EFFECTS OF ACOUSTIC PARAMETER ALTERATION UPON PERCEIVED SPEAKER CHARACTERISTICS

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

ABSTRACT - A set of experiments to observe the relationship between certain acoustic parameters and perceived speaker characteristics are described. Linear predictive analysis was performed upon a sentence spoken by a group of three speakers. Utterances of each speaker were then altered via pitch variance, time alignment and pitch interchange. These altered sentences were then resynthesised and played back to an audience of listeners. The listeners' responses to these sentences were recorded and analysed.

INTRODUCTION

The relationship of acoustic parameters to speaker characteristics is a well known problem in speech analysis (Williams & Stevens,1972; Brown,Strong & Rencher,1973; Takagi & Kuwabara,1986). Understanding of this relationship leads to more natural synthesis of speech and more robust recognition systems.

The approach taken in this paper is to extract the parameters from the speech data, alter them in a consistent way, resynthesise the speech and play it back to a group of listeners. Listeners responses are then analysed.

Following sections will describe the speech data used in the experiments, the means of parameter extraction and resynthesis, the parameter manipulations carried out, experimental procedure, and the results of the listener experiments.

ORIGINAL DATA

Speech data from 2 Australian and 1 North American speakers (with average fundamental frequency of 141 Hz, 88hz, and 117 Hz respectively; here after known as A,B & C) uttering the following sentences was used in the experiments:-

1. "Cool shirts please me."
2. "Pay the man first please."
3. "I cannot remember it." (Wagner,1978).

PARAMETER EXTRACTION AND RESYNTHESIS

Linear predictive analysis was used to determine the source parameters and the 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).

To determine the phoneme timing of each speaker the test sentences were segmented by hand into their component phonemes. This segmentation was performed through observation of the time waveform and spectrum, as well as listening to the digitised data.

Following parameter alteration resynthesis was performed using the inverse filter, source energy and pitch, and a voicing parameter determined by the linear predictive analysis.

PARAMETER MANIPULATION

Four types of parameter manipulation were performed upon the sentences:-

(a) Pitch scaling
(b) Pitch transfer
(c) Time alignment
(d) Pitch transfer and time alignment

Pitch Scaling

The fundamental frequency for each speaker was varied in steps of 10% from 70% to 130% of their original values for each frame.

Pitch Transfer

The fundamental frequency contour of each speaker was interchanged with the contours of each of the other two speakers. The transfer was aligned at phoneme boundaries and linear interpolation was performed between boundaries to take into account differences in articulation time for each speaker (see figure 1).

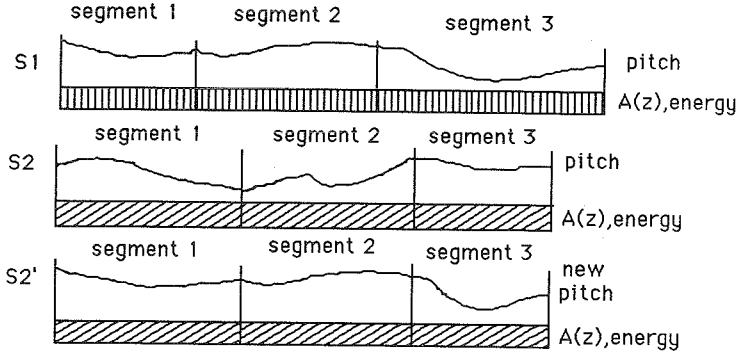

Figure 1. Transfer of pitch from S1 to S2 obtaining S2'. Vocal tract parameters remain constant while pitch becomes the linearly interpolated pitch from S1 (interp. to segment boundaries).

Time Alignment

The phoneme duration of each speaker was altered to conform to that of each of the other two speakers. The alignment was performed at phoneme boundaries and linear interpolation was again used between boundaries to generate new fundamental frequency and linear prediction co-efficients. See figure 2.

Time Alignment and Pitch Transfer

For time alignment and pitch transfer the fundamental frequency contour of each speaker was interchanged with the contours of each of the other speakers along with that speaker's phoneme timing.

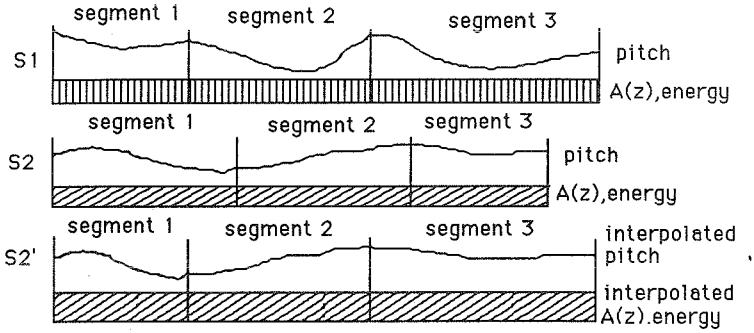

Figure 2. Time alignment of S2 to S1 to obtain S2'. Pitch and vocal tract parameters are derived from S2, but linearly interpolated to conform to the segment timing of S1.

LISTENING EXPERIMENTS

A tape was created recording the resynthesised altered sentences from the three speakers. The tape consisted of the 2 training sentences, followed by the 39 test sentences repeated twice in a random order.

5 native and 2 non-native speakers of Australian English took part in the experiment. These subjects first listened to the training sentences then to each test sentence three times in rapid succession and were asked to determine the speaker's identity and nationality, as well as note any unusual features about the speaker's voice.

RESULTS

Pitch Scaling

Figure 3 shows the results of the pitch scaling experiments. It can be seen that a drop in pitch for speaker A leads many listeners to believe that B was the speaker; similarly a rise in pitch for B led some listeners to believe that the speaker was A. This is probably due to the average pitch of speaker A (141 Hz) being significantly higher than that of speaker B (88 Hz), and absolute pitch therefore being used as the primary cue to distinguish between A & B. Speaker C's personality was far less susceptible to alteration.

The perceived dialect tended to follow the perceived identity, in the sense that if speaker C (the North American) was perceived as speaker B (one of the Australians) he was also perceived as having an Australian accent.

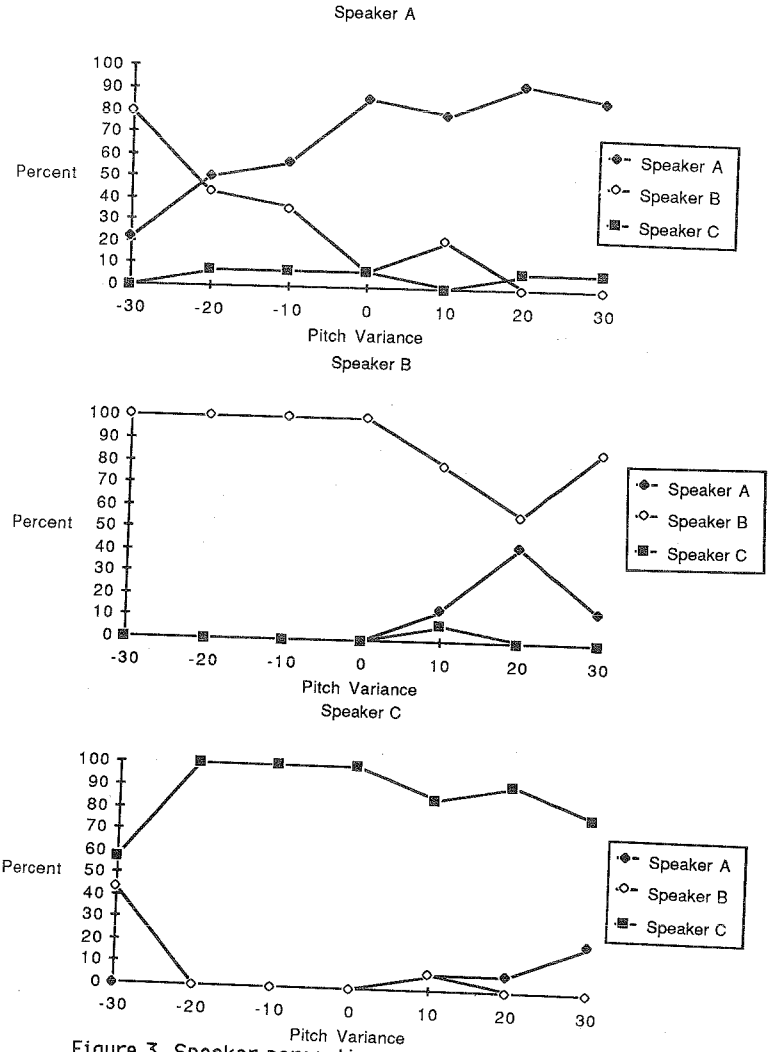

Figure 3. Speaker perception results for pitch variance experiment. Each chart represents the different recognition rates for a particular speaker (A,B and C respectively). Pitch was varied from 70% (-30) to 130% (+30) of its original value. The three lines for each chart represent the percentage of listener responses that identified the talker as one of the three speakers [A=filled diamonds, B=hollow diamonds, C=filled squares].

Pitch Transfer

Table 1 shows the results of the perceived identity of pitch transferred speakers. It can again be seen that speaker B, and particularly speaker A's perceived identity are subject to change. Speaker C, however, remains perceived as C, possibly due to the use of dialect as a cue to speaker identity.

As for pitch scaling, perceived dialect of pitch transferred speakers tended to correspond to perceived identity.

Original Speaker	PITCH TRANSFERRED FROM: A	B	C
A		21/79/0	36/43/21
B	43/43/14		64/36/0
C	7/0/93	0/0/100	

Table 1. Speaker perception rates for pitch transfer experiment. For each partition the three scores represent the percentage of responses that declared A,B or C as the speaker.

Time Alignment

Time alignment of one speaker's sentence with another's did not mask or alter the perception of the speaker and dialect as being that of the original, although the changes in timing were noted by listeners.

Pitch Transfer and Time Alignment

Table 2 shows the results of perceived identity of pitch transferred and time aligned sentences. It can be seen that speaker A was often perceived as B, however pitch transfer and time alignment did little to mask the identities of speakers B and C.

A result of notice is a comparison between tables 1 and 2. It can be seen that table 2 contains a higher percentage recognition of the original speaker. This appears counter-intuitive as one would expect the extra timing information of another speaker to further mask the original speaker's identity, rather than revealing it more clearly.

As for previous alterations perceived dialect tended to correspond to the perceived identity of the speaker.

CONCLUSION

From the experiments conducted it can be seen that perceived speaker identity was affected by changes of pitch as well as pitch transfer. The two

Australian speakers were easily confused by pitch alteration. The North American speaker, however, remained relatively resistant to misidentification. We conclude therefore that dialect is a major cue for cross-dialect identification and that pitch is used as the primary cue for the discrimination of speakers of the same or similar dialects.

Original Speaker	PITCH AND TIMING TRANSFERRED FROM: A	B	C
A		36/64/0	57/43/0
B	14/79/7		7/86/7
C	0/0/100	0/7/93	

Table 2. Speaker perception rates for pitch transfer and time alignment experiment. For each partition the three scores represent the percentage of listeners that declared A,B or C as the speaker.

Time alignment by itself did little to mask the original speaker's identity. Puzzling results were achieved when pitch transfer and time alignment were performed together. This led to a higher recognition rate of the original speaker than for the corresponding pitch only transfer experiment. Further investigation as to the significance of this result is required.

ACKNOWLEDGEMENTS

Thanks to the following listeners who contributed their valuable time: L. Hogarth, J. Hogarth, D. Brennan, Y. Fong, P. Tang, L. Brown and J. O'Neill, and thanks to Willma Nelowkin for programming support.

REFERENCES

Brown, B.L., William, J.S., Rencher, A.C.,(1973)"Perceptions of personality from speech: effects of manipulations of acoustical parameters", J. Acoust. Soc. Am., Vol 54, No 1,29-35.

Takagi, T., Kuwabara, H.,(1986)"Contributions of pitch, formant frequency and bandwidth to the perception of voice personality", IEEE Int'l Conf. Acoustics, Speech, Signal Processing, Tokyo, Japan, March 1986, 889-892.

Wagner, M.,(1978)"The application of a learning technique for the identification os speaker characteristics in continuous speech", Ph. D. thesis, Australian National University, Canberra A.C.T.

Wagner, M., Fulcher, J.,(1986)"An IBM PC based speech research work station", Proceedings 1'st Aust. Conf. Speech Science and Technology.

Williams, C.E., Stevens, K.N.,(1972)"Emotions and speech: some acoustical correlates", J. Acoust. Soc. Am., Vol 2, No 4, 1238-1250.