Comprehensibility and prosody ratings for pronunciation software development.

Paul Warren, Irina Elgort, David Crabbe

Victoria University of Wellington, New Zealand

E-mail addresses and telephones

Paul.Warren@vuw.ac.nz	+ 64 4 463 5631
Irina.Elgort@vuw.ac.nz	+ 64 4 463 5970
David.Crabbe@vuw.ac.nz	+ 64 4 463 5603
Fax	+ 64 4 463 5604

Mailing address

School of Linguistics and Applied Language Studies
Victoria University of Wellington
P.O. Box 600
Wellington 6012
New Zealand

Comprehensibility and prosody ratings for pronunciation software development

In the context of a project developing software for pronunciation practice and feedback for Mandarin-speaking learners of English, a key issue that has arisen is how to decide which features of pronunciation to focus on in giving feedback. The research reported in this paper uses naïve and experienced native speaker ratings of comprehensibility and nativeness as a methodology for establishing the key features affecting comprehensibility of the utterances of a group of Chinese learners of English. Native speaker raters were asked to assess the comprehensibility of recorded utterances, to pinpoint areas of difficulty and then to rate for nativeness the same utterances, but after segmental information had been filtered out. The results show that prosodic information is important for comprehensibility, and that there are no significant differences between naïve and experienced raters on either comprehensibility or nativeness judgements. This suggests that naïve judgements are a useful and accessible source of data for identifying the parameters to be used in deciding what is acceptable and what is not in setting up automated feedback.

INTRODUCTION

Most learners of English need to develop their ability in pronunciation to the point where it does not have any serious effect on comprehensibility when they are engaged in oral communication. For some, this is a skill they develop naturally to a reasonable level of accuracy through imitation of one native speaker norm or another over a period of time. Others need to work harder at it with expert guidance. Morley (1991: 492-495) provides a comprehensive overview of groups of learners in special need of pedagogical support for pronunciation for both ESL and EFL settings. One of the key features of a pedagogy of pronunciation is necessarily feedback on performance, and yet providing feedback on pronunciation is an intensive, time-consuming activity requiring one-to-one work. In addition, Derwing and Munro (2005: 382) lament the "marginalization of pronunciation within applied linguistics". Their informal survey shows either a complete omission of pronunciation from some key publications, such as The Handbook of Second Language Acquisition (Doughty & Long, 2003), or only minimal attention to the topic (as in Hedge, 2000; Nunan, 1999). They also found only few papers on pronunciation published in applied linguistic journals. Research conducted in Canada, Britain and Australia also shows that in addition to the lack of training in pronunciation instruction, English teachers, in general, do not have a strong enough background in phonetics and phonology to feel confident to teach pronunciation (Breitkreutz, Derwing, & Rossiter, 2002; Burgess & Spencer, 2000; MacDonald, 2002; Murphy, 1997). It is not surprising, then, that pronunciation is often given little attention in the classroom, particularly in the communicative curriculum where a focus on meaning has for many years dominated over a focus on form, including phonetic form.

For this reason, the computer-assisted language learning approach appears to be promising, as it can enable students to work on improving their pronunciation independently, outside of the language classroom, focusing on aspects of pronunciation relevant to their individual needs, based on their L1 background and their language learning goals (Pennington, 1999). Unfortunately, it appears that much of the commercially-available pronunciation software does not meet the criterion of being “linguistically and pedagogically sound” (Derwing & Munro, 2005: 391; see also Neri, Cucchiarini, Strik, & Boves, 2002). A key requirement for effective CAP (computer-assisted pronunciation – cf. Pennington, 1999) software is that it provides “immediate, useful feedback, especially for those features that are most important for intelligibility” (Levis, 2007; Neri, Cucchiarini, Strik, & Boves, 2002).

The research results reported in the current paper form part of a larger project, the goal of which was to explore ways of providing automated feedback on learners’ attempts at pronunciation, in the context of pronunciation development as a component of conversational fluency (Pennington & Richards, 1986). This is not of course a new undertaking. Many forms of automated feedback on pronunciation are now appearing, based on a comparison of the acoustic analysis of the learner’s utterance with the target norm stored in the system (for reviews of the issues, see Ehsani & Knodt, 1998; Levis, 2007; Neri, Cucchiarini, Strik, & Boves, 2002; Pennington, 1999). More time will pass before any of these programs are developed to the point where all of the automated responses on performance are useful in guiding the learner towards improving that performance, but this is a productive field in which gradual progress is being made (see, for example, Connected Speech by Protea Textware[1] or ISLE software produced by the European ISLE Consortium[2]).

Our larger project involved the efforts of researchers from three distinct areas of expertise: phonology, computer science and second language pedagogy. The aim of the overall project was to develop pronunciation software to provide automated feedback to learners, software that would be informed by linguistic understanding, particularly of comprehensibility parameters, and by pedagogical understanding of how people can be supported in developing pronunciation skills.

ACCENTEDNESS AND NATIVENESS, INTELLIGIBILITY AND COMPREHENSIBILITY

Levis (2007: 187) identifies “two overlapping and conflicting” principles in pronunciation research and pedagogy (see also Levis, 2005): the nativeness principle and the intelligibility principle. One characterisation of the difference between nativeness and intelligibility is that the former refers to “how strong the talker’s accent is perceived to be” (Munro & Derwing, 1995: 291), or “how different a speaker’s accent is from that of the L1 community” (Derwing & Munro, 2005: 385), while intelligibility is commonly used to refer to the extent to which an utterance is “actually understood” by a listener. Although the nativeness principle continues to be reflected in English teaching curricula and in research concerned with the relationship between foreign accents and identity, the principle of intelligibility has come to the fore in the context of communicative language teaching approaches.

A commonly-used alternative label to “nativeness” is “accentedness”. To an extent these two terms can be seen as reciprocal opposites (with the focus on how much or how little like a native speaker the learner’s pronunciation is, respectively), although it is interesting to note that Derwing and Munro (1997: 6) use both terms for one of their tasks with a response continuum that ranges from “perfectly nativelike” at one end to “extremely accented” at the other. For the current study we have chosen to use the label “nativeness” rather than “accentedness”. Our choice of this label is primarily because the task we used to assess the role of prosody in listener ratings involves low-pass filtering of the speech, in order to focus judgements on prosodic features of the utterances. This results in the loss of the vowel quality and consonantal features that make a major contribution to what is perceived as an accent in a language. (Our use of “nativeness” rather than “accentedness” conveniently also allows us to avoid possible terminological confusion, since “accent” is also used in prosodic analysis to describe one type of prominence.)

Two further terms that need to be carefully disintinguished are “intelligibility” and “comprehensibility”. The former is typically an objective measure, commonly assessed through transcription tasks, while comprehensibility is more usually measured using human rater judgements (Derwing & Munro, 1997; Derwing, Munro, & Carbonaro, 2000; Munro & Derwing, 1995, 1999, 2001). Comprehensibility typically refers to a listener’s *perception* of the amount of *effort* involved in understanding a particular non-native speaker (NNS). The two measures (intelligibility and comprehensibility) appear to be well correlated (e.g., Munro & Derwing, 1999), which suggests that the amount of effort associated with understanding a particular NNS by a native-speaker (NS) listener is likely to be indicative of this listener’s ability to correctly process NNS utterances. In the current study we will be concerned with comprehensibility ratings of utterances, i.e. a measure of the effort required by our raters to understand the utterances they are asked to listen to.

PROSODY, COMPREHENSIBILITY AND NATIVENESS

For reasons outlined below, it was decided early in our project that the main focus would be on prosodic features of pronunciation (including temporal features such as rate and rhythm, as well as loudness, intonation, and vowel quality) and that we would focus on Mandarin-speaking learners of English (MSLEs). This was firstly because they are the largest group of English language learners and secondly because of the considerable linguistic differences between Mandarin and English (particularly in prosody, see Pennington & Ellis, 2000), and the existing evidence that first language transfer is one of the important factors affecting second language development in the area of phonology (for overviews of prosodic transfer see also Hansen, 2001; Pennington & Richards, 1986).

For the initial prototype software module we chose to focus on stress patterns across the utterances, as a key feature that affects comprehensibility (see below). Recognition trials using this prototype (Xie, Andreae, Zhang, & Warren, 2004a, 2004b; Xie, Zhang, & Andreae, 2006) resulted in stress recognition rates – in native-speaker English – of up to 92.6%, using a combination of features based on vowel duration, amplitude, pitch and vowel quality. The acoustic correlates of the prosodic features, especially duration and amplitude, proved to be the most useful to the model. Of the vowel quality features, those that were associated with reduced (therefore unstressed) vowels were most useful.

Although these results are comparable or even slightly better than those produced by similar systems (see discussion in Xie at al., 2004b), they still do not provide an adequate basis for giving feedback to language learners. One way in which these results might be improved is by fine-tuning of the parameters and normalisation methods used. Another is to allow the software development to be better informed by native speaker judgements of non-native speech. This latter approach resonates with a point made in Kim (2006) that feedback provided to language learners by CAP software should be consistent with human feedback (Cucchiarini, Strik, & Boves, 2000a; Derwing, Munro, & Carbonaro, 2000; see also Levis, 2007). In particular, it is unclear whether a set of outputs from the software, in terms of how non-native rhythm and stress deviate from stored native speaker norms, are truly indicative of the types of problems associated with prosodic features of speech, as perceived by native speakers. Thus, it is critical to establish which prosodic features affect the native speaker listener judgements of comprehensibility and nativeness, in order to evaluate the measures used by the software to compare the prosody of the learner's utterance with that of a stored sample. This issue is an important one for speech recognition software if the analysis is to be of any practical use as a pedagogical tool. For this reason we set out to gather data on native speaker perceptions of MSLE utterances with the ultimate purpose of using them to evaluate and fine-tune the computer analysis. The rest of this paper reports on the procedures we used to gather this data, the results and the possible implications for the use of the data.

A number of factors motivated the focus on prosodic features in our research. First, we expect Mandarin English pronunciation to be particularly affected by important differences between these two languages in key aspects of prosodic structure (Pennington & Ellis, 2000). These include the lexical use of tone in Mandarin but not in English; differences in basic rhythmic structures, with Mandarin and English nearer the syllable-timed and stress-timed ends respectively of a continuum of rhythm types (Adams, 1979; Grabe, 2002); and the greater use in Mandarin of tonal range to indicate stress differences (Kratochvil, 1998; Shen, 1990). As an example of the effects of such differences on Chinese English, Chao (1980) showed that through an association of stress with pitch, Chinese learners of English produce phrases with a pitch pattern determined by the stress patterns of the separate words, rather than using an intonation pattern more appropriate to the phrase as a whole. Thus *apple cider* might be pronounced by a Chinese learner of English with a high-low-high-low tonal pattern.

This would give the phrase a double-stress, contrasting with a native speaker's likely pattern of high-high (or mid-high)-high-low, which has a single marked pitch fall from a peak on the first syllable of *cider*, marking this as the stressed word. Similarly, Juffs (1990) found that the most frequent stress errors in Chinese English result from using a tonic stress movement to mark lexical stress. Segmental differences are also important through the impact they have on prosodic structure: for instance Juffs commented that syllable structure influences stress errors in Mandarin English not just because syllables are the domain of lexical stress in English but of tone in Chinese, but also because the syllable structures of the languages differ and syllable structure is crucial to the assignment of stress (see also Ramus, Nespor, & Mehler, 1999). Tajima, Port and Dalby (1997: 18-20) observed many segmental errors in Mandarin English that affect syllable shape, many of these reflecting a tendency to avoid consonant clusters by either deleting consonants or inserting epenthetic vowels (see also Hansen, 2001; Lin, 2001; Weinberger, 1997), which clearly also affects the rhythmic pattern of the utterance. Tajima et al. also noted a reduced difference between stressed and unstressed vowel durations for non-native speakers.

A second reason for our focus on prosodic features is that the effect of the prosodic features of speech on intelligibility has been acknowledged both in longstanding teacher beliefs and, more recently, in pronunciation instruction research (Derwing & Rossiter, 2003; Derwing, Munro, & Wiebe, 1998; Hahn, 2004). This has led to an increased recognition of the role of prosody in both the intelligibility and the comprehensibility of both native and non-native speech (Anderson-Hsieh, Johnson, & Koehler, 1992; Munro & Derwing, 1999). A range of L2 studies have shown that prosodic factors can effect both intelligibility/comprehensibility and nativeness/accentedness, and often with more extreme results than segmental factors. Thus Tiffen (1992) found that the intelligibility of English from Nigerian speakers was found to be more adversely affected by rhythmic and stress factors than by segmental, phonotactic and lexical/syntactic errors. Benrabah (1997) cited data from Indian, Nigerian and Algerian non-native speakers of English whose stress patterns – with 'close enough' segmental qualities – result in misidentifications of words when listened to by native speakers. Indeed, inappropriate timing and patterns of stress alternation are often cited as major contributors to intelligibility deficit (Adams, 1979; Hahn, 1999; Hahn, 2004; Kenworthy, 1987; Nelson, 1982). Anderson-Hsieh et al. (1992) used multiple regression analysis to show that three attributes, namely segmentals, syllable structure and prosody (including stress, rhythm, phrasing and intonation) all play a role in pronunciation ratings, but they found that the prosodic variable had the strongest coefficient. Japanese learners of English have also been shown to use intonational and rhythmic patterns that are judged as “unnatural” (Ono, 1991), and Hutchinson (1973) reported that Spanish speakers who maintain a greater durational difference between stressed and unstressed syllables in their English were given better pronunciation ratings.

Looking specifically at Mandarin learners of English, the target learner group in the current study, we find again a strong effect of prosodic factors. Munro and Derwing (1999) noted that intonation is important in native speaker ratings of comprehensibility and accentedness, and reported in addition that while accentedness correlates with perceived comprehensibility and intelligibility, a strong accent does not necessarily result in reduced comprehensibility or intelligibility. Elsewhere (Derwing & Munro, 1997: 4), these authors indicated that "a strong foreign accent does not necessarily interfere with intelligibility, although NSs may require extra processing time to understand NNS speech, which may lead to lower perceived comprehensibility ratings", as shown in an earlier study (Munro & Derwing, 1995). Rhythmic factors were highlighted by Tajima et al. (1997) who used LPC resynthesis and dynamic time warping to align Mandarin English with native English timing patterns, and found a significant increase in intelligibility from 39% to 58%. Note that Tajima et al.’s alignment procedures (1997: 8-9) also involved so-called “discrete” changes, i.e. removing or inserting segments that were or were not in the original Mandarin, so as to match the English target. In other words, this was not just a temporal re-alignment but involved at least some correction of inappropriate segments that would affect the overall prosodic structure. Tajima et al. (1997) concluded that “there is good reason to believe that non-native speakers would benefit from training programs which focus on various temporal aspects of their speech” (1997: 21).

Finally, there are further pedagogical reasons for a focus on prosodic aspects of non-native speech. For instance, one study of the influence of age, motivation and instruction on phonological performance (Moyer, 1999) varied the type of phonological feedback given to learners, to include either feedback on segmental aspects alone, or feedback on both segmental and suprasegmental aspects of learners’ performance. The study found that type of phonological feedback and learning outcomes were significantly related, and that “subjects who were given both suprasegmental and segmental feedback scored closer to native in a predictably constant relationship”

(Moyer, 1999: 95). In another study (Derwing, Munro, & Wiebe, 1998) three instruction types were used. Two were based on pronunciation, i.e. segmental and global (the latter including stress, intonation and rhythm), and one had no specific pronunciation instruction (providing a control group). Learner utterances and narratives were recorded at the beginning and end of a 12-week course of instruction. The individual utterances were rated by non-expert native speaker raters for comprehensibility and accentedness, while narratives were rated for comprehensibility, accentedness and fluency. In their second experiment, which rated the narrative data, Derwing et al. (1998) found that only “speakers who had had instruction emphasizing prosodic features such as rhythm, intonation, and stress could apparently transfer their learning to a spontaneous production” (1998: 406). Hardison (2004) also found that computer-assisted prosody training using real-time pitch display produced significant improvement in both prosody and segmental accuracy with generalization to novel sentences, as judged by native speaker raters. In a similar vein, Hirata (2004) showed that computer-assisted training for English speakers learning Japanese which included visual feedback based on the fundamental frequency contours improved these learners' ability to both perceive and produce pitch and duration contrasts in Japanese.

A RATING STUDY OF THE COMPREHENSIBILITY AND NATIVENESS OF MSLE SPEECH

Since the overall goal of our larger project was to develop interactive software for pronunciation training with a focus on prosodic aspects of learner speech, we conducted a series of tasks that aimed to establish the links between comprehensibility, nativeness and the segmental and prosodic features of non-native speech. It is the results of these tasks that are presented in this paper.

Comprehensibility and nativeness ratings were collected from both experienced and naïve raters, as described below. In addition, the experienced listeners were asked to identify specific areas of difficulty in the utterances they heard. These areas included both prosodic features such as lexical and sentence stress, rhythm and pitch, and segmental features such as consonant and vowel articulation. We chose to ask experienced listeners because they are more likely than naïve listeners to be able to pinpoint perceived problem areas.

Our ratings of nativeness also focused on prosodic aspects of the isolated sentence utterances. This was achieved by using low-pass filtered speech (see also Derwing & Munro, 1997; Munro, 1995; Trofimovich & Baker, 2006; Van Els & De Bot, 1987). Speech filtered in this way removes detailed information concerning the consonant and vowel segments in the speech and causes listeners to focus on prosodic features such as the timing features of duration, rate and rhythm, as well as amplitude and intonation. The resulting speech is incomprehensible, since it is deprived of any interpretable segmental and lexical content. Participants' judgements of nativeness are therefore based solely on the prosodic features that are preserved under such conditions (Derwing & Munro, 1997). While it can be argued that the intonation pattern of the resulting speech is severely de-contextualised, since for instance listeners cannot know whether pitch accents are being placed on the appropriate words or syllables for the intended meaning of the utterance, we believe that the low-pass filtered speech conveys sufficient non-segmental information for our judges to be able to assess the nativeness of at least the more general prosodic aspects of the utterances. The results we present below seem to bear this out.

Another important aspect of our rating studies is that they included both experienced and naïve judgements of the same utterances. This allows us not only to compare comprehensibility and nativeness judgements from the same groups of listeners but also to evaluate ratings from experienced and naïve listeners in comparable conditions. This is of methodological importance, since it provides some evidence for the relative merits of using trained and experienced vs. naïve listeners for such judgements. For instance, previous research (Thompson, 1991) has indicated higher reliability in accentedness judgments from experienced raters than from naïve raters.

Speech Material

The source materials for the rating study were recordings of five Mandarin Speaking Learners of English (MSLEs) enrolled in 12-week English language courses at the School of Linguistics and Applied Language Studies at Victoria University of Wellington. Only female speaker recordings were used in this study, in order to simplify the speech analysis parameters being used in the computational component of the project. The ages of these five speakers at the time of recording ranged from 21 to 27, and their language proficiency scores were at a level sufficient for entering into university undergraduate study programmes (their local test scores were in each case equivalent to at least IELTS 6.0). They had been in New Zealand for at least 10 weeks, and all went on to enter degree programmes at Victoria University of Wellington.

The materials were based on a set of phonologically-rich isolated sentences used in the New Zealand Spoken English Database (NZSED: Warren, 2002). Pre-selected sentences were chosen for the study rather than excerpts from free narratives (Derwing & Munro, 1997; Derwing, Munro, & Wiebe, 1998), as this methodology is generally used in similar studies that compare listener rating with automatic speech recognition and evaluation (Cucchiarini, Strik, & Boves, 1997, 2000a, 2000b). Using sentence materials based on those in NZSED also meant that we had access to a large set of comparison materials from native speakers, which was exploited in developing materials for the nativeness rating task.

Using the Range program (Nation & Heatley, 2001) we selected from the entire set of NZSED sentences a subset that excluded low frequency words. This reduced the likelihood of word mispronunciation by non-native speakers due to their unfamiliarity with the words. This subset was further scrutinized by an experienced English language teacher to assess the likely familiarity of our target learner population with their lexical content. The resulting recording set of 100 sentences was then read aloud (after quiet reading for familiarisation) by five female MSLEs. From these recordings a final set of fifty utterances (ten from each of five speakers) was chosen so as to optimize the range of segmental and prosodic features of MSLE speech. The sentences in this final set had an average word length of 11.3 words (range 7-15). The sentences were long enough for rhythm and rate characteristics of the speaker to emerge. Examples are given in (1) and (2) below. This final set also excluded any items containing hesitations, repeats or restarts resulting from difficulties in reading the utterance texts.

(1) The price range is smaller than any of us expected

(2) The world is becoming increasingly dangerous but hardly anyone cares

In addition to the non-native speaker recordings, a further fifty utterances recorded by age-matched female native speakers as part of the NZSED project (Warren, 2002) were included in the speech material for the nativeness rating task. Again, this set consisted of 10 sentences from each of five speakers. These fifty utterances were different sentences from the materials selected from the MSLEs, and had an average length of 11.9 words.

For the nativeness rating task, both native and non-native speech materials were subjected to low-pass filtering (with a cut-off frequency of 350 Hz), in order to remove most of the segmental information from the signal, while leaving prosodic features largely intact (see also Derwing & Munro, 1997; Trofimovich & Baker, 2006).

Raters

Ten naïve and six experienced raters were used in the study. The naïve group consisted of staff and students of Victoria University of Wellington whose area of expertise and/or study was not related to language or linguistics. This panel also had no regular contact with Mandarin speakers of English or any other non-native speakers of English. The experienced group consisted of teachers, all native speakers of New Zealand English, teaching on the English Proficiency Programme at Victoria University of Wellington. As is the case with many English language teachers, they had little phonetic training and minimal expert knowledge of intonation and prosody. They had minimal knowledge of Mandarin or other Chinese languages, but had considerable experience of working with Mandarin learners of English, who at the time of the study made up a sizeable proportion of the students on the English Proficiency Programme. In the majority of studies which involve native speaker ratings of L2 pronunciation, either only experts or experienced raters (Anderson-Hsieh, Johnson, & Koehler, 1992; Cucchiarini, Strik, & Boves, 1997, 2000a, 2000b) or only naïve raters (Derwing & Munro, 1997; Munro & Derwing, 1995) are generally used. Studies that use experts sometimes include raters from different expert backgrounds (Cucchiarini, Strik, & Boves, 2000a, 2000b), e.g., phoneticians and speech therapists, to make a comparison and evaluate reliability of expert ratings produced by different groups. However, to our knowledge there is only one study (Thompson, 1991) that compares the ratings of experienced raters with naïve raters. This is a significant issue both because expert or experienced raters are generally harder to recruit, and because some studies show disparity between the judgements of expert and naïve raters, and between those of experienced and inexperienced raters[3].

Procedure

The study consisted of two separate sessions. The sessions differed slightly for the experienced and naïve groups of raters. In their first session (comprehensibility rating), naïve listeners completed three tasks for each utterance, as follows:

i) First, listeners were asked to rate the comprehensibility of the recorded utterances. The following clarification was provided to ensure that all raters had the same understanding of the meaning of the term *comprehensibility*: "In carrying out this rating, please think about how much effort you had to put into working out what was being said." Raters listened to each utterance once, without seeing a transcription of the utterance, before giving a comprehensibility rating on a scale from 1 ("not easy to understand") to 9 ("very easy to understand"). Our use of a 9-point scale is based on that of Derwing and Munro (1997: 6) except that their scale ranged from "extremely easy to understand" to "extremely difficult or impossible to understand".

ii) Raters were then presented with the orthographic transcription in a response booklet and were asked to mark (using underlining or circling) specific areas of difficulty that affected comprehensibility.

iii) Finally, raters were asked to comment in the response booklet on general areas of difficulty affecting comprehensibility across the utterance as a whole.

For tasks ii) and iii), raters were able to listen to the utterance as many times as they needed.

The experienced listeners followed the same procedure as above, except that between tasks i) and ii) they carried out an additional task, as follows:

> These experienced raters heard the utterance one more time, still without seeing the orthographic transcription. Their instruction screen for this part of the study read "Thinking about the utterance as a whole, indicate on the next page of your response booklet whether any of the following areas caused particular difficulty for understanding" after which they were given a list of phonetic and prosodic features to choose from, namely *pronunciation of consonants, pronunciation of vowels, word stress, sentence stress*, *rhythm*, *intonation* and *rate*, i.e. aspects of pronunciation at both segmental and suprasegmental levels that have previously been associated with listener effort in understanding. Our intention was that using these categories would provide us with some structured information about the types of difficulty

experienced by the raters. However, the raters were also able to add other areas of difficulty, in their own words.

This additional task was included in order to obtain more precise data from experienced listeners on aspects of pronunciation and prosody that might affect comprehensibility judgements, for use in our further analysis. We believed that naïve listeners would not have been able to provide data of this type in a readily interpretable form, because of unfamiliarity with the appropriate linguistic terminology. This particular design aspect of our study, which required experienced raters to identify features that caused difficulty, differs from previous comprehensibility studies, where listeners either only rate overall comprehensibility, or are also required to assign specific ratings for identified features, rather than actually identifying features that cause difficulty in comprehension, as such.

In the second session, raters were asked to provide ratings of nativeness ("Enter your rating of how much this was like a native-speaker") for each of the 100 examples of low-pass filtered speech (fifty NS utterances along with the fifty NNS utterances used in the comprehensibility task, presented in random order). Listeners heard each utterance twice, and assigned a rating from 1 ("not at all native-like") to 9 ("very like a native speaker"). Derwing and Munro (1997: 5) similarly used a 9-point scale in their accentedness task, though the endpoints are reversed, ranging from "no accent" to "extremely strong accent". As well as removing segmental cues to lexical content, the low-pass filtering also eliminated voice quality information conveyed by segmental properties (e.g. by vowel quality), and we believe that it is reasonable to assume that this, along with the random mix of the NNS items with previously unheard NS items, made it unlikely that listeners would have been able to base their judgements of nativeness on remembered aspects of the NNS utterances they had heard in the separate comprehensibility rating session. In addition, our nativeness rating task, unlike that used by Derwing and Munro (1997), did not present raters with transcripts of the sentences to refer to while assigning nativeness ratings, ensuring that their 'feel for' nativeness – based on the available prosodic information – was the only reference point in this judgement.[4]

Presentation of speech stimuli and collection of rating data were controlled by *E-Prime* software (Schneider, Eschman, & Zuccolotto, 2002). Raters entered data directly onto response sheets for the more qualitative aspects of the first session. Two presentation orders of the utterances in the first session were used; utterances were placed into two blocks, and the presentation orders differed in how these two blocks were ordered. Within each group of raters (experienced and naïve) half of the participants were randomly allocated to each order, to reduce any impact of practice effects on judgements for individual utterances, particularly effects that might result from increasing familiarity with MSLE pronunciation. For the nativeness rating session, a new random presentation order of utterances was determined for each rater by the software.

RESULTS

This section presents summaries of the results from the two tasks, for both experienced and naïve listeners, as well as comparisons of results for the two rater groups and comparisons of the results for the two tasks. Detailed discussion of the results follows in the next section.

Reliability

Preliminary analyses of our results are necessary to determine that we have good inter-rater agreement, giving us confidence that our rating data will be of use in testing the software. Our first question therefore is whether our rating tasks provide good levels of inter-rater reliability, at least comparable with those reported in the relevant literature. We assessed inter-rater reliability for each rating task by two methods. First we transformed correlations between each pair of raters into Z-scores and calculated the mean (Hatch & Lazaraton, 1991: 533). Second, to allow comparison with other published research using the same method, we calculated Intraclass Correlations (Shrout & Fleiss, 1979). For the comprehensibility rating task, we obtained for the entire group of 16 raters (10 naïve, 6 experienced) a Pearson coefficient (*r*) of .75, significant at $p<0.01$, and an Intraclass Correlation Coefficient (ICC) of .954, $p<0.01$. The equivalent analysis for the nativeness rating data for the entire group of raters over the complete set of 100 utterances (50 native speaker and 50 MSLE) gave a Pearson coefficient (*r*) of .74, significant at $p<0.01$ and an ICC of .931, $p<0.01$. For the native speaker utterances alone the analysis of nativeness ratings gave an *r* of .68 and ICC of .824; for the non-native speaker utterances *r* was .74 and ICC was .937 (all significant at $p<0.01$). The lower figures for native speakers most likely result from a more restricted range of rating values given for these speakers, giving less scope for a clear correlation effect.

However, statistical comparison of the ICC figures showed no significant difference between the reliability scores for ratings of native and non-native speakers. Note that our overall reliability scores compare well with those reported in the literature, e.g. r of .71 and .70 for comprehensibility and accent ratings respectively for the naïve raters reported in Derwing et al. (1998: 402) and ICC of .968 and .987 for comprehensibility and accent ratings reported by Munro and Derwing (2006).

In addition to the overall reliability analysis, we considered the reliability of experienced and naïve rater groups separately. This is because we wish to assess our comprehensibility ratings against identification of problem areas by the experienced listeners. We therefore need to be confident that the comprehensibility ratings given by the group of experienced listeners alone are reliable. Munro and Derwing (1995: 297), for example, point out that according to some previous research (for example, Gass & Varonis, 1984) comprehensibility rating "tends to improve with increased exposure to foreign-accented speech", which is likely to be the case with English language teachers in New Zealand, who have high exposure to MSLE pronunciation. Thompson's (1991) experiment, in which experienced and inexperienced raters evaluated the degree of foreign accent using speech samples from Russian-born NNS of English, also showed that experienced raters (college-educated native speakers who spoke a foreign language fluently, lived and studied abroad, had taken a course in linguistics and had frequent contacts with Russian speakers of English) were significantly more lenient towards deviations in L2 pronunciation as a group than the inexperienced NS raters. However, experienced raters' judgements were more reliable and did not fluctuate as much, compared to inexperienced raters (Thompson, 1991: 195). In addition, we wished to compare ratings from naïve and experienced listener ratings (English language teachers, in our case) in order to determine whether experienced ratings are in agreement with those given by the naïve listeners, and to improve the ecological validity of the study.

Our second analysis therefore tests whether each group of raters shows a good level of reliability, and whether there are measurable differences in the comprehensibility ratings given by experienced and naïve listeners. The Pearson coefficients within each group of raters were .72 and .74 for the 6 experienced and 10 naïve listeners respectively, and the corresponding ICC values were .883 and .929. All values were significant at $p<0.01$, and values for the two rater groups did not differ significantly from one another, indicating good and comparable levels of agreement within each of the groups. Mean ratings within each group were calculated for each of the 50 utterances. The overall means were 5.97 and 6.02 for experienced and naïve rater groups respectively (on the 9-point scale), and a matched-pairs t-test indicated that these did not differ (t=0.528, df:49, p=0.60). In addition, a correlation analysis of the utterance means for each group showed a high level of agreement between experienced and naïve listeners (*r* = .92, p<0.001).

Comprehensibility and areas of difficulty

The above analyses have confirmed that both tasks show good overall levels of inter-rater reliability, and that the level of agreement between the two rater groups in the comprehensibility task is high. These results give us confidence that we can generalize to naïve listeners any association that we may find between the comprehensibility ratings and the indications of areas of difficulty given by the experienced listeners. In the context of the overall project goals and our focus on prosodic features, our next analysis therefore addresses the question of whether the comprehensibility ratings given by experts are reliably associated with these same experts' indications of difficulty in areas related to prosodic structure, namely intonation, rhythm, stress, rate. (It should be noted of course that a positive answer to this question does not necessitate a negative answer to a similar question that might be posed concerning the role of segmental features, that is, it is possible that features in each area are closely associated with comprehensibility.)

To determine whether comprehensibility ratings were associated with specific areas of difficulty identified in the utterances, a *logit* model (Agresti & Liu, 2001; Liang & Zeger, 1986) was applied to the experts' rating data and the seven problem areas open for identification by them on their second hearing of the utterance (recall that this is still prior to seeing the orthographic transcription of the utterance). This analysis revealed a significant association of comprehensibility ratings with identifications of problems in each of the following areas: sentence stress, consonant pronunciation, vowel pronunciation, and intonation (each at p<0.01), as well as rhythm and word stress (each at p<0.05), with the strength of the association with these six factors decreasing in the order given. The association in each case was that a lower rating was more likely to be associated with an indication of a problem in each of the six areas for which the association was significant. Unlike other authors (Munro & Derwing, 2001), we found that problems in speech rate showed no significant association with the comprehensibility rating. Pennington (1992) comments on the relationship between rate and phonological proficiency in the interview data she obtained from non-native speakers, pointing out that slower speech affords fewer opportunities for the ellisions and connected speech processes that are a feature of native-like speech. It is possible in the case of our read materials, which tend to be more carefully produced, that there were fewer instances of connected speech processes in both non-native and native speech, so that the potential effect of rate was reduced.

Factor analysis of the seven problem areas reduced them to five components. The first of these included significant loadings for sentence stress, intonation and rhythm, which we can call a sentence prosody factor. The other components loaded individually for each of the remaining four areas: consonant pronunciation, vowel pronunciation, word stress and rate. Subsequent analysis showed significant correlation of comprehensibility ratings with each of sentence prosody, word stress, consonant pronunciation and vowel pronunciation (with *r* in the range .24-.31).

Nativeness

The next set of analyses relates to the nativeness ratings. These were obtained, as indicated above, in order to force listeners to focus on the prosodic features of the utterances. The reliability statistics reported above have shown that overall inter-rater reliability is good for this task (*r* was .75, ICC was .954, p<0.01). However, more detailed analysis shows reliability to be numerically greater for the naïve listeners than for the experts, with *r* at .69 and .73 and ICC at .822 and .904 for the 6 experienced and 10 naïve listeners respectively (significant at p<0.01). (Note that the similar analysis of the comprehensibility ratings showed a smaller difference between the two rater groups.) In addition, naïve listeners show a greater distinction between native and non-native speakers (mean ratings for each group were 6.11 and 3.90 respectively) than the experienced listeners (5.83 vs. 4.43). However, this difference was not confirmed in Analysis of Variance of each rater's mean ratings for each speaker group. This analysis showed a significant main effect of speaker group (F[1,14]=50.65, p<0.001)[5], but no interaction of speaker group with rater group (F[1,14]=2.55, p>0.1). Since the following analysis of comprehensibility is based on data only from our naïve listeners (recall that our experts were not asked to complete this part of the test), it is reassuring that the results presented in this section fail to show any significant differences between ratings from the experienced raters and those from the naïve raters.

Comprehensibility and nativeness

In the identification of materials that can be used to assess the software, our goal was to isolate utterances that present difficulties on the basis of their prosodic features. The analysis of comprehensibility and the identification of problem areas has gone some way towards achieving this goal. The analysis of nativeness ratings also makes a contribution in this direction, in that we could select items simply on the basis of low scores in this task. However, we are also interested in the relationship between comprehensibility and perceived nativeness, and in particular in whether there is any association between these two. The presence of a positive relationship might suggest that the prosodic features not eliminated by the low-pass filtering are indeed contributing to comprehensibility. So our next question is whether the nativeness ratings (of low-pass filtered speech) and comprehensibility ratings (of unfiltered speech) from our naïve listeners are correlated, as might be predicted by a model of comprehensibility that acknowledges the contribution made by the prosodic features being assessed in the nativeness rating task. Since the same MSLE utterances were used in each rating task, this question can be addressed in a simple correlation analysis of average comprehensibility and nativeness rating scores given to each MSLE utterance. In Figure 1 these rating scores for each utterance in the two tasks are plotted against each other. As well as the significant overall correlation (*r*=.59 p<0.001), it is interesting to note that the data are distributed in a manner that indicates that perceived nativeness, as measured in this study, provides as it were a baseline on top of which comprehensibility appears to be built. That is, comprehensibility ratings most usually exceed nativeness ratings for individual utterances, and are rarely lower than the nativeness ratings. Indeed, our results here mirror those of Derwing and Munro (1997: 11), who observe that “accent ratings are harsher than perceived comprehensibility ratings”. This is clear from clustering of data-points in the top left quadrant of Figure 1, and confirms that prosody is not the only factor that affects comprehensibility, which depends on both segmental and prosodic aspects of speech.

Insert Figure 1 about here

DISCUSSION

The preceding section has presented the main results from our rating study. These show that inter-rater reliability in the rating tasks is good, and that experienced and naïve raters show a high degree of agreement in the comprehensibility rating task, but less so in the nativeness task. In addition, comprehensibility ratings are significantly associated with experienced listeners' identification of problems in sentence prosody (intonation, rhythm and sentence stress) as well as in segmental pronunciation (of both vowels and consonants). Finally, naïve listener ratings in the two tasks (with and without segmental information) are significantly correlated, suggesting that the prosodic information used in the nativeness task is also important in the comprehensibility task, confirming therefore the analysis associating comprehensibility ratings and problem areas.

The results of the rating studies, then, provide useful information for future work towards establishing a framework for designing computer-aided pronunciation training tools. First, the studies show that experienced and naïve raters agree in their judgements of L2 comprehensibility, so that there is no advantage in using language teachers as experts to evaluate comprehensibility, compared to using naïve raters. Second, the studies also show that naïve listeners are no less reliable than experienced raters in distinguishing between native and non-native accents on the basis of prosodic information alone. Note that this pattern differs from that presented by Thompson (1991), who found greater inter-rater reliability in accentedness judgements from experienced raters than from naïve raters. It should be stressed however, that there are important differences between her studies and ours. Most importantly, our raters listened to low-pass filtered speech to arrive at judgements of nativeness, while Thompson's raters made accentedness judgements on unfiltered speech. Recall that we used filtered speech because of our primary interest in the prosodic aspects of speech, which were the chosen target of the computational part of our overall research project. Prosody and intonation are perhaps the least well covered aspects of pronunciation in typical English teacher-training programme, and so it should come as no surprise that our experienced raters, English language teachers, were no more reliable than our naïve raters. In consequence, apart from being able to request ratings of specific aspects of speech production, for which a certain degree of familiarity with phonetic description would be useful, there seems little advantage in recruiting experienced raters rather than using more readily available untrained listeners.

In addition, our rating studies have confirmed that specific features of both prosodic and segmental aspects of speech, as identified by experienced raters, correlate well with the overall judgements of comprehensibility of L2 utterances by naïve speakers. This finding is in line with Munro and Derwing's (2001) conclusion based on previous studies (Anderson-Hsieh, Johnson, & Koehler, 1992; Brennan & Brennan, 1981; Munro & Derwing, 1999) that "simple counts of segmental errors and prosodic assessments correlate well with listeners' ratings of L2 speech on such dimensions as accentedness and comprehensibility, whether or not the listeners are phonetically trained." Cucchiarini et al.'s (2000b) study, which compared automatic scores produced by speech recognition algorithms with expert ratings of pronunciation quality, also shows that specific ratings collected from expert raters (phoneticians and speech therapists) were highly correlated with the overall pronunciation ratings. Cucchiarini et al. (ibid) conclude that these findings "warrant the use of overall ratings of pronunciation as a sole reference for the automatic score".

Finally, the findings of the factor analysis, which group together sentence stress, intonation and rhythm as a sentence prosody factor, warrant an approach to software development that includes all three features in the learning activities aimed to improve sentence prosody. This is of course not to deny that the other significant factors – word stress, consonant pronunciation and vowel pronunciation – also need to be treated within the pedagogical framework used in software development.

SUMMARY

In the context of developing software that would offer useful and effective feedback to Mandarin speaking learners of English on their pronunciation, we have assessed the relative importance of different speech features through the effect they have on the communicative quality of the utterance, measured by comprehensibility ratings. Such data are important to the issue of how to evaluate and fine-tune the acoustic information that the software derives from learner speech and subsequently uses in assessing learner performance.

We have identified a number of issues that need to be addressed in developing pedagogical and software models for learner pronunciation instruction. It was clear that prosodic features have an important effect on comprehensibility, a finding that supports previous studies suggesting that time spent on such features is well justified (see supporting references discussed in our Introduction). Rehearsal of prosodic features in a semi-communicative context can be provided through software that targets features that have the strongest effect on comprehensibility, and a conscious awareness of those features can be raised through a number of explanatory notes associated with the feedback that the software provides. Feedback, rehearsal and language awareness are three learning opportunities that are well supported in curriculum development (Crabbe, 2003).

It has also been acknowledged that accuracy, relevance and ease of interpretation are key issues in the provision of feedback through automated software for CAP (Computer Assisted Pronunciation). The two main problems with existing CAP software are the limitations of automatic speech recognition technologies which are yet to reach maturity, and the lack of clear pedagogical basis in software design. In order to address technological limitations, the research reported on here set out to establish relevant comprehensibility data to be used as a feedback parameter in developing CAP software.

Our exploration of a methodology for incorporating native speaker judgements into decision making on the parameters used in developing pronunciation feedback software offers a useful contribution in this area. Our initial results show that holistic comprehensibility ratings by naïve native speakers provide good information with which to fine-tune CAP software for prosodic features. This would imply that where the development of such software incorporates native speaker judgements in determining acceptability, then using naïve speakers is sufficient for this purpose. We believe that the exploration of how such native speaker judgements can be used as a parameter in selecting features for automated feedback on pronunciation is a productive area for further research.

ACKNOWLEDGEMENTS

The research reported here formed part of a project funded by the New Economy Research Fund administered by the NZ Foundation for Research Science and Technology (contract number: VICX0011). We would like to acknowledge the contributions made to this research by the other project members Peter Andreae, Mengjie Zhang, Jason Xie and Mike Doig, and to thank Ave Coxhead and our panels of raters for their valuable assistance. Our thanks go also to the editors and three anonymous reviewers for their constructive comments on an earlier version of this paper.

REFERENCES

Adams, C. (1979). *English speech rhythm and the foreign learner*. The Hague: Mouton.

Agresti, A., & Liu, I. (2001). Strategies for modeling a categorical variable allowing multiple category choices. *Sociological Methods and Research*, 29, 403-434.

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of non-native pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning*, 42, 529-555.

Benrabah, M. (1997). Word-stress - a source of unintelligibility in English. *IRAL*, XXXV(3), 157-165.

Breitkreutz, J., Derwing, T., & Rossiter, M. (2002). Pronunciation teaching practices in Canada. *TESL Canada Journal*, *19*, 51-61.

Brennan, E., & Brennan, J. (1981). Measurements of accent and attitude towards Mexican-American speech. *Journal of Psycholinguistic Research*, 10, 487-501.

Burgess, J., & Spencer, S. (2000). Phonology and pronunciation in integrated language teaching and teacher education. *System*, *28*, 191-215.

Chao, Y. R. (1980). Chinese tones and English stress. In L. R. Waugh & C. H. van Schooneveld (Eds.), *The melody of language: intonation and prosody* (pp. 41-44). Baltimore: University Park Press.

Crabbe, D. (2003). The quality of language learning opportunities. *TESOL Quarterly*, 37(1), 9-34.

Cucchiarini, C., Strik, H., & Boves, L. (1997). *Automatic evaluation of Dutch pronunciation by using speech recognition technology*. Paper presented at the 1997 IEEE workshop ASRU, Santa Barbara.

Cucchiarini, C., Strik, H., & Boves, L. (2000a). Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms. *Speech Communication*, 30, 109-119.

Cucchiarini, C., Strik, H., & Boves, L. (2000b). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *Journal of the Acoustical Society of America*, 107(2), 989-999.

Derwing, T., & Rossiter, M. (2003). The effects of pronunciation instruction on the accuracy, fluency, and complexity of L2 accented speech. *Applied Language Learning*, *13*, 1-17.

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, 19(1), 1-16.

Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, 39, 379-397.

Derwing, T. M., Munro, M. J., & Carbonaro, M. D. (2000). Does popular speech recognition software work with ESL speech? *TESOL Quarterly*, 34, 592-603.

Derwing, T. M., Munro, M. J., & Wiebe, G. (1998). Evidence in Favour of a Broad Framework for Pronunciation Instruction. *Language Learning*, 48(3), 393-410.

Doughty, C. J., & Long, M. H. (Eds.). (2003). *The handbook of second language acquisition*. Malden, MA: Blackwell.

Ehsani, F., & Knodt, E. (1998). Speech technology in computer-aided language learning: strengths and limitations of a new CALL paradigm. *Language Learning and Technology*, 2(1), 45-60.

Gass, S., & Varonis, E. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning*, 34, 65-89.

Grabe, E. (2002). Variation adds to prosodic typology. In B.Bel & I. Marlin (Eds.), *Proceedings of the Speech Prosody 2002 Conference* (pp. 127-132). Aix-en-Provence: Laboratoire Parole et Langage.

Hahn, L. D. (1999). *Native speakers' reactions to non-native stress in English discourse.* Unpublished dissertation, University of Illinois at Urbana-Champaign.

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, 38, 201-233.

Hansen, J. G. (2001). Linguistics constraints on the acquisition of English syllable codas by native speakers of Mandarin Chinese. *Applied Linguistics*, 22(3), 338-365.

Hardison, D. M. (2004). Generalization of computer-assisted prosody training: Quantitative and qualitative findings. *Language Learning & Technology*, 8, 34-52.

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. New York: Newbury House.

Hedge, T. (2000). *Teaching and learning in the language classroom*. Oxford: Oxford University Press.

Hirata, Y. (2004). Computer-assisted pronunciation training for native English speakers learning Japanese pitch and duration contrasts. *Computer Assisted Language Learning*, 17, 357-376.

Hutchinson, S. P. (1973). *An objective index of the English-Spanish pronunciation dimension.* Unpublished Masters thesis, University of Texas.

Juffs, A. (1990). Tone, syllable structure and interlanguage phonology: Chinese learners' stress errors. *IRAL*, XXVIII(2), 99-115.

Kenworthy, J. (1987). *Teaching English Pronunciation*. New York: Longman.

Kim, I. S. (2006). Automatic speech recognition: Reliability and pedagogical implications for teaching pronunciation. *Educational Technology and Society*, 9(1), 322-344.

Kratochvil, P. (1998). Intonation in Beijing Chinese. In D. Hirst & A. di Cristo (Eds.), *Intonation Systems* (pp. 417-431). Cambridge: Cambridge University Press.

Levis, J. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, 39(3), 369-378.

Levis, J. (2007). Computer technology in teaching and researching pronunciation. *Annual Review of Applied Linguistics*, *27*, 184-202.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

Lin, Y.-H. (2001). Syllable simplification strategies: a stylistic perspective. *Language Learning*, 51(4), 681-718.

MacDonald, S. (2002). Pronunciation-views and practices of reluctant teachers. *Prospect*, *17*(3), 3-18.

Morley, J. (1991). The pronunciation component of teaching English to speakers of other languages. *TESOL Quarterly*, 25, 481-520.

Moyer, A. (1999). Ultimate attainment in L2 phonology. The critical factors of age, motivation and instruction. *Studies in Second Language Acquisition*, 21(1), 81-108.

Munro, M. J. (1995). Nonsegmental factors in foreign accent: ratings of filtered speech. *Studies in Second Language Acquisition*, 17, 17-34.

Munro, M. J., & Derwing, T. M. (1995). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech*, 38(3), 289-306.

Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 49(Supp. 1), 285-310.

Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech. *Studies in Second Language Acquisition*, 23, 451-568.

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System*, 34, 520-531.

Murphy, J. (1997). Phonology courses offered by MATESOL programs in the United States. *TESOL Quarterly*, *31*(4), 741-764.

Nation, P., & Heatley, A. (2001). RANGE. A program for measuring the lexical burden of texts. Wellington: School of Linguistics and Applied Language Studies, Victoria University of Wellington.

Nelson, C. (1982). Intelligibility and non-native varieties of English. *The other tongue: English across cultures*, 15, 59-73.

Neri, A., Cucchiarini, C., Strik, H., & Boves, L. (2002). The pedagogy-technology interface in computer assisted pronunciation training. *Computer Assisted Language Learning*, 15(5), 441-467.

Nunan, D. (1999). *Second language teaching and learning*. Boston: Heinle & Heinle.

Ono, Y. (1991). Experimental Phonetic Analysis of the Speech Sounds and Prosodic Features Produced by Native and Non-native Speakers. *Language and Culture*, 20, 241-288.

Pennington, M. C. (1992). Discourse factors related to L2 phonological proficiency: An exploratory study. In J. L. A. James (Ed.), *Proceedings of New Sounds 92* (pp. 137-155). Amsterdam: University of Amsterdam.

Pennington, M. C. (1999). Computer-aided pronunciation pedagogy: promise, limitations, directions. *Computer Assisted Language Learning*, 12, 427-440.

Pennington, M. C., & Ellis, N. C. (2000). Cantonese speakers' memory for English sentences with prosodic clues. *The Modern Language Journal, 84*, 372-389.

Pennington, M. C., & Richards, J. (1986). Pronunciation revisited. *TESOL Quarterly*, 20, 207-226.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, 73, 265-292.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime User's Guide*. Pittsburgh: Psychology Software Tools, Inc.

Shen, X.-n. S. (1990). *The Prosody of Mandarin Chinese* (Vol. 118). Berkeley: University of California Press.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin*, 86(2), 420-428.

Tajima, K., Port, R., & Dalby, J. (1997). Effects of temporal correction on intelligibility of foreign-accented English. *Journal of Phonetics*, 25, 1-24.

Thompson, I. (1991). Foreign accents revisited: The English pronunciation of Russian immigrants. *Language Learning*, 41(2), 177-204.

Tiffen, B. (1992). A study of the intelligibility of Nigerian English. In A. v. Essen & E. I. Burkart (Eds.), *Homage to W.R.Lee: essays in English as a foreign or second language* (pp. 255-259). Berlin: Foris.

Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition*, *28*, 1-30.

Van Els, T., & De Bot, K. (1987). The role of intonation in foreign accent. *The Modern Language Journal*, 72, 147-155.

Warren, P. (2002). NZSED: building and using a speech database for New Zealand English. *New Zealand English Journal*, 16, 53-58.

Warren, P., & Britain, D. (2000). Intonation and prosody in New Zealand English. In A. Bell & K. Kuiper (Eds.), *New Zealand English* (pp. 146-172). Wellington: Victoria University Press.

Weinberger, S. H. (1997). Minimal segments in second language phonology. In A. James & J. Leather (Eds.), *Second Language Speech: Structure and Process* (pp. 263-312). Berlin: Mouton de Gruyter.

Xie, H., Andreae, P., Zhang, M., & Warren, P. (2004a). Detecting stress in spoken English using decision trees and support vector machines. In M. Purvis (Ed.), *Proceedings of the Australasian Workshop on Data Mining and Web Intelligence (DMWI2004)* (pp. 145-150). Dunedin, New Zealand: Australian Computer Society, Inc.

Xie, H., Andreae, P., Zhang, M., & Warren, P. (2004b). Learning models for English speech recognition. In V. Estivill-Castro (Ed.), *Proceedings of the Twenty-Seventh Australasian Computer Science Conference (ACSC2004)* (Vol. 26, pp. 323-329). Dunedin, New Zealand: Australian Computer Society, Inc.

Xie, H., Zhang, M., & Andreae, P. (2006). Genetic programming for automatic stress detection in spoken English. In F. Rothlauf & e. al. (Eds.), *EvoWorkshops 2006* (pp. 460-471). Berlin: Springer-Verlag.

FIGURE CAPTIONS

Figure 1. Average ratings from naïve listeners for nativeness (horizontal axis) and comprehensibility (vertical axis) for 50 Mandarin English utterances (10 utterances from each of 5 speakers). Rating scales range from 1 to 9 in each case (see text for details). The two sets of ratings correlate significantly (*r*=.59, p<0.001).

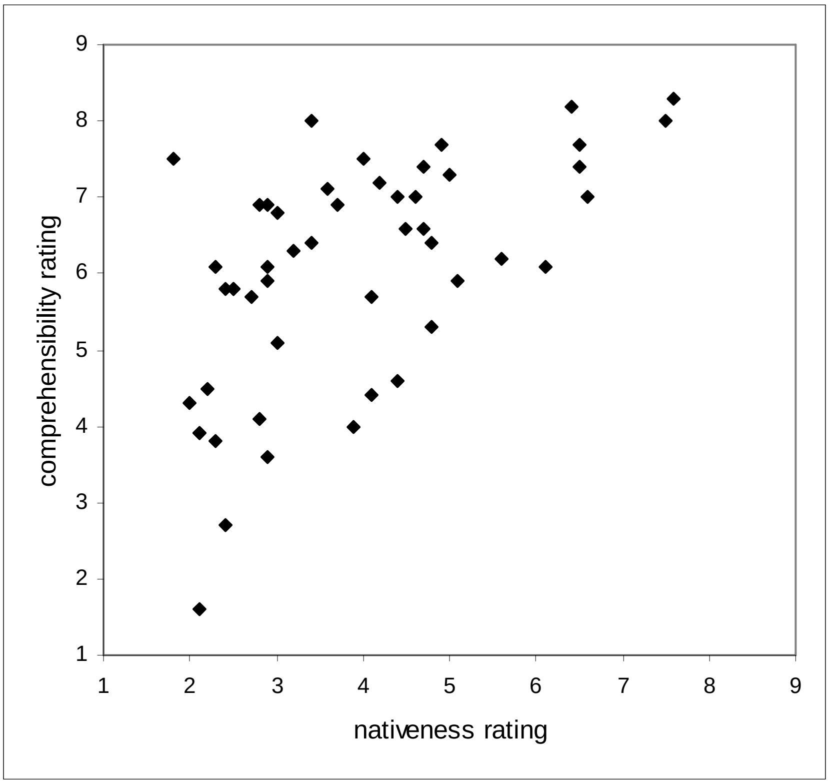

Figure 1

NOTES

[1] Connected Speech (2001). Protea Textware Pty Ltd. http://www.proteatextware.com.au/. Last accessed November, 2008.
[2] ISLE (Interactive Spoken Language Education). The ISLE Consortium. http://nats-www.informatik.uni-hamburg.de/~isle/index.html. Last accessed November, 2008.
[3] For example, some older studies cited in Cucchiarini, Strik, & Boves (2000b: 991) seem to indicate that reliability of expert fluency ratings maybe low, even though their own study did not corroborate this. Thompson (1991: 195), on the other hand, observed that experienced listeners were more reliable and more lenient in their accentedness ratings than inexperienced listeners.
[4] A reviewer has suggested that one of the prosodic differences between the native and non-native recordings used in our experiment might have resulted from the New Zealand tendency to use High Rising Terminals, i.e. rising intonation patterns on statement utterances. In fact, these are extremely rare in sentence readings (and were absent from our recordings), since they function largely as discourse markers in conversations or in longer narratives (see Warren & Britain, 2000).
[5] Levene's test showed no significant difference in the variances for the two rater groups.