On cross-language consonant identification in second language noise (L)

Letizia Marchegiani[a]
Language and Speech Laboratory, Faculty of Art, University of the Basque Country, 01006 Vitoria-Gasteiz, Spain

Xenofon Fafoutis[b]
Department of Applied Mathematics and Computer Science, Technical University of Denmark, 2800 Kongens Lyngby, Denmark

(Received 24 February 2015; revised 16 July 2015; accepted 1 September 2015; published online 16 October 2015)

Speech perception in everyday conditions is highly affected by the presence of noise of a different nature. The presence of overlapping speakers is considered an especially challenging scenario, as it introduces both energetic and informational masking. The efficacy of the masking also depends on the familiarity with the language of both the target and masking stimuli. This work analyses consonant identification by non-native English speakers in N-talker natural babble noise and babble-modulated noise, by varying the number of talkers in the babble. In particular, only English consonants that are also present in all the native languages of the subjects are used. As the subjects are familiar with the consonants used, this study can be considered a step towards a deeper analysis on perception of first language speech in the presence of second language maskers.

© *2015 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4930955]

[MAH] Pages: 2206–2209

I. INTRODUCTION

In everyday life, a massive amount of acoustic stimuli continuously reaches us. Cognitive mechanisms operate on these stimuli as a filter, selecting the most relevant ones and helping the brain to discard the others. O'Sullivan *et al.*[1] recently proved the existence of a correlation between attended speech representation, obtained through electroencephalography (EEG), and the subjects' performance at isolating the target stimulus. This filtering procedure has, as a fundamental prerequisite, the capability of isolating specific signals from the mixture. This capability depends greatly on the nature and the features of the signals involved (see the works of Cherry[2] and Bregman[3] for a general introduction on the topic). Several studies focused on distribution of attention, speech perception, and intelligibility in the presence of other concurrent sounds and, in particular, simultaneous talkers. Drullman *et al.*[4] proved that the separation between two different speakers becomes harder when the voices are of the same gender. Moore *et al.*[5] illustrated the effect of the difference in the fundamental frequency, phase spectrum, intensity, and spatial proximity between the two sound sources on segregation ability. The number of masking talkers also plays an important role on the intelligibility of a target speech. For a review considering up to nine overlapping talkers, see the work of Bronkhorst.[6] Simpson *et al.*[7] explored consonant identification in N-talker natural babble, babble-modulated noise, and speech-shaped noise, proving that intelligibility is a nonmonotonic function of N. Other investigations illustrated the effect of competing speech in different languages, using up to six talkers. Van Engen *et al.*[8] showed that, for native English speakers, as long as the words in the competing babble are detectable, English speech is a more effective masker than speech in unknown languages. Cooke *et al.*[9] analysed the impact of overlapping speech-shaped noise and natural English speech on the identification of English consonants, using subjects with eight different native languages. In this study, non-native English speakers were shown to have more difficulty than native ones performing the task. The authors explain this behaviour by suggesting that even if the confusion decreases when the language of the interference is different than the native language of the listeners, in terms of lexical masking and cognitive load (as was also proved by Hoen *et al.*[10] and Mattys *et al.*[11]), stream segregation still requires more effort due to the lower competence in the language, and this makes the task harder. Consonant identification for native and non-native speakers in the cases of competing talkers, stationary noise, and eight-talker babble was also investigated by Lecumberri and Cooke.[12]

Starting from these analyses, this work extends the investigations of Cooke *et al.*,[9] Van Engen *et al.*,[8] and Lecumberri *et al.*[12] to a wider range of N and, more generally, masking conditions. With this purpose, this paper presents listening experiments on consonant identification, based on the paradigm in Simpson *et al.*[7] Compared to the work of Simpson *et al.*,[7] though, in which the native language of the subjects is not specified, all of our listeners have English as a second language (L2). The choice of investigating identification of consonants aims at avoiding lexical and semantic cues that influence the comprehension

[a] Current address: Department of Engineering Science of the University of Oxford, 23 Banbury Road, Oxford OX2 6NN, UK. Electronic mail: letizia.marchegiani@eng.ox.ac.uk

[b] Current address: Department of Electrical and Electronic Engineering of the University of Bristol, Woodland Road, Bristol BS8 1UB, UK.

of the target stimuli and to, rather, highlight phonetic perception. Another important difference between this work and the one of Simpson *et al.*[7] and previous experiments on English consonant identification by non-native English listeners (such as, for example, Cooke *et al.*[9]) lies in the choice of the consonants used. To avoid any confusion with the identification due to English consonants which do not correspond to any sound in the native language of the subjects, only a subset of English consonants that are present and have the same acoustical features in all the languages of our subjects is considered. The aim is to explore how intelligibility changes with the variation of the number of overlapping talkers, both in the case of natural English babble and modulated babble, and if the use of consonants which are familiar to the listeners has any effect on their performance compared to what was shown in previous investigations. From this perspective, this study can be considered a step towards a deeper investigation of first language (L1) consonant identification and, more generally, speech perception, in the presence of maskers in L2.

II. EXPERIMENTS

The consonant tokens are presented in the form *VCV*, where *V* is the vowel /ɑ/ and *C* is one of the ten consonants used: /b/, /d/, /g/, /p/, /t/, /k/, /m/, /n/, /f/, /v/. The speech stimuli for the consonants were taken from Loizou.[13] The consonants are in American English, but no particular language influence on their pronunciation can be observed. The babbles were generated by combining segments of sentences from the following datasets: CNC_words,[14] CST,[15] IEEE_corpus,[16] NOIZEUS,[17] ELSDR,[18] and AUDITEC.[19]

For each of the ten N-talker masking conditions ($N \in \{1, 2, 3, 4, 8, 16, 32, 64, 128, 512\}$) and for each consonant, five different consonant identification tests are randomly generated. The order in which the consonants are presented in the tests is random. The available speakers are 34. When the babble is composed of up to 32 speakers, none of the speakers is used more than once. In the case of more than 32 speakers, the same speaker is not used more than i times, unless all the others have also been used i times. The root mean square (RMS) energy of all the utterances was normalised before being combined in each test, so that they contribute equally to the babble. The target-to-masker ratio [i.e., signal-to-noise ratio (SNR)] of any single test is -12 dB and all the tests are normalised to have the same RMS energy. Some pilots to empirically choose the most appropriate value of the SNR have been performed, to make sure that the task was neither too easy nor too hard.

From each of the babble segments previously generated, a babble-modulated segment has been created and mixed with the relative consonants. The correspondence between the consonants and the masking fragments and the order of the consonant tokens in the tests are the same as in the tests with natural babble for comparison purposes. Following Simpson *et al.*,[7] the speech-shaped noise is created by processing white noise with a filter, with a magnitude response equal to the long-term magnitude spectrum of a set of sentences from the ones used to build the natural babble (two random sentences, when available, for each of the 34 speakers). The N-talker babble-modulated noise tokens are, then, obtained by multiplying the speech shaped noise by the envelope of the relative N-talker natural babble tokens. The envelope derives from the convolution of the babble waveform with a 7.2 ms rectangular window.[20] In all cases, the consonant tokens and the respective overlapping fragments have the same duration and were gated.

The experiments have been carried out in an acoustically isolated booth. Both consonants and sentences used to create the babble have been resampled and presented at 44 100 Hz. A computer located outside the booth managed the presentation of the stimuli and the result collection. The tests were presented diotically using *Sennheiser HD580* headphones previously calibrated. The listeners executed the task using a MathWorks MATLAB interface containing the list of tests to execute and a set of buttons correspondent to the consonants. Each subject started with a training test, characterised by speech samples from different masking conditions for a total of 22 stimuli. All the tests were randomly ordered in the interface, to avoid any effect due to a specific choice of this order.

Ten listeners (seven males and three females) with no reported hearing or attentional impairment took part in the experiments. In particular, the set of listeners includes three native Italian speakers, three native German speakers, three native Danish speakers, and one native Greek speaker. All the above mentioned languages, in fact, contain the sounds of the consonants considered in this work. In the case of Italian and Greek subjects, the interface matched the sound with the relative symbol in the language, in order to avoid any confusion for the listener. For example, with regards to Italian, the letter *C* replaced the letter *K* in the interface presented to the Italian subjects. All the subjects performed a pretest to recognise the consonants in clean conditions (without any overlapping masker) to be sure they correctly understood the task and were able to identify the stimuli. The identification rate for each subject was above 98%.

III. RESULTS AND DISCUSSION

Figure 1 shows the relative identification rate averaged over the ten subjects (error bars correspond to 90% confidence intervals). A repeated measure analysis of variance (ANOVA) test confirmed the statistical significance of the results, revealing a significant effect of both the number of talkers N ($p < 0.001$) and the kind of masker ($p < 0.003$). A further repeated measures ANOVA test (with 95% Bonferroni correction) was also carried out to investigate the significance of the differences between the results obtained in the case of natural babble and babble modulated noise. Significant differences are observed in the conditions $N = 2$ ($p = 0.001$) and for $4 \leq N \leq 128$ ($p < 0.0001$). The cases $N \in \{1, 3\}$ and $N = 512$ are statistically equivalent. Overall, the general trend of the natural babble looks different from the one given by the babble-modulated noise, as the performance tends to change in a more considerable way, with the identification rate drastically decreasing to lowest points in the case of $N \in \{16, 32, 64\}$ and improving again with

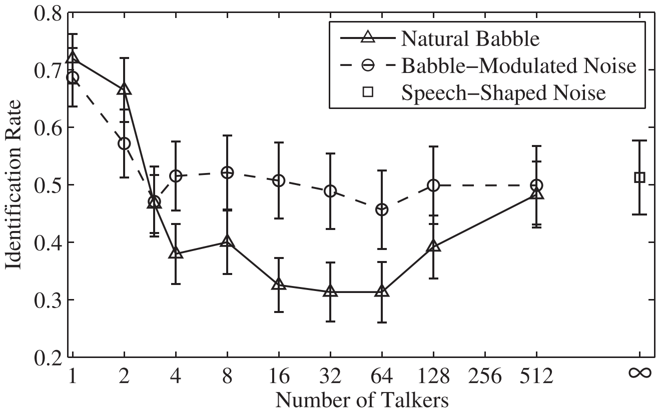

FIG. 1. Consonant identification rates in case of natural babble and babble-modulated noise at varying of the number of overlapping talkers N. Speech shaped noise is indicated as a condition with $N = \infty$ number of talkers.

$N \in \{128, 512\}$. In the case of natural babble, the conditions $N \in \{1, 2\}$ are significantly different from all other masking conditions but not between them. Additionally, $N = 512$ significantly differs from all other masking conditions apart from $N \in \{3, 8, 128, \infty\}$; $N = 3$ differs from $N \in \{16, 32, 64\}$; $N = \infty$ differs from $N \in \{8, 32, 64, 128\}$. In the case of babble-modulated noise, $N = 1$ is significantly different than all other masking conditions. Additionally, $N = 2$ is significantly different than $N \in \{3, 64\}$. Table I shows the consonant confusion matrix of the conducted experiments. The rows represent the consonants presented to the listeners (ground truth); the columns report the confusion rate, averaged over all of the experiments (natural babble, babble-modulate noise, and speech-shaped noise), over all the subjects, and over all the masking conditions. The diagonal shows the identification rate for each consonant. The confusion matrix demonstrates that some consonants are harder to identify than others.

Our results confirmed the findings of Simpson *et al.*[7] with regards to the identification rate being a nonmonotonic function of N, also in the case of non-native speakers. Nevertheless, the babble-modulated noise shows smoother changes in the identification rate on variation of N. It is interesting to note that for $N = 3$ and $N = 512$, the identification rates, both for natural and babble-modulated noise tend to coincide. The condition $N = 3$ likely represents a boundary for the natural babble between conditions in which words are still detectable and conditions in which they are not. According to Hoen *et al.*,[10] language-dependent factors like lexical masking are present when it is possible to detect words from the babble, and this should make the consonant identification harder, due to factors like informational masking, linguistic confusion, and attentive load, as also illustrated in Simpson *et al.*[7] However, it is possible that the language played a double role in this circumstance. In fact, if, on one hand, it has been proved that sound segregation in L2 is more challenging than in L1 (see Cooke *et al.*[9]), having chosen consonants whose sound is familiar to the listeners could have helped them in isolating the target from the mixture. As the number of talkers increases, even though no relevant changes are present in the identification rate, babble-modulated noise appears to be a less effective masker than natural babble. As mentioned by Simpson *et al.*,[7] this is also due to the fact that a more stationary background noise, which is characterised by less acoustical cues (like a considerably lower frequency of onsets), induces less attentional switching from the target. This effect should be particularly significant in the case of non-native speakers, who, as illustrated by Cutler *et al.*,[21] have a harder time recovering from disruption. With $N \geq 128$, the resulting babble tends to sound like stationary noise and the difference in the performance in both conditions tends to disappear. It is possible that this happens simply because the difference in the acoustic features of the stimuli also tend to disappear, leaving energetic masking as the only actual kind of masking still operating. Any language-related factor: semantic, lexical, or phonetical has no additional effect on the identification performance and the distraction introduced by onsets is lost due to the high frequency of them and, consequently, to the more acoustically smooth resulting combination.

TABLE I. Consonant confusion matrix.

	Subjects' response									
Consonant	/b/	/d/	/f/	/g/	/k/	/m/	/n/	/p/	/t/	/v/
/b/	**0.22**	0.24	0.01	0.17	0.01	0.01	0.01	0.01	0.01	0.32
/d/	0.01	**0.81**	0.01	0.10	0.01	0.00	0.01	0.01	0.03	0.01
/f/	0.17	0.03	**0.41**	0.04	0.04	0.01	0.01	0.16	0.04	0.11
/g/	0.05	0.20	0.02	**0.44**	0.06	0.02	0.03	0.02	0.04	0.11
/k/	0.03	0.05	0.11	0.12	**0.32**	0.02	0.02	0.17	0.10	0.05
/m/	0.09	0.05	0.02	0.04	0.04	**0.46**	0.06	0.02	0.02	0.19
/n/	0.04	0.16	0.02	0.09	0.05	0.20	**0.25**	0.03	0.04	0.12
/p/	0.01	0.00	0.09	0.00	0.19	0.00	0.00	**0.64**	0.06	0.00
/t/	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01	**0.96**	0.00
/v/	0.15	0.05	0.03	0.08	0.01	0.01	0.01	0.02	0.01	**0.63**

IV. CONCLUSION

This work investigates L1 consonant identification in the presence of L2 N-talker babble, natural and modulated. All the consonants used exist in the native language of the subjects. As in Simpson *et al.*,[7] the identification rate is a nonmonotonic function of N and the natural babble is a more effective masker than the modulated one, with an increase of the number of talkers. However, the babble-modulated noise shows smoother changes in the identification rate. Further experiments in the same conditions, following the same paradigm, but using English subjects, would better highlight this effect and help with the understanding of how the various factors linked to competence in the language operate on the identification procedure. Moreover, a larger set of subjects from each language group would allow us to investigate and compare the effect of the specific L1s on the performance of the subjects. Finally, it would be interesting to explore in more detail consonant confusion, looking for consistent confusion and mistakes. As shown in Table I, some of the consonants were, indeed, easier to recognise. A deeper investigation of the reasons behind this behavior could also be undertaken. In a different direction, the experimental paradigm described in this paper could be useful to the neuro-imaging community, that could compare neuro-imaging data collected while the subjects were executing the listening experiments to investigate the reasons behind misidentifications and identify the features of the stimuli that affect the identification process.

ACKNOWLEDGMENTS

This work was carried out at the Technical University of Denmark and partly funded by the European Commission FP7 Marie Curie Initial Training Network Investigating Speech Processing In Realistic Environments (INSPIRE). The INSPIRE project was funded with support from the European Commission under Contract FP7-PEOPLE-2O11-290000.

[1] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional selection in a multi-speaker environment can be decoded from single-trial EEG," Cerebral Cortex **2014**, 1697–1706 (2015).

[2] Colin E. Cherry, "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am. **25**(5), 975–979 (1953).

[3] Albert S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT press, Cambridge, MA, 1994).

[4] Rob Drullman and Adelbert W. Bronkhorst, "Speech perception and talker segregation: Effects of level, pitch, and tactile support with multiple simultaneous talkers," J. Acoust. Soc. Am. **116**(5), 3090–3098 (2004).

[5] Brian C. J. Moore and Hedwig Gockel, "Factors influencing sequential stream segregation," Acta Acust. Acust. **88**(3), 320–333 (2002).

[6] Adelbert W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," Acta Acust. Acust. **86**(1), 117–128 (2000).

[7] Sarah A. Simpson and Martin Cooke, "Consonant identification in N-talker babble is a nonmonotonic function of N," J. Acoust. Soc. Am. **118**(5), 2775–2778 (2005).

[8] Kristin J. Van Engen and Ann R. Bradlow, "Sentence recognition in native- and foreign-language multi-talker background noise," J. Acoust. Soc. Am. **121**(1), 519–526 (2007).

[9] Martin Cooke, Maria Luisa Garcia Lecumberri, Odette Scharenborg, and Wim A. Van Dommelen, "Language-independent processing in speech perception: Identification of English intervocalic consonants by speakers of eight European languages," Speech Commun. **52**(11), 954–967 (2010).

[10] Michel Hoen, Fanny Meunier, Claire-Léonie Grataloup, François Pellegrino, Nicolas Grimault, Fabien Perrin, Xavier Perrot, and Lionel Collet, "Phonetic and lexical interferences in informational masking during speech-in-speech comprehension," Speech Commun. **49**(12), 905–916 (2007).

[11] Sven L. Mattys, Lucy M. Carroll, Carrie K. W. Li, and Sonia L. Y. Chan, "Effects of energetic and informational masking on speech segmentation by native and non-native speakers," Speech Commun. **52**(11), 887–899 (2010).

[12] M. L. G. Lecumberri and Martin Cooke, "Effect of masker type on native and non-native consonant perception in noise," J. Acoust. Soc. Am. **119**(4), 2445–2454 (2006).

[13] Philipos C. Loizou, *Speech Enhancement: Theory and Practice* (CRC Press, Boca Raton, FL, 2013).

[14] Gordon E. Peterson and Ilse Lehiste, "Revised CNC lists for auditory tests," J. Speech Hearing Disorders **27**(1), 62–70 (1962).

[15] Robyn M. Cox, Genevieve C. Alexander, and Christine Gilmore, "Development of the connected speech test (CST)," Ear Hear. **8**(5), 119S–126S (1987).

[16] E. H. Rothauser, W. D. Chapman, N. Guttman, K. S. Nordby, H. R. Silbiger, G. E. Urbanek, and M. Weinstock, "IEEE Recommended Practice for Speech Quality Measurements," IEEE Trans. Audio Electroacoust. **17**(3), 225–246 (1969).

[17] Yi Hu and Philipos C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," Speech Commun. **49**(7), 588–601 (2007).

[18] Ling Feng, "Speaker recognition," Ph.D. thesis, Technical University of Denmark, DTU, IMM-THESIS, DK-2800 Kgs. Lyngby, Denmark, 2004.

[19] Auditec dataset, Auditory Tests (Revised), Auditec, St. Louis, MO, 1997.

[20] Douglas S. Brungart, "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**(3), 1101–1109 (2001).

[21] A. Cutler, M. L. G. Lecumberri, and M. Cooke, "Consonant identification in noise by native and non-native listeners: Effects of local context," J. Acoust. Soc. Am. **124**(2), 1264–1268 (2008).