Poznań Studies in Contemporary Linguistics 51(2), 2015, pp. 227–246

doi: 10.1515/psicl-2015-0010

ACCENT EVALUATION FROM EXTEMPORANEOUS CHILD SPEECH

MARÍA LUISA GARCÍA LECUMBERRI
University of the Basque Country, Vitoria
garcia.lecumberri@ehu.es

MARTIN COOKE
University of the Basque Country, Vitoria

CHRISTOPHER BRYANT
University of Edinburgh

ABSTRACT

A key issue in judging foreign accent is to isolate the phonetic component from potentially confounding higher-level factors such as grammatical or prosodic errors which arise when using natural sentence-length speech material. The current study evaluated accent and intelligibility ratings of children's speech for isolated words spliced out of extemporaneous material elicited via a picture description task. Experiment 1 demonstrated that word scores and accent ratings provided by native judges pattern as in earlier studies, validating the use of word-based material derived from natural speech. In a second experiment, listeners rated the degree of foreign accent and comprehensibility for unrelated sequences of 1 to 8 words from the same talker. Degree of foreign accent was judged to increase with sequence length, asymptoting by 2 word sequences, although listeners did not rate the sequence based on the most-accented word it contains. Comprehensibility was judged to be lower as sequence length increased, asymptoting at 4 words. These findings suggest that short sequences of randomly-permuted words extracted from extemporaneous speech can be used for robust accent and comprehensibility judgements whose focus is on the phonetic basis for deviations from the native norm.

KEYWORDS: Foreign accent; extemporaneous speech; intelligibility; comprehensibility.

1. Introduction

Possession of a strong foreign accent (FA) can be unfavourable for both listeners and speakers. Listeners may require significantly more effort to process accented speech, and in extremis may fail to decode the intended message. Aside

from communicative breakdown, accented speakers can suffer from decreased motivation or prejudice (Cargile 1997; Hamers and Blanc 2000). Consequently, understanding the dimensions of foreign accent and their effects on listeners is an important challenge aimed at improving speech communication in its widest sense.

One of the outcomes from the considerable body of work examining measures of foreign accent and their correlation with variables such as processing time, age, length of residence and language activation (e.g., Varonis and Gass 1982; Munro and Derwing 1995b; Flege et al. 1997; Flege et al. 1999; Piske et al. 2001) has been a clear differentiation between *comprehensibility*, which is taken to measure the relative ease of understanding speech, and *intelligibility*, which relates to how much the utterance is actually understood. A consistent finding is that while speech with a perceived strong FA can still be quite intelligible, degree of FA (DFA), comprehensibility and intelligibility nevertheless tend to be related (Munro and Derwing 1995a).

The current article explores whether it is valid to evaluate these three dimensions of FA using isolated words extracted from extemporaneous speech, and, if so, how many word tokens in each stimulus are required to obtain robust judgements. Addressing these issues has clear didactic implications for language teaching, both for instructors, who would benefit from efficient accent testing methods, and for learners, who would receive more realistic feedback, since extemporaneous productions are closer to naturalistic performance levels than speech elicited by other means such as reading or imitation.

A key issue in FA studies concerns the relative contributions to FA made by phonetic, grammatical, lexical, prosodic and other speech features (Gass and Varonis 1984; Varonis and Gass 1985; Flege and Fletcher 1992; Munro and Derwing 1995a; Munro and Derwing 1995b; Derwing and Munro 1997; Levi et al. 2007). Unless native-like competence has been achieved, speech produced by foreign speakers is usually accompanied by deviations from native speech norms at levels other than phonetic, so it is not straightforward to isolate FA as signalled by pronunciation as opposed to other features. Indeed, listeners have been found to be influenced negatively by grammatical errors in DFA estimations (Munro and Derwing 1995a; Derwing and Munro 1997). In the former study, Munro and Derwing found a surprisingly low number of grammatical errors in spontaneously-produced speech which at the same time displayed a wide variety of accent strengths. However, their participants were highly proficient learners for whom grammar may have been less problematic than more resilient phonetic indicators of FA. For an intermediate-level cohort of second language (L2) learners tested by Derwing and Munro (1997), the role of grammatical er-

rors, and their number, was greater. The presence of grammatical errors was among the most important factors affecting comprehensibility and DFA, with grammatical scores correlated to DFA and comprehensibility ratings for many listeners, and with higher correlations than observed for phonemic accuracy, prosodic goodness and speaking rate. When investigating the speech of lower level learners, particularly in foreign language classroom settings, the amount of non-phonetic deviations increases considerably, which makes the use of spontaneous utterance-level speech for FA research problematic.

In order to minimise the influence of other types of error on pronunciation judgements, some researchers have used delayed repetition techniques (Flege, Munro and Mackay, 1995; Imai, Walley and Flege, 2005; Aoyama, Guion and Flege, 2008), scripted read speech (Munro and Derwing 1995b; van Wijngaarden 2001; van Wijngaarden, Steeneken and Hougast, 2002; Bent and Bradlow 2003; Sidaras et al., 2009) or both (Flege et al. 1995). These techniques produce grammatically and lexically correct utterances but they may compromise the naturalness of the speech obtained and can introduce other effects. Scripted speech enhances the likelihood of spelling pronunciations, particularly for words unfamiliar to the reader (Bassetti 2009; Piske et al. 2011). While in repetition techniques an effort is made to control memory effects by using delayed repetition, pronunciations may still at least partly reflect phonological memory. Aoyama et al. (2008) attempted to isolate the imitation effect by combining the technique with visual prompt presentation. Words were elicited three times: the first one with an audio model and image prompt, and the other two with just an image prompt. The authors found no differences in intelligibility scores between imitated and visually prompted productions, which apparently attests to the lack of an imitation advantage. Nevertheless, since the stimuli consisted of only 26 words, speakers could have remembered some aspects of the phonological shape they had heard with the audio model and repeating it in the visual-only mode.

The above discussion concerns non-phonetic contributions which have the potential to result in less native-like accent ratings. Another issue that arises with the use of utterance level speech is the possible *beneficial* influence of higher order information on accent evaluations, particularly when assessing intelligibility and comprehensibility, since listeners may make use of the available syntactic and semantic context to aid in understanding speech (Bradlow and Alexander 2007; Bradlow and Bent 2008) and hence provide a more positive accent rating than would be merited on the basis of phonetic distortions alone. In speech perception research the contribution of context to word recognition is a variable that is frequently taken into account or controlled for. For example,

perceptual studies often include low versus high predictability sentences when evaluating intelligibility (e.g., Mayo et al. 1997; Bradlow and Alexander 2007). On the other hand, although many FA studies have used utterance length material for intelligibility and comprehensibility assessments (e.g., Munro and Derwing 1995a; Flege et al. 1995), there is usually no factoring of the contribution of sentence context to the number of words understood or the predictability of the sentence for comprehensibility ratings. This possibility is hinted at in Munro and Derwing (1995a: 303) when they suggest that an unrecognised word may become intelligible when the rest of the utterance is understood. Similarly, Bradlow and Bent (2008) and Sidaras et al. (2009) recognise that listeners pay attention to higher order information in speech, and that this can be minimised by the use of isolated words or semantically-anomalous sentences.

One potential solution to this problem is to use isolated words for FA evaluation. Some recent FA studies have employed words for FA judgements (Levi et al. 2007; Aoyama et al. 2008; Sidaras et al. 2009). Listeners in Sidaras et al. (2009) evaluated words which had been elicited individually through reading. These words were then presented individually, with a total of 8 per speaker, for intelligibility assessment via orthographic transcription. In the case of Aoyama et al. (2008), individually-produced words were presented for DFA estimation in sequences of 4 or 5 words. Levi et al. (2007) also recorded words read individually. Words were controlled for lexical frequency and were presented individually to listeners for DFA evaluation. The main finding is that higher frequency words were evaluated as less accented than low frequency words. However, the importance of Levi et al. (2007) is in demonstrating that FA judgements are influenced by factors which are not related to speakers' production of the signal but to listeners' processing of it, not only for variables such as accent experience but in terms of the properties of the lexical material being assessed independently of a speaker's particular rendition of that material.

A common factor in the aforementioned studies which have evaluated FA through isolated words is that the words themselves were also *produced* in isolation. Consequently, while supra-lexical factors are controlled, the exemplars are not necessarily typical of natural, spontaneously-produced speech. The primary purpose of the current study is to evaluate FA dimensions using speech material elicited naturally, while simultaneously avoiding the potential confounding contributions of higher-level information to the assessment of accent. One motivation for the proposed approach is didactic. We want to be able to evaluate FA in school children learning English as a foreign language. This is a cohort whose level of grammatical and lexical competence is lower-intermediate and who vary in fluency. Our aim is to find a naturalistic way to test pronunciation which

is appropriate for children starting from a young age and at lower levels of general competence.

We use an age-appropriate storyboard-based picture description task to elicit speech, followed by segmentation of lexical tokens from the continuous signal prior to presentation to judges for accent rating. We argue that this methodology has several advantages over existing approaches: (i) it sidesteps confounding factors such as the grammatical and lexical errors (including code-switching) that we frequently observe with this learner population; (ii) it prevents judges using fluency in their ratings (which might penalise less extrovert children); (iii) it avoids potential pitfalls of imitation techniques, which may unrealistically favour this population (Snow and Hoefnagel-Höhle 1978; Aoyama et al. 2008); and (iv) it avoids the use of orthographic prompts, which might elicit unfamiliar words and spelling-pronunciations.

A second goal of the current study is to determine how many words are needed to be able to detect DFA and comprehensibility robustly. This is an issue of didactic importance since its results will determine the ease of applicability of the above elicitation technique in real classroom situations. Previous studies have produced mixed results. Munro and Derwing (1995a) found no correlation between utterance length and accent measures based on utterances of 4 to 17 words. Another study (Munro et al. 2003) found no differences in accent judgement when listeners were presented with 1, 3, 12 or 36 words in backward speech. Munro et al. argue that in the absence of segmental information, judgements must have been based mainly on voice quality. At the other extreme, Flege (1984) demonstrated that a single segment and even a 30 ms portion of a segment in the case of (/t/) were sufficient for native English speaking listeners to detect a French FA. However, the two segments isolated (/t/ /uː/ in the word *two*) were particularly indexical items for French-accented English and it is not clear whether this result generalises to other segments and languages. Park (2013) showed that the mild FA of highly proficient L2 speakers can be detected even in monosyllables. On the other hand, Ikeno and Hansen (2006) found that FA detection from phrase-length speech in a two-way response task was accurate whereas in isolated words it was at chance level even for native listeners. Unlike Park, whose listeners shared the same L1, Ikeno and Hansen presented speakers from a variety of L1 backgrounds (Canadian, UK from three disparate accent regions, US and L2 speakers from seven different L1s) to three listener groups (English, US and non-natives).

Our investigation of the effect of token length on accent rating is inspired by a study into listener adaptation to novel speakers. Kato and Kakehi (1988; reported in Kakehi 1992) found that it takes listeners 4–5 words to adapt to novel

speakers in terms of intelligibility in adverse conditions, thus indicating that enough information about the speaker has been garnered by that point and that further exposure to the speaker produces no intelligibility benefits. Kato and Kakehi's technique was to vary the number of consecutive words produced by the same talker within a sequence of 100 nonsense monosyllables. In the present study we adapted their procedure to FA estimation based on the hypothesis that listening to foreign-accented speech is in part akin to native-language speaker adaptation (cf. Bradlow and Bent 2008).

In summary, the current study investigated two main research questions. First, do measures of intelligibility, DFA and comprehensibility for extemporaneously-elicited speech in which higher-order information is excluded, pattern as in previous accent studies? Second, how many words from the same talker in each stimulus do listeners require to obtain robust accent judgements? These questions were addressed through two experiments. In Experiment 1, native judges identified isolated words extracted from extemporaneous child speech and went on to rate degree of foreign accent and comprehensibility of these items. In Experiment 2, listeners rated the accent and comprehensibility of sequences of 1, 2, 4 or 8 randomly-juxtaposed words spoken by the same speaker.

Finally, as a subsidiary issue, we were interested in determining whether the finding by Levi et al. (2007) of an effect of orthographic presentation would apply to our cohort. Intuitively, we hypothesise that their finding – that NNs are judged to have a stronger FA when listeners are made aware, by orthography, of the intended target – might be even stronger in our case of low-proficiency early language learners. We also extend the orthographic factor to degree of comprehensibility judgements.

2. Elicitation of speech materials

2.1. Task

In order to elicit natural speech without the influence of orthography or auditory memory, children were presented with two different storyboards, each depicting a narrative as a sequence of images. Each child was then asked to interpret these visual cues as they wished, and hence retold each story as naturally as possible, in their own words, with minimal interference. One of the stories was "The Three Little Pigs" and the other was based on a children's book about a panda. Participants also recounted a story that they had been reading at school.

2.2. Speakers

Twenty Spanish children aged 9–10 were recorded on two separate occasions in a sound studio in the Phonetics Laboratory at the University of the Basque Country. These children all studied at the same school and were in year 4 of primary education. They had been exposed to English used as a vehicular language for approximately 30% of their school hours since the age of 3 (preschool year 1), when they first started attending school. Additionally, eight children aged 7–10 who were either native speakers of English or bilingual in English and another language were recruited to provide "distractor" tokens in order to encourage judges to make full use of the range of accentedness and comprehensibility ratings. These speakers took part in a single recording session in a sound studio in the School of Informatics at the University of Edinburgh where they described the same two storyboards as the Spanish learners and also recounted a film that they had watched recently.

Speakers were recorded individually in semi-anechoic acoustic booths. The researchers instructed the children to speak about the two storyboards, one at a time. When they had finished describing the first story, they were asked to move to the next one. After finishing the second story they were asked to speak about a book that they had read or a film that they had watched. Speech was recorded via a close-talking microphone and a table microphone directly to a Macintosh computer using Audacity.

2.3. Word extraction

Individual words were extracted manually using Praat (Boersma 2002) from continuous speech samples produced by the speakers. Segmentation was facilitated by the fact that the children's speech was not very fast and included abundant pauses, consistent with their age and level of proficiency. Words were extracted taking care that intonation did not suggest discontinuities. Intrinsic differences in pitch amongst speakers were not monitored because, given that the speakers were children, the variation was smaller than in adult populations. The following criteria were employed to extract words: (i) only content words, such as nouns, adjectives, adverbs and verbs, were selected for use in the experiment; (ii) across the cohort as a whole, as many unique (non repeated) words as possible were selected (all children recounted the same stories and hence many potential word items were common); (iii) a minimum of 15 words was extracted from each speaker; and (iv) words chosen were mostly monosyllabic or bisyllabic words (such as *reading*, *children*, *parking*) which present no difficulties

with stress to Spanish learners. The following is a representative sample of the words extracted for the experiment: *blow*, *chair*, *stone*, *crumbs*, *sleep*, *pig*, *soup*, *running*, *yellow*, *rabbit*, *house*, *kitchen*, *wolf*, *bad*, *birds*.

3. Experiment 1: Judgements based on single words

In Experiment 1 listeners first undertook a word intelligibility task followed by an accent assessment task in which they rated both the degree of foreign accent and comprehensibility of isolated words. All tasks made use of the same set of words. For the accent assessment task, half of the listeners saw an orthographic representation of the word at the same time as the auditory stimulus, while the other half heard only the auditory stimulus.

3.1. Methods

3.1.1. Participants

Twenty four listeners (14 female, 10 male) from amongst the students and staff at the University of Edinburgh were recruited using the university's Student and Graduate Employment service. All were born in the UK and had English as their native language. Ages ranged from 17–34 (mean: 22.1 years). All were paid for their participation. Two of the listeners were pursuing beginner-level Spanish courses but were retained for the experiment.

3.1.2. Test items

Test items were chosen based on achieving an approximate balance of speech material from each non-native speaker while minimising the number of repetitions of any given word. Eight words from each speaker (including the eight distractors) were extracted, leading to a test corpus of 224 items (8 × 28 talkers), of which 211 were unique words.

3.1.3 Procedure

Testing was carried out at the School of Informatics at the University of Edinburgh, using individual sound-treated booths and Beyerdynamic DT770 head-

phones. A custom-built MATLAB software application controlled the entire experiment. All stimuli were normalised to have equal root mean square energy and 100 ms half-Hamming ramps were applied at onset and offset to minimise artefacts. All were presented at 44.1 kHz in quiet at a comfortable listening level. All listeners undertook the intelligibility task first followed by the accent assessment task, both in a single session. Listeners were given a short break between the intelligibility and the accent assessment tasks. The whole experiment took between 60 and 90 minutes.

For the word intelligibility task, listeners were told that they would hear English words. Stimuli were presented once only after which participants typed their response into an on-screen text entry box. The intelligibility task was preceded by a short practice session consisting of 8 items which had been extracted with the rest of the words but not used in the main experiments. Stimulus presentation order was randomised for each listener.

For the accent assessment task, listeners were again told that they would hear English words and that they should rate the pronunciation of each word by clicking on two scales. Seven-point scales were used following studies such as Sidaras et al. (2009). The first scale was used to rate the strength of foreign accent and consisted of a row of seven radio buttons numbered 1–7, with the two ends of the scale labelled 1 = "native-like accent" and 7 = "very strong foreign accent". The second scale was used to rate how comprehensible the word was and was identical to the accent scale except for the labels 1 = "very easy to understand" and 7 = "impossible to understand". The scales were organised in this manner so that low-numbered ratings in each case correspond to the scores for more native-like tokens. Listeners could enter the ratings in any order but could only hear the stimulus once. Listeners who belonged to the "orthography" cohort were also told that the word they were asked to rate would appear on the screen during the judgement process. As in the intelligibility task, the accent assessment task was preceded by a short practice made up of 8 words not used in the main experiments. Presentation order was randomised for each listener.

3.1.4. Post-processing

Intelligibility was computed as the percentage of target words correctly identified. Prior to the computation of intelligibility, any very obvious typos were corrected. No attempt was made to correct nonwords which were not clear typos since they were likely to be attempts by the listeners to make sense of heavily-accented speech (e.g., "estro" for "straw", "baroom" for "bathroom", "histor"

for "story"). The scale of comprehensibility judgements was inverted (i.e., higher ratings indicating higher comprehensibility) by subtracting the ratings from 8, in order to provide an intuitive match for the direction of intelligibility scores. All subsequent analyses were based on response to non-native talkers, i.e., ignoring responses to distractors.

3.2. Results

Inter-rater reliability was estimated using intraclass correlation coefficients (ICC; Shrout and Fleiss 1979), computed with the ICC function of the IRR package in R (Gamer et al. 2012). ICCs of 0.93 and 0.94 were estimated for DFA and comprehensibility ratings respectively, suggesting a high level of consistency amongst judges. Mean DFA and comprehensibility ratings were 4.84 and 5.26 respectively.

Comparing the control speakers (mononlingual and bilingual) with the experimental non-native (NN) group, listeners found words produced by the native distractors (monolinguals, 95.1%; bilinguals, 85.7%) more intelligible than words produced by the NN group (71.5%). As expected, the NN group was judged to be strongly accented while the monolingual and bilingual speakers received a low accentedness rating [$F(2,46) = 272$, $p < .001$]. NN listeners were judged to be reasonably comprehensible, though not at the level of the bilingual and monolingual controls [$F(2,46)=176$, $p < .001$].

Judges who were presented with the word both orthographically and auditorily had significantly higher DFA ratings than judges who only heard the auditory stimulus [5.06 vs. 4.62; $t(459.4) = -5.16$, $p < .001$]. However, words were judged to be equally-comprehensible by the two groups [5.31 vs. 5.21; $t(477.8) = -1.34$, $p = .18$].

To illustrate sample correlations amongst intelligibility, DFA and comprehensibility, Figure 1 plots per-speaker values for the three measures in pairs. Judgements based on orthographic + auditory presentation are shown separately from judgements based on the auditory stimulus alone. Accentedness is negatively-correlated with intelligibility [with orthography: $r = -0.70$, $p < .001$; without orthography: $r = -0.58$, $p < .01$] and with comprehensibility [both $r = -0.75$, $p < .001$], while intelligibility and comprehensibility are positively-correlated [with orthography: $r = 0.90$, $p < .001$; without orthography: $r = 0.78$, $p < .001$].

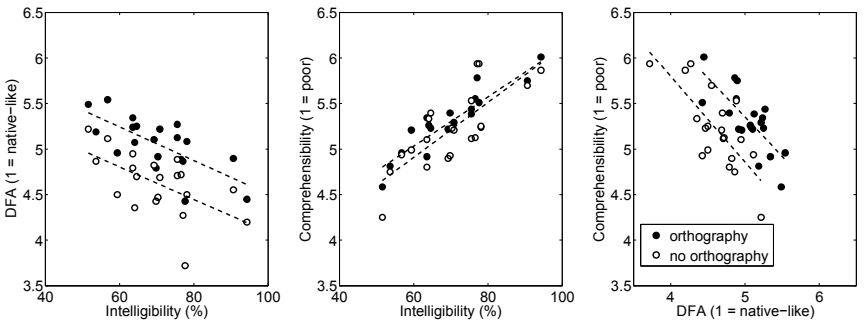

Figure 1. Scatter plots showing the relationship between intelligibility and DFA (left), intelligibility and comprehensibility (middle) and DFA and comprehensibility (right). Each point represents a single speaker and is the mean rating/word score measured over all words and judges. The best linear fits are shown for the orthographic and non-orthographic presentation conditions.

3.3. Interim discussion

A strong positive correlation was in evidence between comprehensibility and intelligibility, with a strong negative correlation between comprehensibility and DFA. The weakest (negative) correlation was between DFA and intelligibility (as in Munro and Derwing 1995a), indicating that a strong FA can be quite intelligible (Munro and Derwing 1995a, b) even though it is found to be harder to understand. Our findings validate the methodology used in the present paper in that words extracted from extemporaneous speech may be used in order to study the main FA correlates whilst avoiding possible biases introduced by higher order information present in longer utterances and task effects related to repetition or imitation techniques.

An additional methodological variable we included was the presence or absence of the orthographic form of the target during the judging of DFA and comprehensibility. We found a small effect of orthography in same direction as Levi et al. (2007) with talkers given harsher DFA ratings when the orthography was present. The presence of orthography also produced stronger correlations between intelligibility and both DFA and comprehensibility of NN speech, indicating that presenting the target when estimating DFA and comprehensibility can enhance the reliability of judgements.

Although we have established that isolated words elicited from extemporaneous speech provoke similar patterning of FA dimensions as those uncovered

in earlier studies, it is unclear whether providing more information to listeners would result in similar levels of DFA and comprehensibility. While isolated items permit a greater focus on the phonetic basis for accent, the task faced by listeners in making accent judgements on short-duration tokens may be challenging. Consequently, it is of interest to determine whether the use of longer speech samples leads to different judgements, and if so, at what point DFA and comprehensibility estimates reach an asymptote. Rather than use contiguous word sequences, which risks re-introducing higher-level (supra-word) carriers of FA, in Experiment 2 we employed random and unrelated sequences of words from the same talker.

4. Experiment 2: Judgements based on word sequences

Experiment 2 required listeners to provide DFA and comprehensibility judgements on both isolated words and sequences of 2, 4 or 8 words spoken by the same talker. The number of words in each sequence was motivated by the findings of Kato and Kakehi (1992) described earlier.

4.1. Methods

4.1.1. Participants

A new cohort of 16 listeners (6 male, 10 female) took part in Experiment 2. Participants were staff and students recruited using the University of Edinburgh's Student and Graduate Employment service. All were paid for their participation. All were born in the UK and had English as their native language. Two participants, both female, were excluded from the analysis: one participant had lived in Spain for 10 years, while another reported daily contact with Spanish speakers of English. Of the remaining 14, three had some knowledge of Spanish at a basic level but were retained. Ages for the experimental cohort ranged from 19–52 with a mean of 23.2 years.

4.1.2. Test items

Test items were based on the same set of 224 word stimuli used in Experiment 1. Words were presented in four distinct blocks: in isolation, and in sequences of

2, 4 and 8 words from the same talker. To enable comparisons between items presented alone and as constituents in multi-word sequences, all words contained in multi-word sequences also occurred individually. As a consequence, test blocks with longer sequences contained fewer test items. Specifically, listeners heard 28 × 8-word sequences (one from each speaker), 56 × 4-word sequences, 112 × 2-word sequences and 224 single words. Within a sequence, word order was randomised, although the same random order was used for each listener. For the multi-word sequences all words were equalised to have the same root mean square level prior to forming the longer sequence. Pauses of one-third of a second were inserted between each word in the sequence.

4.1.3. Procedure

The order of stimulus blocks was counterbalanced across listeners, so that all orders of presentation of the four different word blocks were equally distributed amongst listeners. Half of the listeners saw an orthographic representation of the word or word-sequence while making their judgement. Prior to the main experiment listeners underwent a short practice during which they heard 2 × 1-word, 2 × 2-word, 2 × 4-word and 1 × 8-word sequences. The main experiment required between 35 and 45 minutes. Participants were able to pause between blocks.

4.2. Results

Separate ANOVAs with sequence length as a within-subjects factor and orthographic presentation as a between-subjects factor were carried out on the mean accent and comprehensibility ratings. For DFA, neither orthography [$p = .42$] nor the interaction between sequence length and orthography [$p = .62$] were statistically-significant. However, accent ratings differed as a function of sequence length [$F(3,36) = 21.68$, $p < .001$]. A similar picture emerged for comprehensibility: neither orthography [$p = .18$] nor its interaction with sequence length [$p = .13$] were significant factors, but sequence length was [$F(3,36) = 18.1$, $p < .001$].

Figure 2 depicts DFA and comprehensibility ratings for word sequences. Since we found no effect of orthography, DFA and comprehensibility judgements from listeners who had access to the orthography are combined with those from listeners who used the auditory stimulus alone.

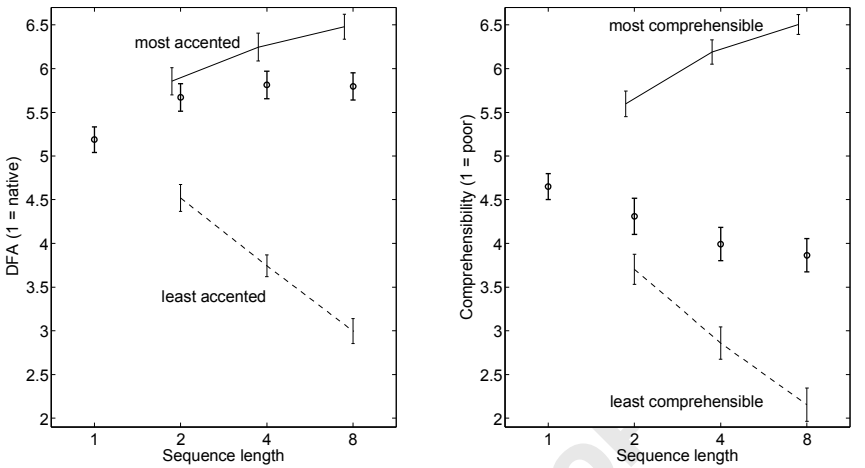

Figure 2. Mean DFA (left) and comprehensibility judgements as a function of the length of the word sequence. Also plotted are hypothetical values for DFA and comprehensibility if listeners were basing their judgements on the most (solid lines) or least (dashed lines) accented/comprehensible member of the word sequence. Errors bars indicate +/− 1 standard error.

Further analysis of sequence length effects for DFA based on a Fisher's Least Significant Difference of 0.18 indicated that sequences of length 1 differed from those of length 2, 4 or 8, which were all equivalent. For comprehensibility, a similar analysis (Fisher's LSD: 0.25) revealed that sequences of length 4 and 8 were equivalent, but that length 1 and length 2 differed from each other and from the remaining sequences.

Since the multi-word sequences were composed of individual words for which ratings were available from the same judges, it is possible to examine the relationship between judgements for sequences of words and judgements for the constituent words. For example, do judges rate properties such as degree of foreign accent based on the most (or least) accented item in the sequence or form a balanced assessment based on the entire sequence? To address this question, we replaced the rating for each multi-word sequence with either the maximum or minimum rating of its constituents, as measured in the single-word presentation condition. These quantities are plotted in Figure 2. For DFA, while listeners' ratings tend towards the most rather than least accented member of the word sequence, participants are clearly not basing their judgement on the maximally-accented item itself [$F(1,13) = 61.1$, $p < .001$], although judgements of 2-word

sequences are reasonably well-predicted by choosing the most accented item (difference of 0.18 re. Fisher's LSD of 0.13). For comprehensibility, listeners similarly tend towards using the least comprehensible member of the sequence but form a weighted judgement since ratings for the least comprehensible member always significantly underestimate actual comprehensibility judgements [$F(1,13) = 178$, $p < .001$].

4.3. Interim discussion

The second experiment investigated the effect of word sequence length on the robustness of DFA and comprehensibility measures. Following the results of Munro and Derwing (1995b), which indicated that order of presentation of the comprehensibility and DFA judgments did not result in significant differences, we did not include this as a variable. The design of our second experiment was inspired by previous research in listener to speaker adaptation by Kato and Kikehi (1988). In the current study, increases in sequence length affected the comprehensibility of NN speakers, with statistically significant decreases (i.e., poorer comprehensibility) for longer sequences, up to 4 words, from which point ratings stabilised. For DFA, increasing the number of words on which the judgement was based had a significant effect for NN speakers, where DFA was judged to be higher. The largest change was from single words to two-word sequences, after which DFA judgements remained constant. Unlike in the first experiment, here we found no effect of orthographic presentation of the stimulus on DFA judgements. We return to this issue in the general discussion.

Additionally, we investigated the possibility that listeners were judging sequences based on the most or least accented/comprehensible word in each sequence. This analysis demonstrates that judges do not simply use the most-accented nor least comprehensible word, but form a more nuanced judgement. For DFA in 2-word sequences there is some indication that the most-strongly accented word is chosen.

5. General discussion

This paper reports two experiments on FA evaluation. The main purpose of the first experiment was to ascertain if isolated words extracted from extemporaneous, naturally produced speech, could be reliably used to assess FA whilst avoiding the influences of higher order information on listeners' judgments. The

second experiment was meant to determine the number of words necessary for such evaluations to be reliable. The first experiment, based on individual words excised from natural speech, revealed correlations between the accent dimensions of intelligibility, comprehensibility and DFA in accordance with those found in previous studies of NN speech (Munro and Derwing 1995a, b). Comprehensibility and intelligibility were strongly correlated whereas comprehensibility was negatively correlated with DFA. As in previous studies we also found a negative correlation between intelligibility and DFA. These correlations indicate that, although strong FAs are found to be less easy to understand, they can still be quite intelligible, although less so than milder accents (Munro and Derwing 1995a, b). Our findings show that, for words elicited in and extracted from extemporaneous speech, the principal dimensions of FA pattern similarly to previous studies using other types of speech material. This is encouraging since, in order to study the phonetic basis for FA, it may be inadvisable to employ utterance-level speech samples produced by lower intermediate learners, which often display abundant grammatical errors as well as code-switching, both difficult to ignore in pronunciation assessment. Excising words from natural speech avoids this problem and at the same time sidesteps potential confounding effects that can arise from imitation or reading techniques with children.

The presentation of target words for evaluation with or without their orthographic form was included as a methodological variable. Being informed of the target ought to permit the disambiguation of certain pronunciations and avoid listeners evaluating against incorrect assumed targets. Experiment 1 did find that NNSs were given harsher DFA ratings when listeners were aware of the intended word, supporting Levi et al. (2007). This outcome agrees with the everyday experience of FL instructors in evaluating learners' productions: being able to see the orthographic representation of the word may highlight the production divergence from the native target and even uncover conflicts between the two word shapes which might not have been apparent had the written form been absent. However, this effect was not replicated in Experiment 2. It is possible that methodological differences between the two experiments are responsible for the discrepancy. In Experiment 2, listeners heard each item 4 times (once in each of the 1, 2, 4 and 8-word sequences), and it is plausible that repetition will have affected their awareness of the target even in the non-orthographically cued condition. Further studies are needed to explore in more depth the effect of target awareness on FA evaluations.

Experiment 2 investigated the number of words per stimulus that are required in order to obtain robust DFA and comprehensibility assessments, following work on listener to speaker adaptation by Kato and Kikehi (1984). These

authors found that listeners' adaptation to different speakers asymptoted at N=4 words, measured in terms of intelligibility improvements for contiguous words coming from the same speaker. Similarly, we found that listeners stabilised their evaluation of non-native speech comprehensibility after four words. For DFA the stabilisation point was found after two words, in agreement with Ikeno and Hansen (2006), who found that a minimum of two words were needed for accurate accent detection.

Speech perception has been linked to phonological short term memory (Jacquemot and Scott 2006). Although there are many and often conflicting views as to the capacity of short term memory, many studies in cognitive psychology typically postulate 4 as the number of chunks that are held by normal adults in their short term memory, albeit with considerable room for task and individual variability (Cowan 2001). Our results for comprehensibility tend to support this quantity. For DFA, the outcome that listeners' assessments reached a plateau at two words is consistent with studies which suggest that phonological working memory is smaller for anomalous items (Jacquemot and Scott 2006). This is also supported by the study of Dupoux et al. (2001) who found that the number of items that may be recalled in phonological short term memory is smaller when the stimuli do not conform to the phonological properties of the native language. Since accented words may possess phonologically-deviant patterns, our results might reflect a reduced phonological short term memory capacity.

As Munro and Derwing (2009) suggest, more classroom studies are needed in FA research. Indeed, we would argue that more research is needed in foreign-language instruction settings, since this reflects the majority experience in post-infancy language testing. Such studies present significant challenges. One of the problems derives from the age of the students, whose cognitive maturity has to be taken into account when designing elicitation tasks. Another concern is the wide range of competences found amongst young learners in foreign language settings in production skills for grammar, lexicon and fluency. Despite the fact that the approach to FA assessment presented in the current article has some disadvantages, such as the time required to excise words and the absence of several suprasegmental components for the evaluation, we believe its advantages outweigh the drawbacks. This method simplifies – both qualitatively, by excluding non-phonetic factors, and quantitatively, by efficient estimation based on short word sequences – the data-gathering process in school conditions, and further is suitable for studies across the spectrum of age and proficiency levels, and can also be used to assess the value of training programmes and test productive competence.

6. Conclusions

The present study was motivated by a need to find robust means for testing foreign language pronunciation proficiency without confounds from other linguistic levels which can interfere in the assessment of foreign accented speech, particularly at intermediate and lower levels of FL development. Our results suggest that the evaluation of isolated words excised from extemporaneous speech is a reliable means of language assessment, and also demonstrate that very short sequences of words are sufficient to reach stable judgements.

REFERENCES

Aoyama, K., S.G. Guion, J.E. Flege, T. Yamada and R. Akahane-Yamada. 2008. "The first years in an L2-speaking environment: A comparison of Japanese children and adults learning American English". *Int. Rev. Applied Linguistics* 46. 61–90.

Bassetti, B. 2009. "Orthographic input and second language phonology". In: Piske, T. and M. Young-Scholten (eds.), *Input Matters in SLA*. Bristol: Multilingual Matters. 191–206.

Bent, T. and A.R. Bradlow 2003. "The interlanguage speech intelligibility benefit". *Journal of the Acoustical Society of America* 114. 1600–1610.

Boersma, P. 2001. "Praat, a system for doing phonetics by computer". *Glot International* 5(9/10). 341–345.

Bradlow, A.R. and JA. Alexander. 2007. "Semantic and phonetic enhancement for speech-in-noise recognition by native and non-native listeners". *Journal of the Acoustical Society of America* 121(4). 2339–2349.

Bradlow, A.R. and T. Bent. 2008. "Perceptual adaptation to non-native speech". *Cognition* 106. 707–729.

Cargile, A.C. 1997. "Attitudes towards Chinese-accented speech. An investigation in two contexts". *Journal of Language and Social Psychology* 16. 434–443.

Cowan, N. 2000. "The magical number 4 in short-term memory: a reconsideration of mental storage capacity". *Behavioral and Brain Sciences* 24. 87–114.

Derwing, T.M. and M.J. Munro 1997. "Accent, intelligibility, and comprehensibility: Evidence from four L1s". *Studies in Second Language Acquisition* 20. 1–16.

Dupoux, E., S. Peperkamp and N. Sebastian-Gallés. 2001. "A robust method to study stress 'deafness'". *Journal of the Acoustical Society of America* 110. 1606–1618.

Flege, J.E. 1984. "The detection of French accent by American listeners". *Journal of the Acoustical Society of America* 76. 692–707.

Flege, J.E. and K. Fletcher. 1992. "Talker and listener effects on the perception of degree of foreign accent". *Journal of the Acoustical Society of America* 91. 370–389.

Flege, J.E., E. Frieda and T. Nozawa. 1997. "Amount of native-language (L1) use affects the pronunciation of an L2". *Journal of Phonetics* 25. 169–186.

Flege, J.E., I.R.A. MacKay and D. Meador. 1999. "Native Italian speakers' perception and production of English vowels". *Journal of the Acoustical Society of America* 106. 2973–2987.

Flege, J.E., M.J. Munro and I.R.A. Mackay. 1995. "Factors affecting strength of perceived foreign accent in a second language". *Journal of the Acoustical Society of America* 97. 3125–3134.

Gamer, M., J. Lemon, I. Fellows and P. Singh. 2012. "irr: Various coefficients of interrater reliability and agreement". R package available from http://www.r-project.org.

Gass, S. and E.M. Varonis. 1984. "The effect of familiarity on the comprehensibility of nonnative speech". *Language Learning* 34. 65–89.

Hamers, J.F. and M.H.A. Blanc. 2000. *Bilinguality and bilingualism.* (2nd ed.) Cambridge: Cambridge University Press.

Ikeno, A. and J.H.L. Hansen. 2006. "Perceptual recognition cues in native English accent variation: Listener accent, perceived accent and comprehension". *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '06)* 1. 401–404.

Imai, S., A. Walley and J.E. Flege. 2005. "Lexical frequency and neighborhood density effects on the recognition of native and Spanish accented words by native English and Spanish listeners". *Journal of the Acoustical Society of America* 117. 896–907.

Jacquemot, C. and K.S. Scott. 2006. "What is the relationship between phonological short-term memory and speech processing?" *Trends in Cognitive Sciences.* 10(11). 480–486.

Kakehi, K. 1992. "Adaptability to differences between talkers in Japanese monosyllabic perception". In: Tohkura, Y., E. Vatikiotis-Bateson and Y. Sagisaka (eds.), *Speech perception, speech production and linguistic structure*. Tokyo: OHM. 135–142.

Kato, K. and K. Kakehi. 1988. "Listener adaptability to individual speaker differences in monosyllabic speech perception". *Journal of the Acoustical Society of Japan* 44. 180–186.

Levi, S.V., S.J. Winters and D.B. Pisoni. 2007. "Speaker-independent factors affecting the perception of foreign accent in a second language". *Journal of the Acoustical Society of America* 121. 2327–2338.

Mayo, L.H., M. Florentine and S. Buus. 1997. "Age of second-language acquisition and perception of speech in noise". *Journal of Speech, Language and Hearing Research* 40. 686–693.

Munro, M.J. and T.M. Derwing. 1995a. "Foreign accent, comprehensibility and intelligibility in the speech of second language learners". *Language Learning* 49. 285–310.

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

Munro, M.J. and T.M. Derwing. 2009. "Putting accent in its place: Rethinking obstacles to communication". *Language Teaching* 42. 476–490.

Munro, M.J., T.M. Derwing and C.S. Burgess. 2003. "The detection of foreign accent in backwards speech". In: Sole, M.J., D. Recasens and J. Romero (eds.), *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS 15)*. Barcelona: Futurgraphic. 535–538.

Park, H. 2013. "Detecting foreign accent in monosyllables: The role of L1 phonotactics". *Journal of Phonetics* 41(2). 78–87.
Piske, T., J.E. Flege, I.R.A. MacKay and D. Meador. 2011. "Investigating native and non-native vowels produced in conversational speech". In: Wrembel, M., M. Kul and K. Dziubalska-Kołaczyk (eds.), *Achievements and perspectives in SLA of speech: New Sounds 2010* (vol. 2). Frankfurt am Main: Peter Lang. 195–205.
Piske, T., I.R.A. MacKay and J.E. Flege. 2001. "Factors affecting degree of foreign accent in an L2: A review". *Journal of Phonetics* 29. 191–215.
Shrout, P.E. and J.L. Fleiss. 1979. "Intraclass correlation: uses in assessing rater reliability". *Psychological Bulletin* 86. 420–428.
Sidaras, S. K, J.E.D. Alexander and L.C. Nygaard. 2009. "Perceptual learning of systematic variation in Spanish-accented speech". *Journal of the Acoustical Society of America* 125. 3306–3316.
Snow, C. E. and M. Hoefnagel-Höhle. 1978. "The critical period for language acquisition: Evidence from second language learning". *Child Development* 49. 1114–1128.
van Wijngaarden, S.J. 2001. "Intelligibility of native and non-native Dutch speech". *Speech Communication* 35. 103–113.
van Wijngaarden, S. J., H.J.M. Steeneken, and T. Hougast. 2002. "Quantifying the intelligibility of speech in noise for non-native listeners". *Journal of the Acoustical Society of America* 111. 1906–1916.
Varonis, E. and S. Gass. 1982. "The comprehensibility of non-native speech". *Studies in Second Language Acquisition* 4. 114–136.

Address correspondence to:
María Luisa García Lecumberri
Universidad del País Vasco
Vitoria-Gasteiz
Spain
garcia.lecumberri@ehu.es