PROSODY, FOREIGN ACCENT AND SPEECH SYNTHESIS

John Ingram
English Department
University of Queensland

ABSTRACT - The contribution of prosody to the perception of foreign accent and the impact of non-native prosody for the intelligibility of speech of second language users (Vietnamese immigrants speaking Australian English) is investigated using speech synthesis.

INTRODUCTION

Speech synthesis has long been used as a tool for systematic investigation of the signal parameters that control the perception of prosody, but much less used in the study of dysprosody and non-native patterns of prosody, though it is of potential value here also. The role of prosody in the perception of foreign accent and its influence upon the intelligibility of non-native speech is controversial and largely unknown (Flege, 1990). However, before the parametric manipulations that speech synthesis makes possible can be brought to bear on the study of foreign accent prosody, it is necessary to have benchmark information on the effects of the synthesis system itself. This was the primary purpose of the present study. These effects are of some practical interest for speech synthesis in their own right, as indicated below.

In the recent speech synthesis literature, prosody is seen as the major remaining obstacle to the goal of natural-sounding speech. While high levels of intelligibility can be achieved by modern synthesis, it is well attested that synthetic speech often conveys some impression of foreign accent, and, more often, an unnatural 'machine voice' quality. While some of this lack of naturalness and native speaker quality is undoubtedly attributable to unsatisfactory source signal modelling, a proportion is also quite possibly due to sub-optimal modelling of prosody, of the kind that non-native speakers themselves employ.

THE EXPERIMENT

Instances of errors in word and phrasal stress were culled from field recordings of the connected speech of a Vietnamese speaker of English as a second language. These tokens were paired with exemplars of the correct stress pattern produced by a phonetician. Both sets of naturalistic utterances (with foreign and native stress patterns) were then used as models for speech synthesis and a listening experiment was conducted with native English listeners.

Listeners were asked to identify the naturalistic and synthetic stimuli and to rate

them for naturalness and foreign accentedness. The design of the study permitted us to factorially separate out the contributions of the synthesis technique, and the speaker characteristics to the intelligibility, naturalness, and foreign accentedness of the tokens that comprised the utterance set.

The speakers and the speech data

The foreign accented items were drawn from a longitudinal speech data base on sound change in second language learning (Pittam & Ingram, 1990). The speaker (P.) was a 26 year old Vietnamese male, who had been resident in Australia for approximately 12 to 36 months, over the period for which speech data used in the present study was gathered. P. spoke some English before leaving Vietnam and had rapidly acquired fluency in English since his arrival in Australia, promoted by frequent contact with customers in the bakery business that he established. Although his comprehension and command of English was good, P. spoke with a strong Vietnamese accent, that seemed to be occasionally coloured with French pronunciation on certain words. French was the first foreign language that he had studied in school. In addition to the typical segmental features of Vietnamese English, P.'s speech contained numerous instances of non-English stress pattern on words and phrases.

Nine examples of stress errors on polysyllabic words (e.g.: 'marshmallow') or phrases (e.g.: 'the refugee centre') were selected from field recordings of P.'s spontaneous speech. The selection criteria were (a) that items had been independently identified by two phoneticians as containing an error or non-standard realisation of stress and (b) the target words were spoken with sufficient emphasis and deliberation to be excisable from context without significant phonetic distortion caused by accompanying coarticulation effects. The set of corresponding natural native-English exemplars were produced by a phonetician (JP), who has considerable broadcasting experience and who speaks with a modified standard British-English accent.

Synthesis method

Synthetic models of the foreign-accented and native-English tokens were fabricated using the phone catenation program provided with the LSI parallel formant synthesiser (Quarmby & Holmes, 1984). The default 'speaker table', which specifies parametric values for phone or sub-phone segments, was basically appropriate for the accent of the native-English model. The speaker table was augmented with a set of reduced-intensity vowel segments to improve the synthesis of unstressed vowels and vocalic segments in phrase final position, and to partly circumvent the limitation that the program provides control over the prosodic parameters of F0 and segment duration, but not intensity.

Spectrogram matching and extensive auditory comparisons were used to fabricate the best synthetic correspondences to the natural utterances, by a process of successive approximation until a point of diminishing returns appeared to have

been reached. Figure 1 illustrates a typical natural-synthetic token pair.

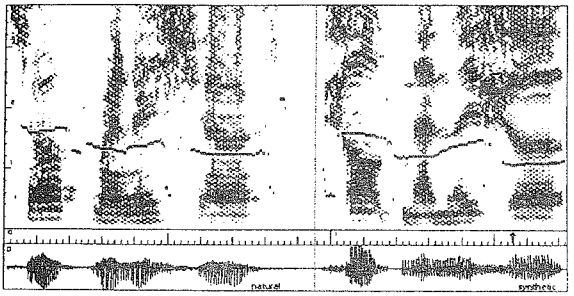

Figure 1. Natural and synthetic tokens: Vietnamese English
Phrase: 'shop manager'

The naturalness of the achieved synthesis was limited chiefly by the characteristics of the source tone and the inability to simulate 'micro-prosodic' effects caused by segmental perturbations of prosodic features. Nevertheless, the stress errors made by the Vietnamese speaker were readily recognisable in his synthetic tokens.

Segmental errors were also present in the tokens of the Vietnamese speaker and these undoubtedly contributed to the perception of 'foreign accent'. Where such errors affected phonemic perception (resulting in a perceived substitution or loss of a phoneme), the substitution was incorporated into the synthetic model of the natural token. Thus, although 'English' segmental parameters were deployed in the speaker-table used to model the natural Vietnamese-English tokens, we did not attempt to model only the non-native prosodic features of the speaker's pronunciation. To do so, would probably have compromised the naturalness of the synthesised foreign-accent tokens unduly. The primary concern of the present study was to assess the feasibility of modelling foreign accent, rather than to test the limits of the respective contributions of segmental and suprasegmental features to the perception of foreign accent.

The perception experiment

The 36 items representing natural and synthetic versions of the 9 foreign-accented and native-English tokens were presented in a perceptual experiment to small groups of listeners over a loudspeaker. Four blocks of stimuli were used to counterbalance familiarity effects. Subjects were students in an introductory linguistics course. The listeners' task was firstly to attempt to identify the stimulus, which was presented three times with an inter-stimulus interval of approximately 2 seconds, and then to rate the stimulus on scales of 'foreign accentedness' and 'naturalness', defined as shown in Figure 2.

FOREIGN ACCENTEDNESS

1	2	3	4
No foreign accent	Slight foreign accent	Moderate foreign accent	Strong foreign accent

NATURALNESS

1	2	3
Definitely human	Slightly machine-like	Definitely not human

Figure 2. Rating scales

Intelligibility scores and mean ratings for 'foreign accentedness' and 'naturalness' were calculated across subjects and used as dependent variables in a series of 3-way Analyses of Variance (ANOVAs), with Speaker (native-foreign), Utterance type (natural-synthetic) and Item (1-9) as categorical variables.

The results of the ANOVAs for the factors of Speaker and Utterance type are shown graphically in Figure 3. All main and interaction effects indicated in these graphs were significant at the $p < .0001$ level.

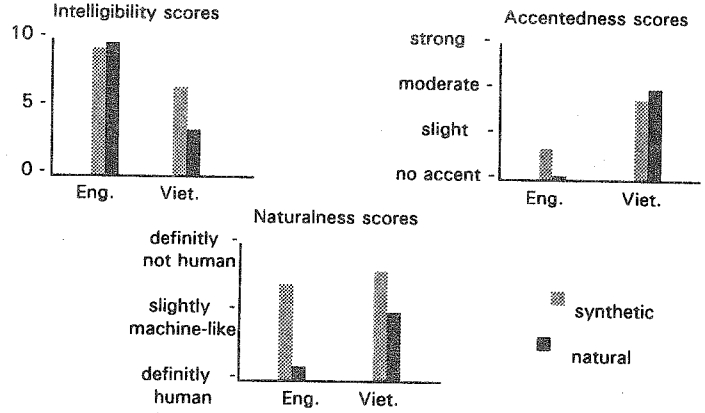

Figure 3. ANOVA results

The natural native-speaker items were correctly identified almost perfectly (97.3%) and their synthetic counterparts, not quite as well (93.3%). There was a significant interaction effect of Speaker's Language Background (SLB) and Synthesis Type (ST) upon Intelligibility scores. For the foreign-accented speaker, the natural-speech items were substantially less intelligible than their synthetic versions (37.2% identification rate vs. 64.7%). We attribute this reversal of the normally expected difference in intelligibility between natural and synthetic speech to the substitution of 'English' for 'Vietnamese-English' segmental parametric values in the synthesised tokens.

While the natural native-speaker tokens received, as expected, a mean rating of 'no accent' on the Foreign Accentedness scale (1.14), their synthetic counterparts

were perceived as 'slightly' foreign accented (1.73). This effect of the Synthesis Type upon accent ratings was also reversed for the Vietnamese-English based tokens, where the natural-speech items were perceived as more strongly foreign accented than their synthetic counterparts. Again, this significant interaction effect may be attributed to the use of 'English' segmental parameters in the synthesised Vietnamese-English tokens. Presumably, the effect of using 'English' segmental parameters outweighed the 'foreign speaker' effect of the synthesiser, which was observed with the native-speaker items.

For the Naturalness ratings, there was an expected main effect for item type, with the synthesised tokens perceived as less 'natural' than their naturalistic counterparts. However, there was also an unexpected significant main effect for the speaker's language background (SLB) upon the perceived naturalness of the speech. The Vietnamese-English speaker's productions were perceived as 'slightly machine-like', unlike those of native-English speaker, which rated, as expected, as 'definitely human'. The interaction effect of speaker's language background and synthesis type upon naturalness ratings was also significant. The discrepancy in perceived naturalness between synthetic and natural speech tokens was greater for native-English items than it was for Vietnamese-English items.

Correlations among the scales of Intelligibility, Foreign accentedness, and Naturalness are shown in Table 1.

	Intelligibility	Foreign A.	Naturalness
Intelligibility	1.000		
Foreign Accent	-0.696	1.000	
Naturalness	-0.184	0.443	1.000

Table 1. Correlations between scales.

Intelligibility showed a high negative correlation with ratings of foreign accent, but a low correlation with naturalness. There was some correlation between foreign accent and naturalness ratings.

DISCUSSION

As stated previously, the primary goal of experiment was to assess the potential suitability of the synthesis method, with its recognised limitations, as a tool for investigating the contribution of prosody to the perception of foreign accent and the impact of non-native prosody on speech intelligibility. Ideally, the effects of parametric manipulations of the speech signal should be transparent to the method of synthesis used. But in practice this is not achieved. The synthesiser contributed a small (5%) loss of intelligibility, a 'slight' but readily detectable foreign accent colouring, and an unmistakably 'unnatural' or 'machine-like' voice quality. Nevertheless, these potentially confounding effects of the synthesiser were relatively small in relation to the perceived differences between the native and foreign-accented speech samples. Also, and perhaps more crucially, synthesiser effects were relatively small in relation to the perceptual effects of manipulating

segmental and prosodic parameters of the synthesised speech.

Some tentative conclusions about the impact of non-native prosody and the relative importance of prosodic and segmental features for word and phrase recognition may be drawn from the experiment. Non-native prosody, or more particularly, errors of stress, in concert with accompanying non-native segmental features of speech yielded a low recognition rate (37.2%) for items produced by the Vietnamese-English speaker. The design of this preliminary study did not permit us to estimate the independent contributions of prosodic and segmental features, and the phonetic structure of speech itself limits the extent to which such a separation is achievable in principle.

However, it is possible to draw some preliminary inferences about the relative importance of segmental and prosodic parameters for the intelligibility of the items. The increase in the recognition rate from 37.2% for the natural Vietnamese-English items to 64.7% for the synthesised Vietnamese-English items reflects the contribution of 'English' sub-phonemic segmental parameters which were substituted for those of the original Vietnamese-English, minus the effects of loss of intelligibility, attributable to the synthesiser, which from the native-speaker data was estimated at approximately 5 percent. Preservation of non-native prosody, plus the phonemic substitutions that were incorporated into the synthesised Vietnamese-English forms, resulted in a decrement of approximately 30 percentage points in the recognition rate. Thus, the effects of non-native prosody plus phonemic substitutions appear to be roughly equivalent to the effects of non-native sub-phonemic segmental parameters, lowering the recognition rate about 30 percentage points in either case. These considerations set an upper limit on the importance of prosodic parameters in the recognition of foreign-accented, and inappropriately stressed isolated words or phrases.

The unexpected finding that natural foreign-accented speech obtained lower naturalness ratings, which fell within the range of the synthesised items may be worthy of sociolinguistic investigation. Current work, however, is focused upon the acoustic correlates that differentiate perceptions of the degree of foreign accentedness from naturalness.

REFERENCES

Pittam, J. & Ingram, J. (1990) Connected speech processes in Vietnamese Australian. Proceedings of the Third Australian International Conference on Speech Science and Technology. ASSTA: Canberra, 84 -89.

Flege, J.E. (1990) The contribution of segmental and prosodic errors to degree of perceived accent change. Unpublished MS.

Quarmby, D.J. & Holmes J.N. (1984) Implementation of a parallel formant speech synthesiser using single chip programmable signal processor. IEEE Proceedings, Vol 131, Pt F, No. 6, 563-569.