

COMPREHENSIBILITY AND PROSODY RATINGS FOR PRONUNCIATION SOFTWARE DEVELOPMENT

Paul Warren, Irina Elgort, David Crabbe

Victoria University of Wellington

In the context of a project developing software for pronunciation practice and feedback for Mandarin-speaking learners of English, a key issue is how to decide which features of pronunciation to focus on in giving feedback. We used naïve and experienced native speaker ratings of comprehensibility and nativeness to establish the key features affecting comprehensibility of the utterances of a group of Chinese learners of English. Native speaker raters assessed the comprehensibility of recorded utterances, pinpointed areas of difficulty and then rated 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 setting up automated feedback.

INTRODUCTION

Most learners of English need to develop their pronunciation ability to the point where it has no serious effect on comprehensibility when they are engaged in oral communication. Some develop this skill naturally over time and to a reasonable level of accuracy through imitation of one native speaker norm or another. Others need to work harder at it with expert guidance. Morley (1991, p. 492-495) provides a comprehensive overview of groups of learners in special need of pedagogical support for pronunciation for both ESL (English as a second language) and EFL (English as a foreign language) settings. Yet, Derwing and Munro (2005) lament the “marginalization of pronunciation within applied linguistics” (p. 382). Their informal survey shows either 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 few papers on pronunciation in journals in applied linguistics. Research in Canada, Britain and Australia shows that in addition to a lack of training in pronunciation instruction, English teachers, in general, do not have a strong enough background in phonetics to feel confident to teach pronunciation (Breitkreutz, Derwing, & Rossiter, 2002; Burgess & Spencer, 2000; MacDonald, 2002; Murphy, 1997). Moreover one of the key features of a pedagogy of pronunciation is necessarily feedback on performance, and yet providing pronunciation feedback is an intensive, time-consuming activity requiring one-to-one work. 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 long dominated over a focus on form, including phonetic form.

In this context, the computer-assisted language learning approach appears to be promising, as it can enable students to work on improving their pronunciation independently, focusing on aspects of pronunciation relevant to individual needs, based on L1 (first language) background and 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, p. 391; see also Neri, Cucchiaroni, 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, p. 186; see also Neri, et al., 2002).

The data reported in the current paper form part of a larger collaborative project, involving researchers in phonology, computer science, and second language pedagogy. The project explores the provision of automated feedback on learners' 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 learner's utterance with a target norm stored in the system (for reviews of the issues, see Ehsani & Knodt, 1998; Levis, 2007; Neri, et al., 2002; Pennington, 1999). Such programs are not yet developed to the point where all of the automated responses are useful in guiding learners towards improving their performance, but this is a productive field in which gradual progress is being made (see, for example, *Connected Speech* by Protea Textware¹ or ISLE software produced by the European ISLE Consortium²). Our aim was to ensure that software development is informed by linguistic understanding, particularly of comprehensibility parameters. The project component presented in this paper aims to identify the principal speech features that contribute to comprehensibility and nativeness.

ACCENTEDNESS AND NATIVENESS, INTELLIGIBILITY, AND COMPREHENSIBILITY

Levis (2007, p. 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, p. 291), or “how different a speaker's accent is from that of the L1 community” (Derwing & Munro, 2005, p. 385), while intelligibility commonly refers 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.” Derwing and Munro (1997, p. 6) use both terms for one of their tasks—their response continuum ranges from “perfectly nativelike” to “extremely accented.” For the current study we have chosen to use the label “nativeness,” primarily because our rating task uses low-pass filtering of speech in order to focus attention on prosodic features, and this results in the loss of the segmental details that contribute strongly to what is perceived as an accent in a language.

Two further terms that need to be carefully distinguished are *intelligibility* and *comprehensibility*. The former is frequently 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 effort associated with understanding a NNS is indicative of the listener's ability to correctly process the NNS utterances. In the current study we were concerned with comprehensibility ratings (i.e., a measure of the effort required by raters to understand the utterances they are asked to listen to).

PROSODY, COMPREHENSIBILITY, AND NATIVENESS

One reason for our focus on prosodic features in this study was that their impact on intelligibility has been acknowledged both in longstanding teacher beliefs and, more recently, in pronunciation instruction research (Derwing, Munro, & Wiebe, 1998; Derwing & Rossiter, 2003; Hahn, 2004). This has led to an increased recognition of the role of prosody in the comprehensibility and accentedness of native and non-native speech (Anderson-Hsieh, Johnson, & Koehler, 1992; Munro & Derwing, 1999), with prosodic factors often producing more extreme results than segmental factors (Anderson-Hsieh, et al., 1992;

Benrabah, 1997; Hutchinson, 1973; Tiffen, 1992). Indeed, inappropriate timing and stress patterns are often cited as major contributors to intelligibility deficit (Adams, 1979; Hahn, 1999, 2004; Kenworthy, 1987; Nelson, 1982) or “unnaturalness” (Ono, 1991).

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 segmental aspects alone, or both segmental and suprasegmental aspects of learners’ performance. The type of phonological feedback significantly affected learning outcomes, and “subjects who were given both suprasegmental and segmental feedback scored closer to native” (Moyer, 1999, p. 95). In another study (Derwing, et al., 1998), three instruction types were used: two based on pronunciation, that is, segmental and global (the latter including stress, intonation and rhythm), and one with specific pronunciation instruction (providing a control group). Sentences read aloud and learner-produced narratives were recorded at the beginning and end of a 12-week course of instruction. Non-expert native speakers rated the sentences for comprehensibility and accentedness, and excerpts from the narratives for comprehensibility, accentedness and fluency. Training resulted in improvement in the read sentences for both the segmental and global groups, while for the narrative data only “speakers who had had instruction emphasizing prosodic features such as rhythm, intonation, and stress could apparently transfer their learning to a spontaneous production” (Derwing, et al., 1998, p. 406). Hardison (2004) also found that computer-assisted prosody training with a real-time pitch display produced significant improvement in both prosody and segmental accuracy, as judged by native speaker raters, and Hirata (2004) found a similar effect for English-speaking learners of Japanese.

Our project focused on Mandarin-speaking learners of English (MSLEs) both as the largest group of English language learners, and also as a group that is likely to be particularly affected by important language differences in key aspects of prosodic structure (Hansen, 2001; Pennington & Ellis, 2000; Pennington & Richards, 1986). These include the lexical use of tone in Mandarin but not in English; differences in basic rhythmic structures (Adams, 1979; Grabe, 2002); and the greater use in Mandarin of tonal range to indicate stress (Kratochvil, 1998; Shen, 1990). Thus 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. Similarly, Juffs (1990) found that the most frequent stress errors in Chinese English result from using a tonic stress movement to mark lexical stress, and that differences in the syllable structure of the languages also affect stress assignment. Tajima, Port and Dalby (1997) observed many segmental errors in Mandarin English that reflect a tendency to avoid consonant clusters by either deleting consonants or inserting epenthetic vowels (see also Hansen, 2001; Lin, 2001; Weinberger, 1997), impacting the rhythmic pattern of the utterance. They also noted a reduced difference between stressed and unstressed vowel durations.

Munro and Derwing (1999) noted that intonation is important in native speaker ratings of comprehensibility and accentedness of Mandarin English. 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%. Their alignment procedures (p. 8-9) also involved so-called “discrete” changes (i.e., removing or inserting segments that were or were not in the original Mandarin, to match the English target). They 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” (p. 21).

The initial prototype software module used in our project focused on stress patterns as one key feature affecting comprehensibility. Recognition trials—using a combination of features based on vowel duration, amplitude, pitch and vowel quality—produced automatic stress recognition rates for NS (native speaker) English of up to 92.6% (Xie, Andreae, Zhang, & Warren, 2004). Duration and amplitude were the most

useful features, along with the vowel quality features associated with reduced (therefore unstressed) vowels. Although these results are comparable to those produced by similar systems, they still do not provide an adequate basis for feedback to language learners. Fine-tuning the parameters used by the software might result in some improvement in recognition. But so too might a strategy of allowing the software development to be informed by native speaker judgements of non-native speech, just as feedback provided to learners by CAP software should be consistent with human feedback (Cucchiaroni, Strik, & Boves, 2000a; Derwing, et al., 2000; Kim, 2006; Levis, 2007). Thus, it is critical to establish which prosodic features affect NS listener judgements of comprehensibility and nativeness, in order to evaluate the analysis measures used by the software. The rest of this paper reports on the procedures we used to gather data on native speaker perceptions of MSLE utterances, and discusses the results and possible implications for the use of the data.

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.

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

Our nativeness ratings focused on prosody. This was achieved by using low-pass filtered speech (see also Derwing & Munro, 1997; Munro, 1995; Trofimovich & Baker, 2006; Van Els & De Bot, 1987), removing detailed information concerning the consonant and vowel segments in the speech and causing 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 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 assess the nativeness of 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 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 versus naïve listeners for such judgements. For instance, previous research (Thompson, 1991) has indicated higher reliability in accentedness judgments from experienced raters.

Speech Material

The source materials were from 5 Mandarin Speaking Learners of English (MSLEs) enrolled in 12-week English language courses at Victoria University of Wellington. Only female speaker recordings were used in this study, in order to simplify the speech analysis parameters used in the computational component of the project. The ages of these 5 speakers 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 equivalent to at least IELTS 6.0). They had been in New Zealand for at least 10 weeks, and all

subsequently entered degree programmes at Victoria University of Wellington. Given their location and their intentions regarding further study, it can be claimed that New Zealand English was at the time of the experiment their target variety.

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 used, rather than excerpts from free narratives (Derwing & Munro, 1997; Derwing, et al., 1998), reflecting similar studies that compare listener rating with automatic speech recognition/evaluation (Cucchiaroni, Strik, & Boves, 1997; Cucchiaroni, et al., 2000a; Cucchiaroni, Strik, & Boves, 2000b). Using sentence materials based on those in NZSED also meant that we had access to a large set of comparison NS materials, which was exploited in developing materials for the nativeness rating task.

We selected a set of 100 sentences from the 200 used in the NZSED project. The selected sentences contained no low frequency words, as determined by the Range program (Nation & Heatley, 2001), and no other words that were likely to be unfamiliar to our target learner population, as judged by an experienced English language teacher. This reduced the likelihood of word mispronunciation by non-native speakers due to unfamiliarity. The 100 sentences were then read aloud (after quiet reading for familiarisation) by 5 female MSLEs. A final set of 50 utterances (10 from each of 5 speakers) was chosen so as to optimize the range of segmental and prosodic features of MSLE speech and to exclude hesitations, repeats or restarts. The sentences in this final set had an average word length of 11.3 words (range 7-15), and were long enough for rhythm and rate characteristics of the speaker to emerge. Examples are given in (1) and (2) below.

- (1) The price range is smaller than any of us expected
- (2) The world is becoming increasingly dangerous but hardly anyone cares

A further 50 utterances from age-matched female native speakers in the NZSED project were included in the nativeness rating task. Again, this set consisted of 10 sentences from each of 5 speakers. They were different sentences from the materials selected from the MSLEs, and had an average length of 11.9 words. In other respects (lexical frequency, etc.) they were comparable with the MSLE sentences.

For the nativeness-rating task, the native and non-native speech materials were subjected to low-pass filtering (with a cut-off frequency of 350 Hz), removing most of the segmental information, while leaving prosodic features largely intact (see also Derwing & Munro, 1997; Trofimovich & Baker, 2006). In addition to forcing the judgement of nativeness to be based on prosodic features, this also has the advantage of reducing the impact of any possible mismatch between the target English variety of the learners and that of the raters, since such a mismatch is likely to be carried by segmental features such as vowel quality.³

Raters

Ten naïve and six experienced raters were used in the study, all native speakers of New Zealand English. 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 group had no regular contact with Mandarin speakers of English or any other non-native speakers of English. The experienced group consisted of teachers in 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 in 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. The majority of studies which involve native speaker ratings of L2 (second language) pronunciation use either only expert raters (Anderson-Hsieh, et al., 1992; Cucchiaroni, et al., 1997, 2000a, 2000b) or only naïve raters (Derwing & Munro, 1997; Munro & Derwing, 1995). Studies that use experts sometimes

include raters from different expert backgrounds (Cucchiaroni, et al., 2000a, 2000b), for example, 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 and 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 or experienced raters on the one hand, and naïve or inexperienced raters on the other. For example, older studies cited in Cucchiaroni et al. (2000b) indicate low reliability for expert fluency ratings. However, Thompson (1991) observed that experienced listeners were more reliable and more lenient in accentedness ratings than inexperienced listeners.

Procedure

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

- i) First, they rated the comprehensibility of the recorded utterances. The following clarification was provided to encourage raters to focus on comparable criteria: “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) except that theirs ranged from “extremely easy to understand” to “extremely difficult or impossible to understand” (p5). Note also that in their methodological study of scales used in accent rating, Southwood and Flege (1999) indicate that using anything with fewer than 9 points is likely to yield unsatisfactory outcomes.
- ii) Raters were then presented with the orthographic transcription in a response booklet and were asked to mark 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 the following additional task:

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., a range of segmental and suprasegmental features 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 be able to provide such data in a readily interpretable form, because of unfamiliarity with the appropriate linguistic terminology. This additional task distinguishes our study from previous comprehensibility studies, where listeners either only rate overall comprehensibility, or are required to assign specific ratings for identified features, rather than actually identifying features that cause difficulty in comprehension.

In the second session, raters were asked to provide nativeness ratings (“Enter your rating of how much this was like a native-speaker”) for each of the 100 examples of low-pass filtered speech (50 NS utterances along with the 50 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) similarly used a 9-point scale in their accentedness task, but with reversed endpoints, 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). 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 based their judgements of nativeness on remembered aspects of the NNS utterances previously heard in the 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 was based solely on the available prosodic information.

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 summary 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

So that we could be confident that our rating data would be of use in software development, we first assessed inter-rater agreement, by two methods. First we transformed correlations between each pair of raters into Z-scores and calculated the mean (Hatch & Lazaraton, 1991). 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 $p < .01$, and an intraclass correlation coefficient (ICC) of .954, $p < .01$. 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 $p < .01$ and an ICC of .931, $p < .01$. 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 $p < .01$). The lower figures for native speakers most likely resulted 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, and ICC of .968 and .987 for comprehensibility and accent ratings reported by Munro and Derwing, 2006).

Because we wished to assess our comprehensibility ratings against identification of problem areas by the group of experienced listeners, we also needed to assess the reliability of comprehensibility ratings given by this group alone. Munro and Derwing (1995), for example, pointed out that previous research (e.g., Gass & Varonis, 1984) has shown that comprehensibility rating “tends to improve with increased exposure to foreign-accented speech” (p. 297), 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 NNSs 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). In addition, we wished to compare ratings from naïve and experienced listeners (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 showed a good level of reliability, and whether there were 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 $p < .01$, 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 ($t(49) = 0.528$, $p = .60$). In addition, a correlation analysis of the utterance means for each group showed a high level of agreement between experienced and naïve listeners ($r = .92$, $p < .001$).

Comprehensibility and areas of difficulty

The above analyses have confirmed good overall levels of inter-rater reliability in both tasks, and a high level of agreement between the two rater groups in the comprehensibility task. 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 addressed the question of whether the comprehensibility ratings given by experts were 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 $p < .01$), as well as rhythm and word stress (each at $p < .05$), 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 (e.g., Munro & Derwing, 2001), we found that

problems in speech rate showed no significant association with the comprehensibility rating. The generally slow rate of the NNS utterances may have made it difficult for the listeners to discriminate between them in terms of speech rate issues. (The mean rates in syllables/second were 3.08 and 5.48 for NS and NNS respectively, $t(98) = 20.64$, $p < .001$.)

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 related to the nativeness ratings. These were obtained, as indicated above, in order to require 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, $p < .01$). However, more detailed analysis shows numerically greater reliability for the 10 naïve listeners than for the 6 experts, with r at .73 and .69 and ICC at .904 and .822 for the two groups respectively (significant at $p < .01$). (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 ($F(1,14) = 50.65$, $p < .001$)⁴, but no interaction of speaker group with rater group ($F(1,14) = 2.55$, $p > .1$). Since our subsequent analysis of comprehensibility was based on data only from our naïve listeners (recall that our experts were not asked to complete this part of the test), we were reassured that the results presented in this section failed 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 went some way towards achieving this goal. The analysis of nativeness ratings also contributed in this direction, in that we could select items simply on the basis of low scores in this task. However, we were also interested in the relationship between comprehensibility and perceived nativeness, and in particular in any association between the two. The presence of a positive relationship might suggest that the prosodic features not eliminated by the low-pass filtering were indeed contributing to comprehensibility. So our next question was whether the nativeness ratings (of low-pass filtered speech, so based largely on prosodic features) and comprehensibility ratings (of unfiltered speech, so including segmental features) from our naïve listeners were 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, we addressed this question 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. There was a significant overall correlation ($r = .59$, $p < .001$), confirming a positive relationship between nativeness and comprehensibility. Note also that the data are distributed in a manner that indicates that perceived nativeness provides a baseline on top of which comprehensibility appears to be built. That is, comprehensibility ratings usually exceeded nativeness ratings for individual utterances, and were rarely lower than the nativeness ratings. Indeed, our results here mirror those of Derwing and

Munro (1997), who observed that “accent ratings are harsher than perceived comprehensibility ratings” (p. 11).

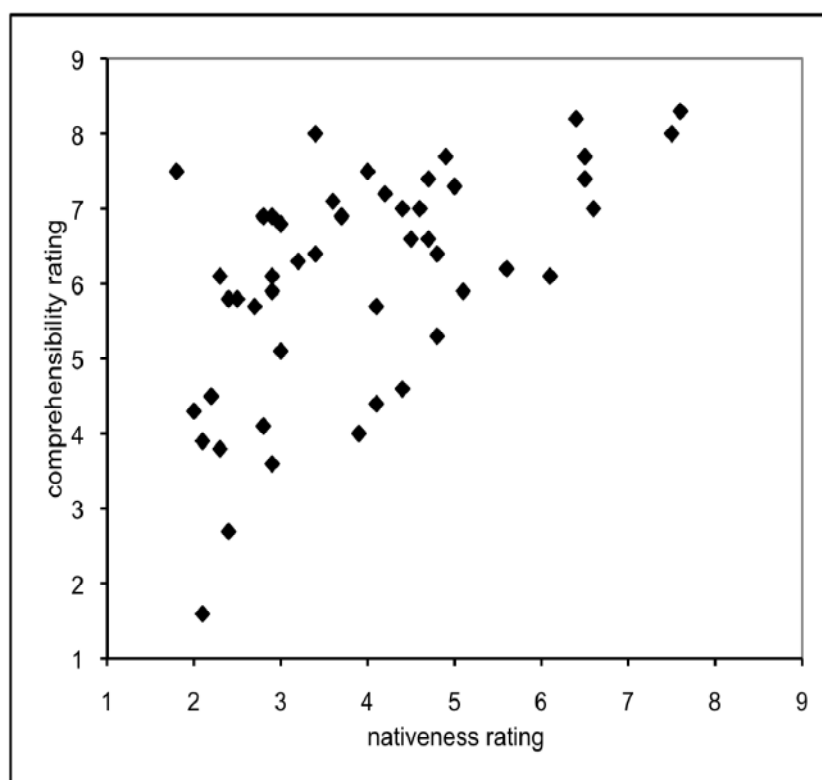


Figure 1. Average ratings from naïve listeners for nativeness (horizontal axis) and comprehensibility (vertical axis) for 50 Mandarin English utterances (10 utterances from each of 5 speakers). Rating scales range from 1 to 9 in each case (see text for details). The two sets of ratings correlate significantly ($r = .59$, $p < .001$).

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, and confirming 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 evidence of an advantage in using language teachers. 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 programmes, 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, et al., 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" (p.453). Cucchiari et al.'s (2000a) 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. Cucchiari et al. conclude that these findings "warrant the use of overall ratings of pronunciation as a sole reference for the automatic score" (p.118).

Finally, the findings of the factor analysis, which groups 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. 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 a clear pedagogical basis in software design. In order to address technological limitations, the research reported 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.

NOTES

1. Connected Speech (2001). Protea Textware Pty Ltd. <http://www.proteatextware.com.au>
2. ISLE (Interactive Spoken Language Education). The ISLE Consortium. <http://nats-www.informatik.uni-hamburg.de/~isle/index.html>
3. A reviewer has suggested that a prosodic difference between the native and non-native recordings used in our experiment—and therefore a potential difference between target varieties for the learners and the raters—might lie in 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).
4. Levene's test showed no significant difference in the variances for the two rater groups.

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 Andrae, 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.

ABOUT THE AUTHORS

Paul Warren is Associate Professor in the School of Linguistics and Applied Language Studies at Victoria University of Wellington, New Zealand. Paul's primary research interests are in psycholinguistics, in particular spoken word recognition and the use of intonation in sentence processing. Since moving to New Zealand in 1994, he has combined these interests with a growing fascination in the development of New Zealand English.

E-Mail: paul.warren@vuw.ac.nz

Dr. Irina Elgort is a lecturer in academic development at Victoria University of Wellington, New Zealand. Her research interests include L2 vocabulary acquisition, reading and computer assisted language learning (CALL). She teaches a CALL paper in the MA in TESOL/Applied Linguistics programme at Victoria.

E-Mail: irina.elgort@vuw.ac.nz

David Crabbe is Associate Professor in the School of Linguistics and Applied Language Studies at Victoria University of Wellington, New Zealand. David works on language curriculum development and learner autonomy. He teaches and supervises in the graduate applied linguistics programme at Victoria University of Wellington and has a broader management role at the university in the area of learning and teaching.

E-Mail: david.crabbe@vuw.ac.nz

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.
- Cucchiaroni, 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.
- Cucchiaroni, 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.
- Cucchiaroni, 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. 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.
- Derwing, T. M., & Rossiter, M. (2003). The effects of pronunciation instruction on the accuracy, fluency, and complexity of L2 accented speech. *Applied Language Learning*, 13, 1-17.
- 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 & Technology*, 2(1), 45-60. Retrieved from <http://llt.msu.edu>
- 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. Retrieved from <http://llt.msu.edu>
- 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, Austin, TX.
- 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., Cucchiari, 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. (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.

- 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.
- Southwood, M. H., & Flege, J. E. (1999). Scaling foreign accent: Direct magnitude estimation versus interval scaling. *Clinical Linguistics and Phonetics*, 13, 335-349.
- 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. (2004). 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.