PERCEPTIONS OF IDENTITY, GENDER AND IDIOLECT IN PROSODICALLY ALTERED SPEECH USING A COMPOSITE MODEL APPROACH

Michael Barlow & Michael Wagner*
{ spike@cs.adfa.edu.au, michael.wagner@canberra.edu.au }
School of Computer Science, University of NSW/ADFA
* School of Computing, University of Canberra

ABSTRACT – The paper describes a series of perceptual experiments in which the prosodic parameters F_0, energy, voicing and timing of utterances were systematically altered and the resulting speech resynthesised and evaluated by a group of listeners. Using the linear-prediction source-filter model of production intermediate spectral models were composed from the speech of two or more speakers. Listener perceptions of identity, gender and idiolect were correlated with the systematic alterations of the prosodic parameters. Perceptions of identity were found to be correlated with both the static and dynamic properties of all parameters, though listeners employed different cues for different speakers. Perceptions of gender were based solely on the mean value of F_0. Perception of idiolect was found only to correlate with utterance duration: shorter duration utterances were perceived as more cultivated, while longer duration was perceived as broader.

INTRODUCTION

A fundamental question in the investigation of speaker characteristics concerns the details of their encoding in acoustic parameters. Approaches to finding answers to these question fall into two broad categories: those that take an analytical approach and those that are perceptually motivated.

One novel method of the perceptual examination of speaker characteristics is the acoustic alteration of parameters: utterances are synthesised in whole or part while parameters are systematically altered. Perceptions are then correlated with the alterations performed. Such schemes in the literature have ranged from the relatively simple (Lass et. al., 1976) through to the more sophisticated manipulation of individual parameters (Van Lancker et. al. 1985; Dommelen 1990).

This paper introduces a powerful method for the parameter alteration, and resynthesis that allows tight focussing upon individual characteristics of an utterance and individual acoustic parameters. In particular the method is applied to investigate the correlates of the prosodic parameters F_0, energy, voicing, and timing with the speaker characteristics identity, gender and the idiolect of Australian English. The following sections introduce the speech material employed, before describing the composite-utterance method and the results of the perceptual experiments.

SPEECH MATERIAL

A group of nineteen (19) speakers of Australian English recorded a set of fifteen (15) sentences on five (5) separate occasions over a period of not less than one week. Recordings were made in a soundproof studio with a random sentence order on each occasion. The recordings were low-pass filtered at 7.6kHz, before being digitised at 12-bit quantisation and a sampling frequency of 16kHz.

A trained linguist scored speakers on their relative position on the spectrum of Australian English idiolects (Bernard, 1967). Each speaker received a score between 0 and 10, with low scores reflecting a cultivated idiolect, and high scores reflecting a broad idiolect.

A single sentence: "I cannot remember it." was selected on the basis of a series of analysis experiments (Barlow, 1991) as the most suitable for resynthesis and parameter alteration. Utterances were hand-segmented at phone boundaries. These boundaries were then used as anchor-points in subsequent parameter and timing alterations.

PROSODIC PARAMETERS

Four prosodic parameters: F_0, energy, voicing, and timing were extracted, and subsequently altered in the resynthesis experiments.

Voicing and fundamental frequency (F_0) values were extracted using a time-domain, parallel pitch detector (Gold and Rabiner, 1969) with a 25ms window and 10ms frame shift after low-pass filtering: 300Hz for male speakers, and 400Hz for female speakers.

Log Mean Square Energy (LMSE) values were extracted for each 25ms frame of an utterance.

Timing values were based on the segmental boundaries assigned through listening and visual analysis of the waveforms and spectra.

COMPOSITE MODEL APPROACH

Two significant problems need to be addressed when investigating the perceptual correlates of speaker characteristics in prosodic parameters. Firstly, a method is required to neutralise or normalise the contribution of segmental or spectral information in an utterance so that the influence of prosodic parameters can be isolated. Secondly, for speaker characteristics other than identity, there is a problem in creating a single 'archetype' of that characteristic free from the influences of a single speaker. For instance, how to investigate the perception of idiolect free from the influences of the identities of the speaker(s)?

A partial step towards addressing these issues is to adopt a linear predictive (Markel and Gray, 1976) source-filter model of speech production which provides a clear separation between the spectral parameters (filter) and prosodic parameters (source).

A further means of addressing these two problems is to generate a composite utterance, consisting of a number of component utterances from different speakers. Such a method 'averages' the spectral properties of the utterances while prosodic parameters may still be varied to that of one or the other of the component utterances. Further, combining a number of speakers, all with a single characteristic, such as male gender or broad idiolect, leads to an 'archetype' utterance for that characteristic. Alternatively, combining utterances from speakers of opposed characteristics, such as male and female speakers, may lead to an utterance which is neutral for that characteristic. Alterations of parameters of the neutral utterance may then be correlated with listener percepts of the opposing characteristics. Figure 1 shows a schematic of the composite model approach.

An autocorrelation linear prediction algorithm (Markel and Gray, 1976) of order p=20 was applied to each 25ms frame with a 12.5ms frame shift and the resulting reflection coefficients were averaged in order to synthesise the composite utterance.

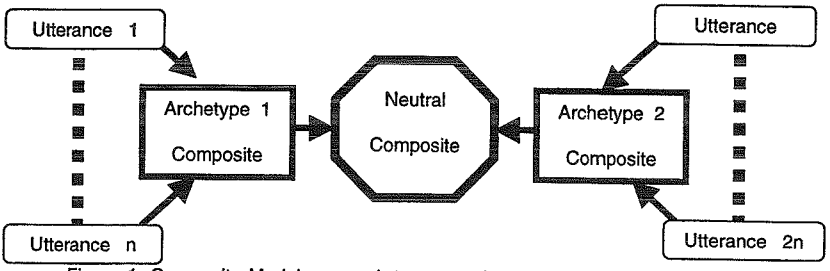

Figure 1: Composite Model approach to composing archetype and characteristic-neutral utterances.

For the identity perception experiments two male speakers, with the same idiolect score (towards the cultivated end of the scale), were selected and a composite utterance constructed. The choice of speakers of the same idiolect and gender was made to eliminate the influence of these characteristics on the listeners' perceptions of identity.

Speaker A's utterance had a duration of 0.93 seconds and a mean F_0 of 114Hz, while speaker B's utterance was 1.20 seconds in duration with a mean F_0 of 108Hz.

In order to minimise the influence of the spectral properties on the perception of identity, an exploratory experiment was conducted using a small set (6) of listeners who were not employed for the main experiment. The prosodic parameters were held fixed at the mean of the two speakers while the reflection coefficients were varied in weight from 25% to 75% between the two speakers. It was found that an equal weighting (50%) of each speaker's spectral characteristic led to an equal distribution in listener responses on the identity of the speaker. This weighting was then used for the main experiment.

For the gender perception experiments two male and two female speakers with the same idiolect score contributed to the composite utterance as shown in Fig. 1, thus reducing the influence of individual identity on the perception of gender.

For the idiolect perception experiments four male speakers contributed to the composite utterance. Two speakers were selected from the most "cultivated" end of the spectrum (with assigned scores of 1.5 and 2.5), and the other two speakers were taken from the most "broad" end of the spectrum (scores of 9 and 8.5), in order to create an approximation of an "idiolect neutral" utterance.

ACOUSTIC PARAMETER ALTERATIONS

In order to build a composite utterance it is necessary to time align utterances and parameter contours of different durations. Such an alignment was achieved by performing a piecewise linear alignment for each segment that composed the utterance. All utterances had previously been hand segmented at phone boundaries. Transformation of LP coefficients to reflection coefficients prior to alignment and subsequent weighting ensured filter stability.

The substitution of one or more prosodic parameter contours onto an extant utterance was achieved by aligning the parameter contour with the target utterance via the linear alignment and substitution between phone boundaries. The parameter contour would then simply be substituted for that associated with the utterance prior to resynthesis. Such a scheme was used for the F_0, voicing and energy contours, alone and in combination.

In a series of analysis experiments utilising the same data (Barlow & Wagner, 1998; Barlow, 1991) it was found that the relative dynamics of the prosodic parameters encoded identity, gender and idiolect. In order to investigate the perceptual significance of relative dynamics, a Dynamic Time Warping (DTW) algorithm was employed to warp or align a contour as closely as possible with another. Listener responses to the warped contour utterance were then contrasted with those for the original and target contour utterances.

Finally, the gender perception experiments involved shifting the mean F_0 of an utterance to 165Hz. This was performed by finding the current mean for the utterance and either adding a constant value to each term, or multiplying each term by a constant so as to achieve the new mean of 165Hz. No significant difference in perception results was found between the two approaches.

PERCEPTUAL EXPERIMENTS

A series of perceptual experiments were carried out to analyse the correlation between listener perceptions of speaker characteristic (identity, gender and idiolect) and the prosodic alterations performed.

A total of sixteen adult listeners, eight male and eight female, ranging in age from 20 to 38, participated in all experiments. No listener reported hearing defects and all bar two listeners were native speakers of Australian English. The two exceptions were native speakers of British English and their judgements of speaker idiolect were not incorporated into the results.

Under the control of a software script, listeners were presented with the digitised utterances through stereo headphones (Dick Smith C-4101). All experiments were forced-decision, double-blind trials conducted in a random order with each sample occurring twice during the session.

Speaker identity experiments were conducted using utterance triplets: reference utterances from speaker A and speaker B, followed by an unknown utterance. Listeners were required to designate the unknown utterance as being from speaker A or speaker B.

For the gender identification experiments each utterance was presented in isolation. Listeners were asked to identify the speaker as being female or male.

Similarly, for the idiolect perception experiments each utterance was presented in isolation. Listeners were asked to mark a point on a line reflecting their perception of the dialect or 'accent' of the speaker, with one end of the line being identified as broad or 'thick' accent and the other as cultivated.

RESULTS

Identity

As a baseline to quantify shifts in listener perceptions with altered prosodic parameters, listeners were presented with a totally 'average' utterance: 50% spectral (filter) properties of each, and 50% each of the prosodics from each speaker. Contrary to the exploratory experiments with six listeners, perception of the identity of the speaker originating the utterance was not equally divided: 34.1% of responses declared the originator as speaker A, with 65.9% as speaker B; a ratio of almost 1:2.

Table 1 shows the shift, as positive percentages towards the favoured speaker, of encoding one, two, or all of a speaker's prosodic parameters upon the composite utterance. These, and subsequent figures are normalised relative to the response of the totally neutral utterance.

Transformation		Perception	
Originating Speaker	Parameter(s)	Shift Towards A	Shift Towards B
A	F_0	8%	
	Energy	50%	
	Timing	35%	
	Voicing + F_0	17%	
	Voicing + Timing	64%	
	Energy + Timing	64%	
	F_0 + Timing	55%	
	All 4 parameters	77%	
B	F_0		41%
	Energy	8%	
	Timing		0%
	Voicing + F_0		71%
	Voicing + Timing	2%	
	Energy + Timing		32%
	F_0 + Timing		50%
	All 4 parameters		62%

Table 1: Shifts in listener perception of the originator of an utterance as 1, 2 or all 4 prosodic parameters from a speaker are encoded on a 'neutral' composite utterance.

In an attempt to quantify the contribution of parameter dynamics, as opposed to static/mean values, parameters were DTW (Dynamic Time Warping) warped to match those of the other speaker. The warped parameters were then encoded on the composite utterance. Table 2 shows the results of those experiments.

Gender

In an exploratory studying employing the same group of six listeners as that in the determination of appropriate weighting of speaker filter properties, a study was conducted of the relationship of mean F_0 to perceived gender. F_0 was varied in 5Hz steps from 130Hz to 200Hz. It was found that the point of even distribution between perceptions of male and female was 165Hz. This value was then employed for the composite model in the larger gender perception experiments: the composite F_0 contour was shifted to a mean of 165Hz.

Transformation		Perception	
Originating Speaker	Parameter(s)	Shift Towards A	Shift Towards B
A warped to B	F_0	18%	
	Energy	50%	
	Energy + Timing	27%	
	F_0 + Timing		32%
	3 + B's Timing	21%	
B warped to A	F_0	6%	
	Energy	26%	
	Energy + Timing	73%	
	F_0 + Timing	33%	
	3 + A's Timing	73%	

Table 2: Shift in listener perception of the identity of the speaker of an utterance as one or more prosodic parameters from a speaker are encoded on a 'neutral' composite utterance subsequent to the parameters being warped to match those of the other speaker.

As for the speaker identity experiments, all combinations of parameter substitution and warping were undertaken and resynthesised to listeners. Only alterations to F_0 were significant (5% level) in affecting perceptions of gender. All other parameters (energy, voicing, timing) and their combinations had no significant impact on perceived gender.

Listener responses to the neutral utterance with the mean of the composite F_0 contour at 165Hz identified the originator as male in 34.4% of cases and 65.6% as female. When a composite male F_0 was substituted perception shifted 37% (normalised) towards male. When the composite female F_0 was substituted perception shifted 73% (normalised) towards female.

Idiolect

As for identity and gender perception experiments, idiolect perception experiments were conducted in which the four prosodic parameters, in isolation and combination, were altered (substitution and warping) on the composite utterance. No significant (5% level) change to the perceived idiolect was found.

However listeners were found to be consistent in their perceptions of idiolect, matching the findings of others (Brennan et. al, 1975). Further, rate alteration of the utterance was found to significantly alter perception of idiolect as illustrated by Figure 2.

DISCUSSION

A novel analysis-resynthesis technique, known as the composite model, has been introduced and shown to be useful in analysing the correlation between listener perceptions of identity, gender and idiolect with a range of acoustic parameters. The approach of forming composite utterances has the advantage of being capable of emphasising a single characteristic of the utterance or speaker while 'normalising' other characteristics not under investigation. This, combined with the myriad of parameter alterations supported by the model facilitates fine-level investigation of the encoding of speaker characteristics.

Investigations of the encoding of identity in prosodic parameters showed that all four parameters examined encoded identity to a significant degree. Energy and segmental timing were the most significant cues (from listeners' perspective) to speaker A's identity, while F_0 was the most significant cue to speaker B's: tending to indicate that listeners utilise different cues for different speakers. Warping of the parameters, in which the static (mean) properties of the contour remained unchanged but the dynamic (time-varying) were altered to those of the other speaker showed interesting results: in all cases bar one, listener perception shifted towards speaker A, particularly any transformations involving energy (alone or in conjunction with others). Inspection of the energy contours showed significant differences between the two speakers; with A's contour being flat and B's far more variable and dynamic. The one exception to this trend: warping A's F_0 and timing to conform to B's led to a

significant increase in the identification of B as the originator, appears to indicate that listeners utilised the dynamics of B's F_0 as a significant cue.

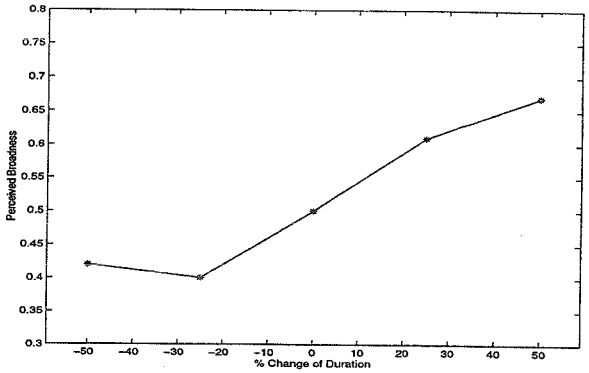

Figure 2. Listener perception of the idiolect of an utterance (vertical axis- high is broad end of spectrum, low is cultivated end) as utterance duration is altered from -50% (half the duration) to +50%.

Gender perception experiments confirmed the well-known significance of F_0 in listeners' perception of gender; no other parameters showed a significant effect.

Idiolect perception experiments showed that while naive listeners appeared consistent in their judgements of idiolect (accent) that no prosodic parameter other than rate (speed) played a significant part in the perception of idiolect; with slower utterances perceived as significantly broader and faster as significantly more cultivated.

REFERENCES

Barlow, M. (1991) "Prosodic Acoustic Correlates of Speaker Characteristics", Ph.D. thesis, University of NSW.

Barlow M. and Wagner M. (1998) "Measuring the Dynamic Encoding of Speaker Identity and Dialect in Prosodic Parameters", Proc. ICSLP-98, Vol. 2, pages 81-84, Sydney..

Bernard J. (1967) "Some Measurements of Some Sounds of Australian English", Ph.D. thesis, Sydney University.

Brennan E.M., Ryan E.B., and Dawson W.E. (1975) "Scaling of Apparent Accentedness by Magnitude Estimation and Sensory Modality Matching", J. Psycholinguistic Research, 4(1), 27-36.

Dommelen W.A.V. (1990) "Acoustic Parameters in Human Speaker Recognition", Language and Speech, 33(3), 259-272.

Gold, B., and Rabiner, L.R. (1969) "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain", JASA, 46, 442-448.

Lass N.J, Hughes K.R., Bowyer, M.D., Waters L.T., and Bourne, V.T. (1976) "Speaker Sex Identification from Voiced, Whispered and Filtered Isolated Vowels", J. Acoust. Soc. Am., 59(3), 675-678.

Markel, J.D., and Gray, A.H. (1976) "Linear Prediction of Speech", Springer-Verlag.

Van Lanker D., Kreiman J., and Emmorey K. (1985) "Familiar Voice Recognition: Patterns and Parameters part I: Recognition of Backward Voice", J. Phonetics, 13(1), 19-33.