Consonant identification in noise by native and non-native listeners: Effects of local context

Anne Cutler[a]

Max Planck Institute for Psycholinguistics, Nijmegen 6500 AH The Netherlands and MARCS Auditory Laboratories, University of Western Sydney, Sydney 1797, Australia

Maria Luisa Garcia Lecumberri

Department of English Philology, University of the Basque Country, Vitoria, 01003, Spain

Martin Cooke

Department of Computer Science, University of Sheffield, Sheffield, S1 4DP, United Kingdom

Speech recognition in noise is harder in second (L2) than first languages (L1). This could be because noise disrupts speech processing more in L2 than L1, or because L1 listeners recover better though disruption is equivalent. Two similar prior studies produced discrepant results: Equivalent noise effects for L1 and L2 (Dutch) listeners, versus larger effects for L2 (Spanish) than L1. To explain this, the latter experiment was presented to listeners from the former population. Larger noise effects on consonant identification emerged for L2 (Dutch) than L1 listeners, suggesting that task factors rather than L2 population differences underlie the results discrepancy.
© 2008 Acoustical Society of America. [DOI: 10.1121/1.2946707]

I. INTRODUCTION

Users of a second language (L2) know well that listening to speech is more disrupted by background noise in the L2 than in the native language (L1). The differential disruption has also been solidly demonstrated under laboratory conditions (Garcia Lecumberri and Cooke, 2006; Mayo *et al.*, 1997; Nábělek and Donahue, 1984; Takata and Nábělek 1990). However, it is not yet fully clear how the phenomenon should be explained.

One class of explanations locates L1/L2 differences chiefly in phoneme identification. On this account, L2 listeners use other or fewer acoustic cues for phonemes than L1 listeners; if their attention were to be drawn to the cues used by L1 listeners, their speech perception would improve both in general (Jamieson and Morosan, 1986) and in noisy conditions in particular (Hazan and Simpson, 2000). Failure to exploit the full range of phonemic identity cues will render phoneme identification less secure than that of L1 listeners; if phonemes are inaccurately recognized, then in turn no word recognition is possible, and the whole speech perception process more or less falls apart.

Another class of explanations locates differences chiefly at higher processing levels. On such an account (e.g., Bradlow and Alexander, 2007), auditory acuity of L1 and L2 listeners may be equivalent, so that noise disrupts phoneme identification comparably for both L1 and L2 input, but the crucial difference lies in the degree to which recovery from disruption is possible. In the L1, extensive experience with the distributional patterns of phonemes, words, phrases, and constructions pays off in realistic expectations of the probabilities for replacement of missing or misperceived portions of the input. In the L2, insufficient experience has been accrued for realistic expectations to be rapidly derivable. Recovery from disruption is thus too slow for speech perception to be repaired.

Prompting the present study were two specific preceding investigations, which produced different patterns of results. In both cases, phoneme identification by native English-speaking listeners and by L2 listeners was contrasted, and in both cases, the materials were American English phonemes presented in minimal nonsense contexts. Thus both studies attempted to measure phonetic identification performance in the absence of support from lexical or other higher-level information. But in one case the effect of noise on the phoneme identification performance of L1 versus L2 listeners was parallel, while in the other case, noise affected the L2 performance far more than the L1 performance.

In the first study, by Cutler *et al.* (2004) the materials were American English vowels and consonants in CV and VC syllables; the L1 listeners were American English, and the L2 listeners were Dutch; the noise mask was six-talker babble presented at three signal-to-noise ratios (SNRs): 16 dB (very low noise), 8 dB (medium noise), and 0 dB (for L2 listeners, quite high noise). No evidence was found for differential effects of noise on L2 listening; the noise affected L1 and L2 listeners to the same extent, with L2 phoneme identification staying at about 80% of L1 performance across all noise levels.

In the second study, by Garcia Lecumberri and Cooke (2006) the materials were again American English, but only consonants were identified; these were presented in an unvarying a_a context. The L1 listeners were British English, the L2 listeners Spanish. The noise mask was a single competing talker, or speech-shaped noise, or eight-talker babble,

---

[a]Author to whom correspondence should be addressed. Electronic mail: anne.cutler@mpi.nl
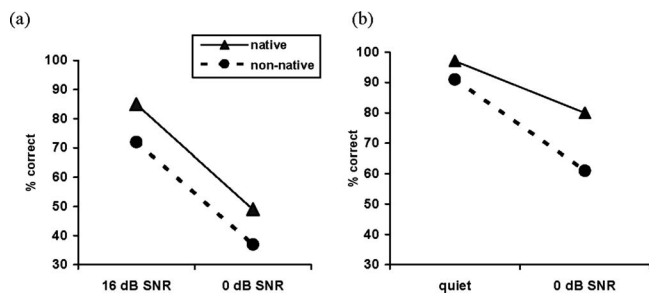
FIG. 1. Percentage of correctly recognized American English consonants in easy versus difficult listening conditions (a) for Dutch-L1 non-native listeners and American English native controls, at 16 dB and 0 dB SNR in six-talker babble; (b) for Spanish-L1 non-native listeners and British English native controls, in quiet and at 0 dB SNR in eight-talker babble.

always at 0 dB SNR, and contrasted with a quiet no-mask condition. In this case L2 performance was 92% of L1 performance without masking, but with the maskers, at 0 dB SNR, it varied from 90% (single competing talker) to 78% (multi-talker babble) – i.e., the noise affected the L2 listeners significantly more than the L1 listeners. Figure 1 contrasts the two studies' results, for consonant identification only, in the most similar conditions: quiet or low-noise versus 0 dB multi-talker babble.

Obviously, there are many differences between the two studies to which the contrast in the result patterns might be ascribed. First, one L1 group was British, the other American, while one L2 group was Spanish and the other was Dutch. The L1 difference is unlikely to be decisive, given that Australian English listeners matched the American performance with the Cutler et al. (2004) stimuli (Cutler et al. 2005). The L2 groups, however, differ in proficiency, as Garcia Lecumberri and Cooke (2006) pointed out; Dutch users of English have been repeatedly shown to perform at high standards in this L2, both in perception (Broersma, 2005, 2008; Cooper et al., 2002) and production (Bongaerts, 1999). This is partly because Dutch and English are closely related languages, with similar phonology, and partly due to the wide availability of English in Dutch daily life. The Spanish L2 group was less proficient and enjoyed less regular exposure to English outside formal learning. Greater proficiency could help to insulate L2 listeners from the impact of noise beyond that affecting L1 listeners.

Further, Dutch and Spanish differ greatly in the makeup of their phoneme repertoires. Dutch has (unusually among languages, especially those of Europe) a near-balanced repertoire of 19 consonants versus 16 vowels, while Spanish has a highly unbalanced repertoire, quite typical of the world's languages: 20 consonants but only five vowels. As lexico-statistical comparisons have demonstrated (Cutler et al. 2004, Cutler and Pasveer 2006), phoneme repertoire structure has far-reaching consequences for the composition of vocabularies and the similarity between words, and hence for the task of word recognition and all the steps involved in it.

Thus although vowels and consonants contribute similarly to word retrieval in Dutch and Spanish (Cutler et al., 2000), the difference in C:V ratio affects phoneme perception in context. In Dutch, effects of contextual (un)predict-

ability in phoneme detection are the same for vowel and consonant detection, but in Spanish, effects of consonant context on vowel detection are greater than effects of vowel context on consonant detection (Costa et al., 1998). Thus Spanish listeners recognize the greater potential for consonantal than for vocalic variation in their language, and this awareness directly affects their phonemic decision making. There are still other differences between these language groups in speech perception; for instance, the type of information used in consonant identification varies, with Spanish listeners paying greater attention to transitions in fricative identification than Dutch listeners do (Wagner et al., 2006). If a particular source of information is more susceptible to masking, then such inter-group differences could also surface as differences in identification success for the affected phonemes in noisy listening conditions. In short, L2 population differences must be reckoned a likely source of the varying results patterns of the Cutler et al. (2004) and Garcia Lecumberri and Cooke (2006) studies.

However, differences in the experiments could also have played a role. Even the most comparable conditions, shown in Fig. 1, were not identical across the studies, although the babble talker Ns used fell in the same performance range (Simpson and Cooke, 2005), and Garcia Lecumberri and Cooke (2006) argue that the range of masking effectiveness in the two studies was similar. Cutler et al. (2004) participants identified both vowels and consonants, but these identification tasks were performed in separate experimental blocks, and as Fig. 1 shows, the consonant results alone showed the clear inter-experiment difference. The greatest difference was perhaps in the type of stimuli used. The syllables of Cutler et al. (2004) were centrally embedded in noise, and since the syllables differed in duration, the precise amount of leading noise varied across the 645 tokens. The syllables were also made from 24 consonants and 15 vowels, so that inter-token variability was high. The Garcia Lecumberri and Cooke (2006) stimuli, in contrast, had a constant leading noise and a constant preceding and following vocalic context for the target consonants.

To test the source of the different results, we presented the Garcia Lecumberri and Cooke (2006) materials to Dutch listeners as tested by Cutler et al. (2004). We chose the quiet and multi-talker babble conditions, which showed, respectively, the smallest and largest differences between the Garcia Lecumberri and Cooke (2006) listener groups. If the discrepancy in the Cutler et al. (2004) versus Garcia Lecumberri and Cooke (2006) results was due mainly to the L2 listener groups, then we will here replicate the Cutler et al. (2004) result (see left panel of Fig. 1): compared to GLC's L1 control group, the listeners will maintain a roughly constant performance decrement across the two conditions. If the discrepancy is due mainly to task differences, then we will here replicate the GLC result (right panel of Fig. 1): the listeners will differ more from the L1 control group in the babble condition than in quiet.

II. METHOD

A. Participants

Sixteen students at the Radboud University Nijmegen (three male; mean age 21 years) took part in return for a small payment. All were native speakers of Dutch.

B. Materials

Two of the five Garcia Lecumberri and Cooke (2006) conditions were presented to the Dutch subjects. Speech tokens were American English VCVs from Shannon et al. (1999) corpus. The vowel preceding and following the consonant for identification was always /ɑ/; the consonant was one of 16 possibilities /p b t d k g m n l r f v s z ʃ ʃ/. Two tokens of each VCV from each of five male talkers were used, for a total test set of 160 items. An additional two examples of each VCV formed an initial set of (unscored) practice items. In one condition the speech tokens were presented without mask. In the other condition the speech was masked with eight-talker babble produced by summing utterances by male talkers from the TIMIT corpus. The babble started 1 s before the VCV and ceased at VCV offset. The SNR of the speech in babble was 0 dB. For further details of these materials see Garcia Lecumberri and Cooke (2006).

C. Procedure

Stimulus presentation and response collection were controlled by a computer running MATLAB. Participants were told that the test involved identification of English consonants; they responded on each trial by clicking on one cell of a grid of English words or phrases representing the 16 response options (e.g., "a Path," "aLarm," etc.). Subjects were tested individually, in a sound-attenuated room; the materials were presented binaurally over Sennheiser HD 497 headphones, and the quiet condition was always presented first.

III. RESULTS

The mean percent correct responses in quiet were 93.63%, in babble noise 58.78%. Thus in this experiment the Dutch listeners performed very much worse with the noise-masked stimuli than they did in the quiet condition. The comparable results for British English L1 listeners in Garcia Lecumberri and Cooke (2006) were 98.31% and 80.35%, respectively; their Spanish L2 listeners averaged 90.95% and 62.38%. As can be clearly seen in Fig. 2, the Dutch group's performance resembles that of Garcia Lecumberri and Cooke's (2006) L2 group, not that of their L1 group.

Analyses of variance across subjects (F1) and across items (F2), comparing the performance of the present Dutch group with that of the two groups tested by Garcia Lecumberri and Cooke (2006) on the present subset of the items only, revealed a significant main effect of language (F1 and F2 $p < 0.001$), and a main effect of testing condition (F1 and F2 $p < 0.001$).

The major difference between the results of Cutler et al. (2004) and Garcia Lecumberri and Cooke (2006) was that in the data of Cutler et al. (2004) no interaction was observed between the language factor and the presentation
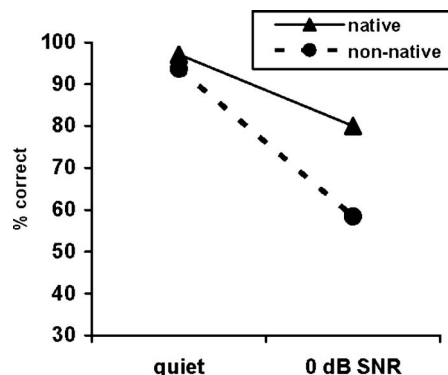

FIG. 2. Percentage of correctly recognized American English consonants for Dutch-L1 non-native listeners and British English native controls, in quiet and at 0 dB SNR in eight-talker babble.

condition (p. 3670), while in the data of Garcia Lecumberri and Cooke (2006) these effects interacted significantly (p. 2449). In the present data this interaction was significant for a comparison of the Dutch and the English groups (F1 $[1,35]=153.77$, $p < 0.001$, F2 $[1,159]=76.08$, $p < 0.001$).

A comparison of the Dutch and Spanish L2 groups alone revealed no significant overall difference in their performance (F1 and F2 < 1); however, the small advantage (2.7 percentage points) for the Dutch listeners in quiet was significant (F1 $[1,73]=5.27$, $p < 0.05$; $t2$ $[159]=2.93$, $p < 0.005$), as was also the 3.6 percentage points advantage for the Spanish listeners in babble noise (F1 $[1,73]=4.72$, $p < 0.05$; $t2$ $[159]=2.16$, $p < 0.05$).

To assess performance across the experiment, we compared first and second halves of each condition. Both in quiet and in babble, performance of each group improved from first to second half, and though the gain was quite small, it was significant across all three groups together ($p < 0.025$). Importantly, though, neither for the whole experiment nor for either condition was there any interaction of this inter-half difference with listener group.

IV. DISCUSSION

With the materials from the study of Garcia Lecumberri and Cooke (2006), Dutch L2 listeners from the Cutler et al. (2004) population produced performance which looked like Garcia Lecumberri and Cooke (2006) L2 results, not like Cutler et al. (2004) L2 results. The L2-L1 difference in the babble noise condition was much greater than the difference in quiet. This is the same asymmetry found by Garcia Lecumberri and Cooke (2006), and in other studies with more word- or sentence-like materials (e.g., Mayo et al., 1997; Nábělek and Donahue, 1984). Cutler et al. (2004), in contrast, observed similar L1-L2 differences in difficult and easy listening conditions.

Thus the difference in results pattern in the preceding studies does not seem to stem from the difference in listener population. It is not the case that Dutch listeners to English are simply able to resist the well-known L2 disadvantage in noise. Instead, the Garcia Lecumberri and Cooke (2006) task has produced analogous results with Spanish and with Dutch

Cutler et al.: Local-context effects in consonant identification

listeners. This outcome points, as we forecast above, to task differences as the source of the difference in results in the preceding studies. The task of Garcia Lecumberri and Cooke (2006) (like those of most other researchers) produces an interaction between listener group (L1, L2) and listening conditions (easy, difficult), while the task of Cutler et al. (2004) does not. The difference seems at least as much due to poorer performance by Cutler et al.'s (2004) L1 listeners as to resistance of their L2 listeners to effects of noise (compare Figs. 1 and 2). The results overall thus suggest that while masking effects for L1 and L2 speech can be equivalent, L1 listeners are definitely better at recovering from disruption. They use even the slightest low-level cues provided by a particular experimental context.

The aim of Cutler et al. (2004) was an experiment offering no contextual support at all for phoneme identification. In particular, the constant duration of the leading noise and the constant preceding vowel context in the present materials were not available in the materials of Cutler et al. (2004). We suggest that these local sources of contextual predictability can, in the absence of other contextual support, be exploited in phoneme identification tasks, at least by proficient (L1) listeners. From other research it is clear that language experience affects use of contextual information to compensate for the effects of noise masking. Mayo et al. (1997) presented high- and low-predictability sentences from the Speech Perception In Noise (SPIN) test to monolinguals, early bilinguals, and late learners of English as L2, and observed a significantly greater predictability benefit for the two former groups than for the last. Van Wijngaarden et al. (2002) found that L2 users' scores in a letter-guessing task on written text predict their speech recognition scores in noise, suggesting that individual differences in ability to exploit contextual redundancy affect relative resistance to noise masking.

Here, no higher-level (lexical or sentential) context was available to assist listeners in the consonant identification task. However, it seems that even low-level predictability, of the sort provided by a constant vocalic context and a constant duration of leading noise, can aid those listeners better able to use it. Especially in a difficult listening task such as consonant identification against a high level of babble noise, with no assistance available from any type of higher-level information, cues, of any kind, which reduce the uncertainty in the task will be very valuable. Constant vocalic contexts simplify consonant judgments in a variety of decision tasks (Costa et al., 1998; Swinney and Prather, 1980). Leading noise could make the speech onset temporally predictable. Intelligibility of plosive-vowel tokens improves with continuous noise (Ainsworth and Meyer, 1994) or leading noise (Ainsworth and Cervera, 2001) compared to co-gated noise; and predictability leads to more accurate stimulus discrimination in a visual gap detection task (Rolke and Hofmann, 2007) and in an auditory pitch discrimination task (Bausenhart et al., 2007).

It is plausible that both for this and for the role of vocalic context, beneficial effects will be larger for L1 listeners and detrimental effects larger for L2 listeners. Predictable onsets and predictable vocalic environment will both be bet-

ter exploited by listeners with greater accrued experience of the phonemic processing task in question. We assume that total accrued amount of experience with a given language translates, ceteris paribus, to ability to exploit redundancy at all levels of speech processing. Although it might be reasonable to expect performance ceilings where differences between higher-ability groups disappear, in fact the literature concerning speech in noise suggests that effects of total language experience are still observable at high performance rates. Thus bilinguals will in general have accrued less experience with either of their languages separately than monolinguals will have accrued with the one language; and although Mayo et al.'s early bilinguals performed much better in noise than did their late learners, the early bilinguals' performance did not match that of monolinguals. Similarly, Rogers et al. (2006) presented monosyllabic words, without higher-level context, to monolinguals and to early bilinguals with no perceptible foreign accent in English; they found that the two groups performed equivalently in quiet, but in noise the scores of these bilinguals were much lower.

Finally, note that the full story of speech identification in noise must even allow for differing effects of relevant experience from the L1 in L2 listening. In a separate study (Cutler et al., 2007) we compared the recognition of American English consonants by Dutch and Spanish and British English listeners, using different maskers than were used here, and a larger set of consonants. Crucially, the consonants included affricates, and /θ/, none of which were used in the present materials. The vowel context was constant, but the speech and noise were co-gated, and there was thus no leading noise to serve as a prior cue to the moment of stimulus onset (note, though, that there was predictability in that stimulus presentation was triggered by the participant's keypress). With most consonants, it was again the case that the performance of Dutch and Spanish listeners was parallel: both were more seriously affected by noise than L1 listeners were. However, a different pattern appeared with the fricative/affricate subset of the materials; there Dutch listeners were less seriously affected by noise than either L1 listeners or the Spanish group. This result was interpretable in the light of the cross-linguistic differences reported by Wagner et al. (2006); listeners attend to transitional cues for fricatives when their native phoneme inventory contains confusable pairs of fricatives. This is true of English and Spanish (both these languages contrast /f/ and /θ/) but not of Dutch. Gating studies (Wagner, 2008) showed that differential cross-language sensitivity to transitional cues does not generalize across phoneme classes (e.g., to stops). If the presence of a noise mask seriously disrupted the use of transition information, then the paradox of the better performance of the Dutch group with these consonants would be explained: in their case, their native experience with fricative identification, not relying on the cues which were fragile under noise, served them better than the other two groups' experience which required attention to the cues which had been disrupted.

Ainsworth, W. A., and Meyer, G. F. (**1994**). "Recognition of plosive syllables in noise: Comparison of an auditory model with human performance," J. Acoust. Soc. Am. **96**, 687–694.

Ainsworth, W. A., and Cervera, T. (**2001**). "Effects of noise adaptation on the perception of voiced plosives in isolated syllables," *Proc. Eurospeech 2001*, Aalborg, Denmark, pp. 371–374.

Bausenhart, K. M., Rolke, B., and Ulrich, R. (**2007**). "Knowing when to hear aids what to hear," Q. J. Exp. Psychol. **60**, 1610–1615.

Bongaerts, T. (**1999**). "Ultimate attainment in L2 pronunciation: The case of very advanced late L2 learners," in *Second Language Acquisition and the Critical Period Hypothesis*, edited by D. Birdsong (Erlbaum, Mahwah, NJ), pp. 133–159.

Bradlow, A. R., and Alexander, J. A. (**2007**). "Semantic-contextual and acoustic-phonetic enhancements for English sentence-in-noise recognition by native and non-native listeners," J. Acoust. Soc. Am. **121**, 2339–2349.

Broersma, M. (**2005**). "Perception of familiar contrasts in unfamiliar positions," J. Acoust. Soc. Am. **117**, 3890–3901.

Broersma, M. (**2008**). "Nonnative listeners use vowel duration less than English listeners for English final v-f," J. Acoust. Soc. Am. **123**, 3183.

Cooper, N., Cutler, A., and Wales, R. (**2002**). "Constraints of lexical stress on lexical access in English: Evidence from native and non-native listeners," Lang Speech **45**, 207–228.

Costa, A., Cutler, A., and Sebastián-Gallés, N. (**1998**). "Effects of phoneme repertoire on phoneme decision," Percept. Psychophys. **60**, 1022–1031.

Cutler, A., Cooke, M., Garcia Lecumberri, M. L., and Pasveer, D. (**2007**). "L2 consonant identification in noise: Cross-language comparisons," in *Proc. INTERSPEECH 2007*, Antwerp, pp. 1585–1588.

Cutler, A., Norris, D., and Sebastián-Gallés, N. (**2004**). "Phonemic repertoire and similarity within the vocabulary," in *Proc. 8th International Conference on Spoken Language Processing*, Jeju, Korea, Vol. **1**, pp. 65–68.

Cutler, A., and Pasveer, D. (**2006**). "Explaining cross-linguistic differences in effects of lexical stress on spoken-word recognition," in *Proc. 3rd International Conference on Speech Prosody*, edited by R. Hoffmann and H. Mixdorff (TUDpress, Dresden), pp. 237–240.

Cutler, A., Sebastián-Gallés, N., Soler-Vilageliu, O., and Ooijen, B. van (**2000**). "Constraints of vowels and consonants on lexical selection: Cross-linguistic comparisons," Mem. Cognit. **28**, 746–755.

Cutler, A., Smits, R., and Cooper, N. (**2005**). "Vowel perception: Effects of non-native language versus non-native dialect," Speech Commun. **47**, 32–42.

Cutler, A., Weber, A., Smits, R., and Cooper, N. (**2004**). "Patterns of English phoneme confusions by native and non-native listeners," J. Acoust. Soc. Am. **116**, 3668–3678.

Garcia Lecumberri, M. L., and Cooke, M. P. (**2006**). "Effect of masker type on native and non-native consonant perception in noise," J. Acoust. Soc. Am. **119**, 2445–2454.

Hazan, V., and Simpson, A. (**2000**). "The effect of cue-enhancement on consonant intelligibility in noise: Speaker and listener effects," Lang Speech **43**, 273–294.

Jamieson, D. G., and Morosan, D. E. (**1986**). "Training non-native speech contrasts in adults: Acquisition of the English /ð/-/θ/ contrast by francophones," Percept. Psychophys. **40**, 205–215.

Mayo, L. H., Florentine, M., and Buus, S. (**1997**). "Age of second-language acquisition and perception of speech in noise," J. Speech Lang. Hear. Res. **40**, 686–93.

Nábělek, A. K., and Donahue, A. M. (**1984**). "Perception of consonants in reverberation by native and non-native listeners," J. Acoust. Soc. Am. **75**, 632–634.

Rogers, C. L., Lister, J. J., Febo, D. M., Besing, J. M., and Abrams, H. B. (**2006**). "Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing," Appl. Psycholinguist. **27**, 465–485.

Rolke, B., and Hofmann, P. (**2007**). "Temporal uncertainty degrades perceptual processing," Psychon. Bull. Rev. **14**, 522–526.

Shannon, R. V., Jensvold, A., Padilla, M., Robert, M. E., and Wang, X. (**1999**). "Consonant recordings for speech testing," J. Acoust. Soc. Am. **106**, L71–L74.

Simpson, S. A., and Cooke, M. P. (**2005**). "Consonant identification in N-talker babble is a nonmonotonic function of N," J. Acoust. Soc. Am. **118**, 2775–2778.

Swinney, D. A., and Prather, P. (**1980**). "Phonemic identification in a phoneme monitoring experiment: The variable role of uncertainty about vowel contexts," Percept. Psychophys. **27**, 104–110.

Takata, Y., and Nábělek, A. K. (**1990**). "English consonant recognition in noise and in reverberation by Japanese and American listeners," J. Acoust. Soc. Am. **88**, 663–666.

Van Wijngaarden, S., Steeneken, H., and Houtgast, T. (**2002**). "Quantifying the intelligibility of speech in noise for non-native listeners," J. Acoust. Soc. Am. **111**, 1906–1916.

Wagner, A. (**2008**). "Phoneme inventories and patterns of speech sound perception," Ph.D. thesis, University of Nijmegen.

Wagner, A., Ernestus, M., and Cutler, A. (**2006**). "Formant transitions in fricative identification: The role of native fricative inventory," J. Acoust. Soc. Am. **120**, 2267–2277.