inter-speaker versus intra-speaker variability experiment the contours were clustered into five sets of five corresponding to their speaker. For each set two distances were calculated. The first, the intra-distance, is the mean of the distances between contours within the set. The second, the inter-distance, is the mean of the distances between each contour within the set and all contours outside the set. For each speaker a ratio of (inter-speaker mean)/(intra-speaker mean) was calculated and the average of this ratio over all speakers was determined.

RESULTS

Viability Investigation

Tables 1A and 1B show the effects of manipulation of smoothing and warping window size on the speaker identity measure for each of the four contour types: energy, voicing, LPC error and F0.

Table 1A shows the effect of window size variation; with 5%, 10%, 20% and 50% of contour duration being investigated while no smoothing or other variation in distance measuring parameters occured. It can be seen that a window size of 10% for energy and voicing; and 20% for LPC error and F0 produce the highest ratios.

Table 1B shows the effect of pre-distance measure smoothing; with none, mean 3, median 5, and the combination median 5 mean 3 being investigated. Using the results obtained from the warping window size variation experiment (table 1A) a window size of 10% for energy and voicing, and 20% for LPC error and F0 was selected. Maintaining these warping window sizes as constant, the three smoothers were applied to the contours before contour distances were computed. It may be observed that for energy, voicing, and LPC error the combination smoother median 5 mean 3 yielded significantly higher ratios and that for F0 median 5, and median 5 mean 3 both yielded the highest

ratios.

WARP WINDOW SIZE	CONTOUR											
	Energy			Voicing			LPC error			F0		
	s2	s3	s4	s2	s3	s4	s2	s3	s4	s2	s3	s4
5%	1.36	1.32	1.36	1.31	1.32	1.16	1.33	1.35	1.25	1.16	1.20	1.08
10%	1.44	1.44	1.46	1.35	1.38	1.18	1.43	1.51	1.36	1.20	1.25	1.13
20%	1.35	1.39	1.38	1.31	1.31	1.15	1.46	1.51	1.47	1.24	1.33	1.19
50%	1.19	1.24	1.24	1.22	1.17	1.06	1.43	1.36	1.41	1.32	1.27	1.14

Table 1A

SMOOTHING	CONTOUR											
	Energy			Voicing			LPC error			F0		
	s2	s3	s4	s2	s3	s4	s2	s3	s4	s2	s3	s4
None	1.44	1.44	1.46	1.35	1.38	1.18	1.43	1.51	1.36	1.24	1.33	1.19
Mean 3	1.49	1.58	1.55	1.46	1.58	1.24	1.51	1.61	1.50	1.17	1.42	1.21
Median 5	1.41	1.55	1.53	1.37	1.49	1.23	1.49	1.60	1.41	1.23	1.46	1.22
Median 5 Mean 3	1.46	1.63	1.61	1.43	1.60	1.26	1.54	1.70	1.44	1.22	1.46	1.21

Table 1B

Table 1. DTW Parameter analysis results. For each of the parameters (4) and each sentence (3) the ratio of mean inter-speaker to intra-speaker distance was calculated as various parameters of the DTW process were altered. Table 1A shows results of DTW warp window size variance. Table 1B shows results of pre-processing smoothing.

Speaker Identity Results

Speaker identity as manifest in prosodic features was investigated in two ways. Firstly, the ratio measure of mean inter-speaker distance to mean intra-speaker distance was calculated for all four contour types using sentence numbers 2, 3, and 4. Secondly, a speaker identification experiment was performed using sentence number 1 and the three contours energy, voicing, and LPC error.

Table 2 summarises the results obtained from the parameter variation experiments (Tables 1A and 1B) and shows for each sentence and contour type combination the ratio obtained using a window size of 10% for energy and voicing, and 20% for F0 and LPC error; and pre-distance calculation smoothing of median 5 mean 3. It may be observed that energy, voicing, and LPC error yield the highest ratio measures, with F0 significantly lower.

SENTENCE NUMBER	CONTOUR			
	Energy	Voicing	LPC error	F0
2	1.46	1.43	1.54	1.22
3	1.63	1.60	1.70	1.46
4	1.61	1.26	1.44	1.21
Average	1.57	1.43	1.56	1.30

Table 2. Mean inter-speaker distance versus intra-speaker distance ratios. For each sentence/ contour combination the ratio of mean inter-speaker to mean intra-speaker distance was calculated using Median 5 Mean 3 pre-smoothing and a warp window size of 10% for energy and voicing, and 20% for LPC error and F0.

CONTOUR	PERCENTAGE CORRECT IDENTIFICATION
Energy	92%
Voicing	75%
LPC error	96%

Table 3. Percentage correct speaker identifications for the sentence: "Cool shirst please me" using energy, voicing and LPC error contours with median 5 mean 3 pre-distance measure smoothing and a warp window size of 10% for energy and voicing, and 20% for LPC error.

Table 3 shows the results of the speaker identification experiment run using sentence number 1 and the contour types: energy, voicing and LPC error. Again, the smoother median 5 mean 3 was applied prior to distance calculation and a warp window size of 10% for energy and voicing, and 20% for LPC error was set. For each of the three contour types investigated a recognition score is given showing the percentage of unknown contours (5 speakers by 5 repetitions giving 25 test contours) that were identified as belonging to the correct speaker based upon minimum distance measure. It can be seen that all three contour types yielded identification rates significantly above chance, with energy, and LPC error yielding rates in excess of 90%.

CONCLUSION

From the experiments conducted we conclude that contour matching is a meaningful approach to the examination of speaker characteristics in continuous speech.

The mean inter-speaker to mean intra-speaker ratio, and speaker identification results both show the usefulness of this technique for extracting speaker identity. While the inter-speaker to intra-speaker ratios are somewhat lower than expected we attribute much of this to poor sentence boundary

selection via the automatic system. We believe that hand boundary selection, or an automatic system written specifically for a sentence structure would yield far higher ratios.

Results show that appropriate pre-processing of contours (smoothing) and appropriate selection of distance measuring parameters (warp window size) led to higher recognition rates and a general increase in the ability to extract more speaker specific data.

We conclude that contour matching as a means of measuring speaker characteristics in continuous speech is a valid technique and that manifestations of speaker identity are readily detectable in voicing, F0, energy, and LPC error contours at a sentence level. Further application and refinement of this technique should yield further significant results.

ACKNOWLEDGEMENTS

We would like to thank the following speakers who contributed their valuable time in performing the recordings: Alan Beswick, Lawrie Brown, Jeff Colin, and David Purdue.

REFERENCES

Atal, B.S. (1972) *Automatic speaker recognition based on pitch contours,* J. Acoust. Soc. Am., Vol 52, No. 6, 1687-1697.

Barlow, M., Wagner, M. (1986) *Effects of acoustic parameter alteration upon perceived speaker characteristics,* Proc. 1st Aust. Conf. Speech Sci. & Techn., Canberra.

Hess, W.,(1983) *Pitch determination of speech signals,* (Springer-Verlag: Berlin).

Markel, J.D., Gray, A.H. (1976) *Linear prediction of speech,* (Springer-Verlag: Berlin).

Sakoe, H., Chiba, S. (1978) *Dynamic programming algorithm optimisation for spoken word recognition,* IEEE Trans. Acoust. Speech and Signal Proc., vol. 26, No. 1.

Rabiner, L.R., Sambur, M.R. (1975) *An algorithm for determining the end-points of isolated utterances,* Bell Syst. Tech. J., 54, 297-315.

Wagner, M., Fulcher, J. (1986) *An IBM PC based speech research work station,* Proc. 1st Aust. Conf. Speech Sci. and Techn., Canberra.

Williams C., Stevens, K. (1972) *Emotions and speech: some acoustical correlates,* J. Acoust. Soc. Am., Vol. 52, No. 4, 1238-1250.