Validation of a training method for L2 continuous-speech segmentation

Anne Cutler[1,2], *Janise Shanley*[1]

[1] MARCS Auditory Laboratories, University of Western Sydney, NSW 1797, Australia
[2] Max Planck Institute for Psycholinguistics, 6500 AH Nijmegen, The Netherlands
a.cutler@uws.edu.au/anne.cutler@mpi.nl, j.shanley@uws.edu.au

Abstract

Recognising continuous speech in a second language is often unexpectedly difficult, as the operation of segmenting speech is so attuned to native-language structure. We report the initial steps in development of a novel training method for second-language listening, focusing on speech segmentation and employing a task designed for studying this: word-spotting. Listeners detect real words in sequences consisting of a word plus a minimal context. The present validation study shows that learners from varying non-English backgrounds successfully perform a version of this task in English, and display appropriate sensitivity to structural factors that also affect segmentation by native English listeners.

Index Terms: continuous speech, L2, segmentation, training

1. Introduction

Listening to spoken language is normally one of the easiest tasks to perform. But although this is true of listening to speech in the native language (L1), listening in a second language (L2) can be noticeably hard. This can even produce the paradoxical situation where an L2 user may read written texts in the L2 with ease even where there is an orthographic mismatch, but may nonetheless have great difficulty following spoken lectures in the L2. What makes listening in L2 so hard?

Of course, there is no single answer. Phonetic confusions (*write* heard as *light*, etc.), vocabulary limitations that make input totally unfamiliar, misprocessing of syntax, pragmatics and discourse structure – all play a role. Training methods exist for some of these factors (e.g., improving phonetic discrimination, or increasing vocabulary). However, as psycholinguistic research summarized in this section shows, a significant part of the problem resides at the level of speech segmentation.

Spoken utterances are made up of sequences of words; the words are the meaningful units which are stored in memory. Speakers do not pause between the words that make up their utterances; utterances are effectively continuous. Segmentation of speech into its component words is therefore necessary for an utterance to be understood. Segmentation is not simply a matter of recognising words one after another as they come in, because words may contain other words embedded within them; this is an inevitable consequence of the construction of huge vocabularies from a phonetic repertoire of a few dozen speech sounds, and hence it is true of all languages.

Distributional and phonetic information correlated to word boundaries can help listeners segment speech. For instance, some phoneme sequences never occur within a syllable or word, so the occurrence of such a sequence must imply a boundary. Listeners more rapidly spot embedded words whose edges are aligned with such a boundary-correlated sequence (e.g., *rock* is recognised more easily in *foomrock* than in *foogrock* because *gr-* could be a word-initial sequence but *mr-* could not, so a boundary must fall before *rock* [12,10,20]).

Such sources of information are language-specific. It is a property of the English vocabulary that sequences such as [pf] or [zw] or [ml] cannot occur inside a syllable; but each of these three sequences is legitimately syllable-internal in some language ([pf] for instance in German: *Pferd, Kopf*). Other language-specific information is also used in segmentation – notably, rhythmic structure. In languages like English and Dutch, most words begin with stressed syllables, and listeners find it easier to segment speech at the onset of stressed syllables [5,19]. This can be clearly seen in segmentation errors, as when a pop song line *She's a must to avoid* is widely misperceived as *She's a muscular boy* – the strong syllable *void* is taken to be the onset of a new word, while the weak syllables *to* and *a-* are taken to be non-initial [2]. The rhythmic structures of other languages are as useful for segmentation as stress rhythm. Syllable-based rhythm in French and Korean is accompanied by syllabic segmentation in French and Korean listening experiments [13,7,9,8], while moraic rhythm in Japanese and Telugu likewise underlies moraic segmentation by Japanese and Telugu listeners [16,6,14]. Years of experience in an L1 induce automatic application of the procedures which work well in that L; but if listeners encounter an L2 with different structure to the L1, the use of the L1 procedures will obviously not be appropriate.

Thus segmentation is a necessary part of continuous-speech recognition, and is highly language-specific. Problems of segmentation could render listening to continuous L2 speech inordinately hard. The question then arises: could the operation of segmentation itself be subjected to targeted training, to improve this facet of L2 listening? This is the object of the present long-term project. The first step, the subject of the current report, is to locate a suitable methodology.

2. Testing speech segmentation

Psycholinguists have developed methods for studying speech perception in the laboratory and among these are methods focusing on speech segmentation. Some segmentation methods use nonsense input such as artificial languages [17] or require focus on non-lexical targets [13], and would thus not present the L2 listener with a readily generalisable training experience. There is however a method that measures detection of known words and separation of them from an adjacent speech context. This is the word-spotting task, introduced in pyscholinguistic research more than two decades ago [5], and in consequence the source of a now extensive body of research evidence. Indeed, much of the evidence described in section 1 is from this task (e.g., 7,10,12,19,20).

In a word-spotting experiment, listeners are presented with a sequence of nonsense utterances such as *obzel*, *foogrock*, *crithnish*, *bookving*. Some of these utterances may contain a real word. In this case, for instance, the English words *rock* and *book* can be found in *foogrock* and *bookving* respectively. When a word is spotted, the listener signals this by pressing a response key, and then repeats the word aloud.

The task addresses segmentation because separation of the known word from a nonsense context is the sole operation; no additional syntactic or semantic processing is involved. By comparing detection of the same word in different contexts, researchers can use the word-spotting task to determine what aspects of the phonetic context surrounding a word affect the ease with which the word (or, more precisely, the boundaries of the word) can be recognised in running speech. Thus the role of distributional correlates of boundaries can be revealed by comparing detection of words such as *rock* in contexts such as *foomrock* and *foogrock*, as described above. More on the range of effects observed with the task can be found in [11].

The task reflects natural listening processes in that, just as in normal listening, no prior information is available as to what the input will contain. The listener does not know what words will be presented or where they will occur. Training with this task will thus encourage listeners to learn the L2 patterns associated with word boundaries in continuous speech. Both the focus on segmentation and the high degree of ecological validity make word-spotting thus a good choice for addressing segmentation problems in L2. Adaptation of the task for L2 listening training is the aim of the present project.

3. Test Construction

3.1. Materials

Materials construction for L2 listening raises a multiplicity of issues where assumptions based on L1 listening cannot simply be transferred to the L2 case. These affect the choice of target words, of anticipated segmentation effects to be varied in the training materials, as well as other parameters that can be varied in the experimental situation.

3.1.1. Target words

The first issue to be addressed is the available L2 vocabulary for such a listening task. Obviously, L2 listeners do not command the extensive vocabulary that can be assumed for the L1 listeners who have participated in published studies of L1 segmentation using word-spotting. Learner vocabulary lists are, however, available, at least for English, in English-learner materials; moreover, they are classified in difficulty grades. In the present project, all potential target words are sorted into three difficulty levels, based on such existing gradings. This three-way classification allows for future variation in difficulty across graded sets of training materials.

Not specific to the L2 case is the important constraint that the construction of word-spotting materials always imposes: listeners must not be confronted with multiple possible targets. Consider, for example, the input *plainoyts*; this could be the word *play* followed by the nonsense context *noyts*, or the word *plain* followed by the nonsense context *oyts*. Such ambiguities must be avoided, and given the extent of embedding in the vocabulary, this requirement rules out a large number of words. In the present project, automatic procedures were applied to electronic dictionary resources, to ensure that such ambiguities did not arise among the selected target words. An initial set of over 500 words satisfying these joint constraints was compiled.

3.1.2. Segmentation effects

The word-spotting literature offers a wide variety of structural parameters that can potentially be varied in training materials. Some exploit features particular to certain languages [e.g., vowel harmony in Finnish: 18]; others involve segmentation effects that are putatively universal [e.g., the requirement that stand-alone words contain at least one vowel: 15, 4]. The most attractive segmentation effects for training purposes are those that are language-specific realisations of general effects, such as the distributional features of lexical items. Listeners in all languages exploit such distributional factors in segmenting speech; it can be guaranteed that L2 listeners thus possess, as a result of experience with their L1, the necessary skills to exploit the same type of information in L2. Distributional factors which can be learned for L2 include both phonotactic constraints on syllable structure (such as the non-occurrence of syllable-internal [pf] in English) and patterns of word structure (such as highly frequent versus infrequest word-initial and word-final patterns, typical stress placement, etc.).

For the present project we have initially chosen to manipulate two such factors: stress placement and syllable boundary transition difficulty. In English, stress falls predominantly on word-initial syllables [3]. A subset of the materials comprises stressed words (monosyllabic or initially stressed), preceded or followed by a monosyllabic context syllable containing a reduced vowel. Examples are *thousand* in *nelthousand*, *neck* in *treneck*, *signal* in *signaltev*, *bed* in *bedesh*; the vowel in each context syllable is schwa. A further subset of the materials contrasted easy versus difficult transitions between target word and context. An easy transition is one in which the word boundary is unambiguous: e.g., *well* in *wellrurp* or *vungwell*, because the sequences [lr] and [ngw] cannot occur within words. A difficult transition is one which is ambiguous, so that the speech sounds adjacent to each edge of the real word contained in the sequence could combine in another word with the speech sound at the word's edge: e.g., *well* in *welljuv* or *vutwell*, where the transitions potentially allow word-initial or word-final sequences (cf. *twitter*, *bulge*).

3.1.3. Further structural parameters

As the examples cited earlier make clear, the construction of word-spotting materials allows for variation in position; in *bookving* the target word precedes the appended context, while in *foogrock* the target follows the context. Medial embedding is also in principle possible (consider *book* in *zimbookving*). The positional parameter can be held constant (to make a particular set of materials easier), or varied (making it harder). For the current set of materials target words were placed both in initial and final position, but not in medial position.

Further parameters that can be varied include the number of speakers presenting the materials, the sex of the speakers, the rate of speech and the presentation rate (inter-item interval), as well as the language variety (e.g., for English: US, British, etc.). For the present project six speakers were recorded (three male, three female). All were speakers of standard educated Australian English, and all had phonetic training. They used a normal rate of speech. Recordings were made direct to disc, and each recorded item (word+context) was extracted, checked, and stored as a separate audio file. Thus more than 3000 audio files are already available for potential presentation in the project.

3.2. Procedure

In standard word-spotting with L1 listeners, responses are timed and both accuracy and reaction time are measured. This double response measure has been instrumental in making the method widely useful. For L2 training purposes, responses need not be rapid; accuracy is the principal aim, and hence in the present project no reaction times will be collected. In order to enable checking of accuracy, all responses in word-spotting are recorded and, if need be, analysed offline in detail.

4. Test Validation

4.1. Participants

Thirty-three advanced-level students of English for Academic Purposes at UWS College, Sydney, took part in the validation test. The L1 of 17 participants was Arabic (dialects of the Arabian peninsula), the L1 of the other 16 was Mandarin Chinese. All participants had normal hearing and their average length of residence in Australia was 20 weeks (for one-third of the participants less than one month).

4.2. Materials and Procedure

From the available audio files, 240 files were selected for use in the validation study. Half of these were spoken by one of the female speakers, the other half by one of the male speakers. All target words were checked by staff members at UWS College, who confirmed that participants should know the words in question. The majority of the words (75%) were monosyllabic (*ring*, *book*, *put* etc.), with the rest bisyllabic (*allow*, *noodle*) or, rarely, trisyllabic (*cigarette*, *medicine*).

The selected items included equal numbers of exemplars from the conditions arising from the stress and boundary difficulty factors built into the materials. In the phonology of English, stress placement and syllable boundary effects are not independent. Accordingly, the effects cannot be orthogonally manipulated in the materials we have constructed. The result is thus six conditions, in which position of the word is varied orthogonally with the three-way comparision of the condition Stressed, Easy Transition and Difficult Transition. The six conditions, in full, are (1) Stressed, Initial Position (e.g., *bed* in *bedesh*); (2) Stressed, Final Position (e.g., *neck* in *treneck*); (3) Easy Transition, Initial Position (e.g., *sock* in *socknal*); (4) Easy Transition, Final Position (e.g., *noodle* in *lemnoodle*); (5) Difficult Transition, Initial Position (e.g., *free* in *freeblom*); (6) Difficult Transition, Final Position (e.g., *cost* in *enscost*).

Four lists of 60 items each were constructed, with speaker and condition equally represented across lists and randomised within each list. This length corresponds to that in an average L1 word-spotting study. There were no filler items without target. At least seven participants heard each of the four lists.

Audio files were presented from a computer running DMDX experimental control software. Participants were tested individually; they were instructed to listen to each item and find the real English word it contained as quickly as they could, then to say that word aloud. The audio was presented over closed-ear Sennheiser headphones. After the first 10 items participants were given an opportunity to ask questions if necessary. Their spoken responses were recorded direct to disc and scored in situ by the experimenter. After completing the audio list, partipants filled in a questionnaire concerning the study, in which they chose descriptive terms from a range of possibilities concerning difficulty, length and value to them of the test materials. Each testing session (instructions, audio list, written questionnaire) took approximately 30 minutes.

4.3. Results

In each list, the initial 10 items were treated as practice and not scored. Every participant scored higher than 50% correct on the 50-item scored list. Although well below the expected L1 performance (always near 100% in word-spotting studies), this is very promising. Mean scores across the four lists and the two speakers did not significantly differ. The mean percent correct detection for the six conditions is shown in Figure 1.

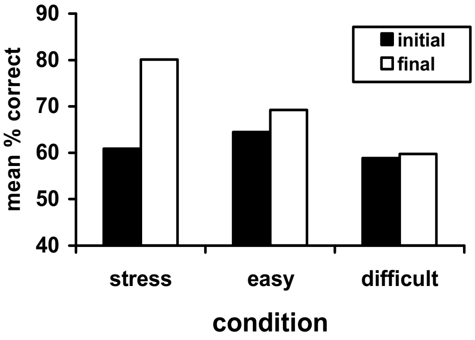

Figure 1: Percent correct word detection, averaged across participants, for the Stress, Easy Transition and Difficult Transition conditions, as a function of target word position in the nonsense item.

For participants with Arabic as L1, the mean percent correct was 65% and for participants with Chinese as L1 66%. This difference was not significant, and L1 group did not interact with any other factors in the analysis of the results. We thus report analyses collapsed over list, speaker and L1 group.

The effects of the structural factors manipulated in the materials were subjected to an analysis of variance. This revealed a significant overall effect of position: words were more accurately detected in final position than in initial position (F [1,32] = 5.52, p < .03). There was also a significant main effect of condition (F [2,64] = 11.24, p < .001). Crucially, however, these effects interacted, as can be clearly seen in the figure (F[2,64] = 6.26, p < .004). Further analyses therefore explored the components of this interaction. The effect of position was highly significant in the Stress condition (*treneck* vs. *bedesh*; t[32] = 4.66, p < .001) but insignificant in the two boundary conditions (t < 1 in both). A separate comparison of the two boundary conditions revealed a significant higher detection rate for words in the Easy Transition condition (*socknal*, *lemnoodle*) than for words in the Difficult Transition condition (*freeblom*, *enscost*); F [1,32] = 14.27, p < .001).

The results of the questionnaire indicated an overall positive evaluation by the participants. The most positive terms on offer ("interesting" and "helpful") were each chosen by over 72% of participants, the most negative terms ("boring", "tiring") by fewer than 7%. Participants were divided on whether the test was difficult (35%) or easy (32%), but no participant expressed any notably negative reaction to the experience.

4.4. Discussion

The results of the validation study have been highly informative for our purpose of assessing the suitability of the word-spotting task for use as a training method.

First, in every respect the task has proved itself amenable to use in L2 listening studies (training or assessment). In its standard form, with a large number of filler items in which no real word is detectable, the task is extremely difficult for L2 listeners, since the uncertainty level is unpleasantly high. The present version clearly does not have this negative effect. Note that the study that we have run was far from the easiest version of the task one could imagine. Not only is the response selection open (i.e., listeners have no pre-specified response set and no multiple-choice answer options are provided), but we also mixed different item types, and a male and a female speaker, in the audio input. On any trial, listeners could not know beforehand whether the target word was at the beginning or at the end of the input, they could not know what the word was or how long it was, and they could not know which voice would be speaking. Even under these conditions, our listeners managed to score extremely well, and even – on the evidence of their questionnaire responses – to enjoy the task. This finding suggests that, given the range of possible ways in which the task could be modified to render it easier or harder, the fundamental requirements for a useful training task have been met.

Second, we observed no difference in the pattern of results as a function of the native language of the participants. In the present study we did not incorporate particular features of the materials which should have been especially hard for either L1 group, since such differences have been well established in the literature and were not at issue here. Nonetheless, the two L1 backgrounds (Arabic, Mandarin) are very different, and vary significantly in the ways in which they contrast with English and in the problems that learning English presents. It is thus very gratifying that L1 background had no observable effect at all, either as a main effect or in interaction with any other factor, on the performance of the task.

Third, the results showed clearly that the structural factors we manipulated had the intended effects. Our listeners were sensitive to both the stress pattern of the items they heard, and the relative difficulty of the transition from one syllable to another. Moreover, their responses were influenced by these factors in the same way as responses by native English listeners had been affected in prior investigations. The condition with the highest probability of correct response was the one in which a stressed word was preceded by a reduced syllable (e.g., *treneck*, in which the word to be detected was *neck*); exactly this pattern of segmenting speech at the onset of a stressed syllable is an appropriate strategy for English [2,5]. Similarly, the use of clearly marked syllable boundaries such as in *lemnoodle* and *socknal* made detection rates in the Easy Transition condition higher than those in the Difficult Transition condition (*freeblom*, *enscost*); again, this is a pattern which has been repeatedly found in segmentation studies with listeners using their L1 [12,20]. It has also been shown that explicit training with such boundary patterns leads to improvement in L2 segmentation [1], and that high-performing L2 listeners make use of L2 boundary patterns in listening [20]. The present study shows that the word-spotting task can encourage use of such patterns if they are present in the input, and thus suggests again that the task lends itself well to use as a training procedure.

5. Conclusions

This study has provided proof of concept that the word-spotting task can be used with L2 listeners, that its use does not create an unpleasant or tedious listening situation, that listeners from widely different L1 backgrounds respond in much the same way, and that the segmentation results that they produce are orderly and affected in a predictable manner by structural factors built into the materials. Use of the task as a training method is crucially dependent on all the above criteria. The validation of the method having thus produced promising results, the project can now move to its main phase. In this phase long-term training with the task will be assessed as a method for inducing the learning of structural patterns associated with boundaries between words in an L2. The learning should result in an improvement in segmentation of continuous L2 speech.

6. Acknowledgements

This project was supported by UWS grant RGS 20211.80666 to the first author and J. Kim. We thank Johnson Chen for programming assistance, and we further thank the staff of UWS College, Westmead, NSW, for their co-operation with this research.

7. References

[1] Al-jasser, F., "The effect of teaching English phonotactics on the lexical segmentation of English as a foreign language", System, 36: 94-106, 2008.

[2] Cutler, A. and Butterfield, S., "Rhythmic cues to speech segmentation: Evidence from juncture misperception", J. Mem. Lang., 31: 218-236, 1992.

[3] Cutler, A. and Carter, D. M., "The predominance of strong initial syllables in the English vocabulary", Comput. Speech Lang., 2: 133-142, 1987.

[4] Cutler, A., Demuth, K. and McQueen, J. M., "Universality versus language-specificity in listening to running speech", Psychol. Sci., 13: 258-262, 2002.

[5] Cutler, A. and Norris, D. G., "The role of strong syllables in segmentation for lexical access", J. Exp. Psychol.: Hum. Perc. and Perf., 14: 113-121, 1988.

[6] Cutler, A. and Otake, T., "Mora or phoneme? Further evidence for language-specific listening", J. Mem. Lang., 33: 824-844, 1994.

[7] Dumay, N., Frauenfelder, U.H., and Content, A., "The role of the syllable in lexical segmentation in French: Word-spotting data.", Brain Lang., 81: 144-161, 2002.

[8] Kim, J., Davis, C. and Cutler, A., "Perceptual tests of rhythmic similarity: II. Syllable rhythm", Lang. Speech., 51: 343-359, 2008.

[9] Kolinksy, R., Morais, J. and Cluytens, M., "Intermediate representations in spoken word recognition: Evidence from word illusions", J. Mem. Lang., 34, 19-40, 1995.

[10] Lugt, A. van der., "The use of sequential probabilities in the segmentation of speech", Percep. Psychophys., 63: 811-823, 2001.

[11] McQueen, J. M., "Word spotting", Lang. Cognitive Proc., 11: 695-699, 1996.

[12] McQueen, J. M., "Segmentation of continuous speech using phonotactics", J. Mem. Lang., 39: 21-46, 1998.

[13] Mehler, J., Dommergues, J. Y., Frauenfelder, U. H. and Segui, J., "The syllable's role in speech segmentation", J. Verb. Learn. Verb. Behav., 20: 298-305, 1981.

[14] Murty, L., Otake, T. and Cutler, A., "Perceptual tests of rhythmic similarity: I. Mora rhythm", Lang. Speech, 50: 77-99, 2007.

[15] Norris, D., McQueen, J. M., Cutler, A., and Butterfield, S., "The possible-word constraint in the segmentation of continuous speech", Cognitive Psychol., 34: 191-243, 1997.

[16] Otake, T., Hatano, G., Cutler, A., and Mehler, J., "Mora or syllable? Speech segmentation in Japanese", J. Mem. Lang., 32: 258-278, 1993.

[17] Saffran, J. R., Newport, E. L., and Aslin, R. N., "Word segmentation: The role of distributional cues", J. Mem. Lang., 35: 606-621, 1996.

[18] Suomi, K., McQueen, J. M., and Cutler, A., "Vowel harmony and speech segmentation in Finnish", J. Mem. Lang., 36: 422-444, 1997.

[19] Vroomen, J., van Zon, M., and de Gelder, B., "Cues to speech segmentation: Evidence from juncture misperceptions and word spotting", Mem. Cognition, 24: 744-755, 1996.

[20] Weber, A. and Cutler, A., "First language phonotactics in second-language listening", J. Acoust. Soc. Am., 119: 597-607, 2006.