Word segmentation in phonemically identical and prosodically different sequences using cochlear implants: A case study

Anahita Basirat

Word segmentation in phonemically identical and prosodically different sequences using cochlear implants: A case study

Anahita Basirat

Univ. Lille, CNRS, CHU Lille, UMR 9193 - SCALab - Sciences Cognitives et Sciences Affectives, F-59000 Lille, France

Address correspondence to:

Anahita Basirat

Département d'orthophonie

Faculté de Médecine (pôle formation)

Université Lille 2

F-59045 Lille Cedex

France

Email: anahita.basirat@univ-lille2.fr

Abstract

Cochlear implant (CI) users frequently achieve good speech understanding based on phoneme/word recognition. However, there is a significant variability between CI users in processing prosody. The aim of this study was to examine the abilities of an excellent CI user to segment continuous speech using intonational cues. A post-lingually deafened adult CI user and twenty-two normal hearing (NH) subjects segmented phonemically identical and prosodically different sequences in French such as “l’affiche” (the poster) vs. “la fiche” (the sheet), both [lafiʃ]. All participants also completed a minimal pair discrimination task. Stimuli were presented in auditory-only and audiovisual presentation modalities. The performance of the CI user in the minimal pair discrimination task was 97% in the auditory-only and 100% in the audiovisual condition. In the segmentation task, contrary to the NH participants, the performance of the CI user did not differ from the chance level. Visual speech did not improve word segmentation. This result suggests that word segmentation based on intonational cues is challenging when using CIs even when phoneme/word recognition is very well rehabilitated. This finding points to the importance of the assessment of CI users’ skills in prosody processing and the need for specific interventions focusing on this aspect of speech communication.

Key words: word segmentation; prosody; cochlear implant; audiovisual speech

INTRODUCTION

In clinical and experimental studies with cochlear implant (CI) users, their general abilities in speech recognition, such as disyllabic word recognition and sentence comprehension, are often reported while their abilities in processing speech prosody usually remain unknown. However, the correct recognition of prosody is essential in speech communication since it conveys meaning, to some extent. For example, intonational cues enable us to distinguish between questions and statements with the same syntactic structures (e.g. “She is sleeping.” versus “She is sleeping?”). The role of prosody is extremely important in tonal languages such as Mandarin in which prosodic cues provide listeners with information to differentiate word meanings. Moreover, prosody cues play a role in emotion recognition in speech. The present study focused on speech segmentation using prosodic cues in French. In order to segment a continuous speech signal into words, listeners use different types of information including lexical, segmental (e.g. phonotactic constraints) and prosodic cues (Mattys, White and Melhorn, 2005). In English, most content words (nouns, adjectives, etc.) begin with stressed syllables (Cutler and Carter, 1987). This stress pattern is used by English listeners to detect word boundaries and to segment speech (Cutler and Norris, 1988) especially in noisy conditions (Mattys et al., 2005). In French, this word stress pattern does not exist. However, a specific prosodic cue can be used by French listeners to segment speech: a nonsense sequence like [me.la.mɔ̃.din] is interpreted by French listeners as a single sequence “mélamondine” if the fundamental frequency (F0) rise starts at the first syllable (i.e. [me]) and as two distinct sequences “mes lamondine” (my lamondine) when the F0 rise starts at the second syllable (i.e. [la]) (Welby, 2007). In sentences without contextual information, French listeners make use of this cue to segment sequences of words that are phonemically identical and

prosodically different such as “l’affiche” (the poster) and “la fiche” (the sheet), both [lafiʃ] (Spinelli, Grimault, Meunier and Welby, 2010). In fact, the vowel /a/ has a greater F0 in a word beginning with a vowel (e.g. “l’affiche”) than in its phonemically equivalent word beginning with a consonant (e.g. “la fiche”) (Spinelli, Welby and Schaegis, 2007). These results confirm that in the absence of semantic/lexical/segmental cues, French listeners rely on prosodic cues to segment speech.

Several studies have shown that adult and pediatric CI users perform less well in prosody recognition compared to normal hearing (NH) listeners (e.g. Chatterjee and Peng, 2008; Van Zyl and Hanekom, 2013; Peng, Tomblin and Turner, 2008; See, Driscoll, Gfeller, Kliethermes and Oleson, 2013). This poor performance can be due to the limitations of CI devices in conveying fine structure information (for a review, see Wilson and Dorman, 2008). However, it seems that prosodic cues could, to some extent, be available to CI users. Spitzer, Liss, Spahr, Dorman and Lansford (2009) studied the performance of adult CI users in using lexical stress cues to find boundaries between words in continuous speech. The authors analyzed the lexical boundary errors of CI users when they listened to sentences in English. If the listeners used syllabic stress, they would commit some predicted errors. For example, they would insert lexical boundaries before strong syllables as these indicate word-onsets in English (Cutler and Carter, 1987). Their results showed that adult CI users committed the same type of segmentation errors as predicted, suggesting that they use lexical stress cues for speech segmentation. These cues seem to be processed by CI users very early during development. Using a habituation-test procedure, Segal, Houston and Kishon-Rabin (2016) observed that CI infants (mean age = 18 months) were able to discriminate nonsense CVCV sequences, which differed only in their stress pattern. In this study, infants were presented with several repetitions of a CVCV sequence with a stress on either the first or the second

syllable (habituation phase). Then, they were presented with the same structure as in the habituation phase with either the same or a different stress pattern. Infants detected the change in the stress pattern when the stress pattern of the novel sequence differed from that of the habituated sequence. This suggests that CI infants can process lexical stress cues. Intonational cues also seem to be processed by CI users to some extent. In a longitudinal study, Snow and Ertmer (2012) analyzed the accent range (i.e. the amount of F0 change used in falling or rising contours) during the spontaneous speech production of CI children. They observed similar skills in CI children to those predicted by a model of early intonation development in NH children. In a sentence recognition task, Meister, Landwehr, Pyschny, Grugel and Walger (2011) showed that adult CI users could extract intonational cues in adverse listening conditions. The authors presented CI users with sentences with regular and inverted F0-contours in noise. Participants were asked to recognise sentences and to repeat as many words as possible. Despite the poorer performance of CI users compared to NH listeners, speech recognition was better in the regular F0 condition than in the inverted F0-contours condition in both populations. These results suggest that CI users could make use of intonational cues despite the limitations of CI devices in conveying these cues. Interestingly, there is a significant variability between CI users in processing intonational cues, which has been reported in both adult (Chatterjee and Peng, 2008) and pediatric CI users (Peng, Tomblin and Turner, 2008). The variability between adult CI users could be related to their hearing experience before deafness or their residual hearing: intonation recognition seems less challenging for post-lingually deaf adults (Peng, Chatterjee and Luc, 2012) and those with residual hearing after implantation (Marx et al., 2015).

To our knowledge, the abilities of CI users in word segmentation relying only on intonational cues are not known. Given that these cues are, to some extent, available to CI users, our hypothesis was that exceptional CI users might be able to segment continuous speech through

intonation identification. We thus tested the performance of an exceptional CI user in segmenting phonemically identical and prosodically different sequences (e.g. “l’affiche” (the poster) versus “la fiche” (the sheet), both [lafiʃ]) embedded in neutral sentence contexts. Sentences were presented in both auditory-only and audiovisual presentation modalities. NH listeners seem to be able to extract prosodic cues such as F0 variations from visual speech (Munhall, Jones, Callan, Kuratate, and Vatikiotis-Bateson, 2004; Cavé et al., 1996). Moreover, seeing lip movements can bias the segmentation of ambiguous French sequences (Strauß, Savariaux, Kandel and Schwartz, 2015). In the latter study, NH listeners were presented with auditory sequences that were phonemically and prosodically ambiguous as they were compatible with two French words such as “l’affiche” (the poster) and “la fiche” (the sheet). These sequences were deliberately produced in an ambiguous way (and not in a natural way) such that listeners could use no auditory intonational cues to segment them. Thus, two word segmentations were possible for each auditory sequence. The authors dubbed the visual lip movements of hyper-articulated words beginning with a vowel (e.g. “l’affiche”) or beginning with a consonant (e.g. “la fiche”) onto the corresponding auditory ambiguous sequences. In a segmentation task, participants reported more segmentation compatible with vowel-beginning words when they were presented with the lip movements of vowel-beginning words and more segmentation compatible with consonant-beginning words when they were presented with the lip movements of consonant-beginning words. This suggests that NH listeners can make use of visual cues to segment auditory ambiguous sequences. In the present study, we expected that word segmentation using intonational cues would be possible, to some extent, for exceptional CI users. Moreover, we expected that audiovisual speech would facilitate the extraction of these cues and thus improve word segmentation.

MATERIALS AND METHODS

Subjects

An exceptionally good CI user (23-year-old female) participated in this study (“Ms. C”). She was a French native speaker. She reported a sudden profound hearing loss at 22 years of age caused by meningitis. She was bilaterally implanted 1 month after her hearing loss. Both devices were a Neurelec Digisonic SP implant with a Saphyr processor. Seven months after CI implantation, her scores using the PAV2L assessment tool, which is used in Lille Hospital in France (Tourmel, 2007), were very good (e.g. disyllabic word recognition = 100% correct without lip-reading; oral text comprehension (level 3) = 96% without lip-reading and 100% with lip-reading). In everyday life, she communicated regularly on the telephone, watched TV and listened to music. She had corrected-to-normal vision. She took part in this study 13 months after CI implantation while she was a graduate student. Twenty-two adults (mean age = 26.6 years, SD = 7, fourteen females) without a diagnosed hearing problem served as controls (NH group). They were native French speakers. They all had normal or corrected-to-normal vision. The study was conducted in accordance with the Helsinki Declaration and the ethical guidelines of the Department of Speech and Language Therapy of Lille University. Before testing, all participants were informed about the experiment by a written document and signed a consent statement.

Stimuli

Twenty-nine phonemically identical and prosodically different pairs from Spinelli et al. (2010) such as “l’affiche” (= the poster) and “la fiche” (= the sheet), both [lafiʃ], and twenty-nine minimal pairs such as “la bouche” [labuʃ] (= the mouth) and “la douche” [laduʃ] (= the shower) were selected. The minimal pairs differed in articulation place (16 pairs) or in voicing (13 pairs). These sequences were inserted into a neutral carrier sentence “c’est ... dont je t’ai parlé” (= it’s … about which I spoke to you). Stimuli were produced by a female native

French speaker and recorded audiovisually by a camera in front of her. The head and neck of the speaker were visible in the videos.

Procedure

The experiment was conducted in two sessions, one in the auditory-only and the other in the audiovisual modality. In the auditory-only session, NH participants were asked to listen to the sentences through headphones. In the audiovisual session, they were asked to listen to the stimuli through headphones and to watch the videos. The sound was presented at a comfortable level and the screen was placed at about 50 cm from the participants. Each trial began with the presentation of a sentence containing one member of the fifty-eight pairs. Then, the two members of each pair were presented on the screen. The participants were asked to identify the word they perceived. They made a forced choice between two possible words by pressing one of the two response keys on the keyboard (F and J keys). Only one member of each pair was presented to each participant. The participants were presented with the same items in the auditory-only and audiovisual sessions. The word presentation order was randomised. The order of the auditory-only and audiovisual sessions was counterbalanced across NH participants. The second session was about two weeks after the first session in order to minimise the response bias. All participants carried out 6 familiarisation trials before the experimental trials.

“Ms C” followed the same presentation procedure. The stimuli were presented through two loudspeakers. During the first session, she was presented with the audiovisual stimuli. Eighteen days later, she was presented with the auditory-only stimuli. “Ms. C” took part in the experiment after the participation of the NH group in the study.

RESULTS

The mean percentages of correct responses of the NH group are presented in Table 1. We will focus on the performance of participants in the identification of phonemically identical and prosodically different sequences. The performance of the NH group was above the chance level in both the auditory-only and audiovisual presentation modalities (A: mean = 80.56%, $t(21) = 15.3$, $p<0.001$; AV: mean = 82.13%, $t(21) = 13.65$, $p<0.001$). An ANOVA with modality as within-subject and session order as between-subject factors showed no significant effect (modality: $F(1,20) = 0.47$, n.s.; session order: $F(1,20) = 0.72$, n.s.; two-way interaction: $F(1,20) = 4.03$, $p<0.1$).

	Phonemically identical and prosodically different sequences	Minimal pairs
Auditory-only	80.6% (SE = 2)	99.5% (SE = 0.4)
Audiovisual	82.1% (SE = 2.3)	99.1% (SE = 0.3)

Table 1: Mean percentages of correct responses of the NH group for the identification of phonemically identical and prosodically different sequences and minimal pairs. SE = standard error.

The percentages of correct responses of “Ms. C” are presented in Table 2. Her performance against the chance level was examined using binomial tests. She was above chance in identifying minimal pairs in both the auditory-only and audiovisual modalities (A: 28/29 correct responses, $p<0.001$; AV: 29/29 correct responses, $p<0.001$). Her performance did not differ from chance in identifying phonemically identical and prosodically different sequences (A: 18/29 correct responses, n.s.; AV: 16/29 correct responses, n.s.). McNemar tests were carried out to compare her performance in the auditory-only and audiovisual modalities for phonemically identical and prosodically different sequences and for minimal pairs. Visual speech did not seem to improve the performance of “Ms. C” (minimal pairs: McNemar $\chi^2 = 0$, n.s.; phonemically identical and prosodically different sequences: McNemar $\chi^2 = 0.1$, n.s.).

	Phonemically identical and prosodically different sequences	Minimal pairs
Auditory-only	62.1%	96.5%
Audiovisual	55.2%	100%

Table 2: Percentages of correct responses of "Ms. C" for the identification of phonemically identical and prosodically different sequences and minimal pairs.

DISCUSSION

The results show that NH listeners can segment phonemically identical and prosodically different sequences above the chance level, at 80%, in the auditory-only presentation modality. This is in line with the study of Spinelli et al. (2007). Our results show no differences between the auditory-only and audiovisual presentation modalities. In fact, Strauß et al. (2015) observed that seeing the speaker's lip movements plays a role in word segmentation. The absence of such an effect in our study can be explained by the fact that our stimuli were acoustically and visually different from theirs. Strauß et al. (2015) used hyper-articulated visual lip movements with ambiguous auditory sequences without relevant acoustical cues for segmentation. Therefore, the visual cues were very important. In our study, we asked the speaker to utter the sentences in a natural way. Thus, compared to the stimuli of Strauß et al. (2015), the auditory cues were more important and the visual cues less so. Compared to auditory speech, natural visual speech does not seem to provide NH listeners with additional intonational cues to segment phonemically identical and prosodically different sequences. Regarding the performance of "Ms. C", as expected from clinical observations, she was very good at discriminating between minimal pairs. However, her performance in segmenting phonemically identical and prosodically different sequences did not differ from the chance level. She was unable to make use of intonational cues in word

segmentation. Although CI users are better than NH listeners at using visual speech (Rouger et al., 2007), no performance gain was observed in the audiovisual presentation modality. The absence of visual benefit in this study could be explained by the fact that “Ms. C” received her bilateral CIs soon (one month) after her hearing loss and did not practice lip-reading alone for a long period. However, NH individuals, without any explicit lip-reading practice, benefit from visual speech especially in adverse conditions such as noise (Peelle and Sommers, 2015, for a review). Regarding the segmentation of phonemically identical and prosodically different sequences, we believe that reliable intonational cues for segmenting these sequences are not very important in visual speech. This is why visual hyper-articulated stimuli dubbed onto auditory ambiguous sequences influence segmentation in NH listeners (Strauß et al., 2015) but not when utterances are produced naturally as in our study. In summary, despite an excellent rehabilitation of some aspects of speech processing, “Ms. C” seemed unable to segment natural speech by using auditory intonational cues. Visual cues were not salient enough to improve her performance.

This finding has implications for speech-language pathologists. It points to the importance of assessing the skills of CI users in the fine processing of auditory and visual prosody, as recommended in other speech and language pathologies (Peppé, 2009; Swerts, 2009). In fact, an excellent rehabilitation in speech understanding based on phoneme/word recognition with CIs is not necessarily accompanied by a good performance in the extraction of speech prosodic cues. Although fine structure information transmitted by CI devices is degraded, prosodic cues are not completely absent (Wilson and Dorman, 2008) and CI users can process auditory prosodic cues to some extent (e.g. Spitzer et al. 2009; Meister et al., 2011; Snow and Ertmer, 2012; Segal et al., 2016). Training in this aspect of speech is therefore feasible and needs to be systematically included in rehabilitation programmes.

ACKNOWLEDGMENTS

The author thanks “Ms. C” and normal-hearing subjects for their participation in this study, Pierre-François Tournade and Amandine Lepachelet for their help in recording the stimuli, and Fanny Heurtebise and Muriel Lefeuvre for their help in collecting the data.

REFERENCES

Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., & Espesser, R. (1996). About the relationship between eyebrow movements and F0 variations. In *Proceedings of the International Conference on Spoken Language Processing*, 2175-2178.

Chatterjee, M., & Peng, S. C. (2008). Processing F0 with cochlear implants: Modulation frequency discrimination and speech intonation recognition. *Hearing Research*, *235*(1), 143-156.

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech & Language*, *2*(3), 133-142.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human perception and performance*, *14*(1), 113.

Marx, M., James, C., Foxton, J., Capber, A., Fraysse, B., Barone, P., & Deguine, O. (2014). Speech Prosody Perception in Cochlear Implant Users With and Without Residual Hearing. *Ear and Hearing*, *36*(2), 239-248.

Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of multiple speech segmentation cues: a hierarchical framework. *Journal of Experimental Psychology: General*, *134*(4), 477.

Meister, H., Landwehr, M., Pyschny, V., Grugel, L., & Walger, M. (2011). Use of intonation contours for speech recognition in noise by cochlear implant recipients. *The Journal of the Acoustical Society of America*, *129*(5), EL204-EL209.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility. Head movement improves auditory speech perception. *Psychological Science*, *15*(2), 133-137.

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, *68*, 169-181.

Peng, S. C., Chatterjee, M., & Lu, N. (2012). Acoustic cue integration in speech intonation recognition with cochlear implants. *Trends in Hearing*, *16*(2), 67-82.

Peng, S. C., Tomblin, J. B., & Turner, C. W. (2008). Production and perception of speech intonation in pediatric cochlear implant recipients and individuals with normal hearing. *Ear and Hearing*, *29*(3), 336-351.

Peppé, S. J. (2009). Why is prosody in speech-language pathology so difficult?. *International Journal of Speech-Language Pathology*, *11*(4), 258-271.

Rouger, J., Lagleyre, S., Fraysse, B., Deneve, S., Deguine, O., & Barone, P. (2007). Evidence that cochlear-implanted deaf patients are better multisensory integrators. *Proceedings of the National Academy of Sciences*, *104*(17), 7295-7300.

See, R. L., Driscoll, V. D., Gfeller, K., Kliethermes, S., & Oleson, J. (2013). Speech intonation and melodic contour recognition in children with cochlear implants and with normal hearing. *Otology & Neurotology*, *34*(3), 490-498.

Segal, O., Houston, D., & Kishon-Rabin, L. (2016). Auditory discrimination of lexical stress patterns in hearing-impaired infants with cochlear implants compared with normal hearing: influence of acoustic cues and listening experience to the ambient language. *Ear and Hearing*, *37*(2), 225-234.

Snow, D. P., & Ertmer, D. J. (2012). Children's development of intonation during the first year of cochlear implant experience. *Clinical Linguistics & Phonetics*, *26*(1), 51-70.

Spinelli, E., Welby, P., & Schaegis, A. L. (2007). Fine-grained access to targets and competitors in phonemically identical spoken sequences: The case of French elision. *Language & Cognitive Processes*, 22(6), 828-859.

Spinelli, E., Grimault, N., Meunier, F. & Welby, P. (2010). An intonational cue to segmentation in phonemically identical sequences. *Attention, Perception & Psychophysics*, *72*(3), 775-787.

Spitzer, S., Liss, J., Spahr, T., Dorman, M., & Lansford, K. (2009). The use of fundamental frequency for lexical segmentation in listeners with cochlear implants. *The Journal of the Acoustical Society of America*, *125*(6), EL236-EL241.

Strauß, A., Savariaux, C., Kandel, S., & Schwartz, J. L. (2015). Visual lip information supports auditory word segmentation. In *FAAVSP 2015*. Vienna, Austria.

Swerts, M. (2009). The relevance of visual prosody for studies in language and speech-language pathology. *International Journal of Speech-Language Pathology*, *11*(4), 282-286.

Tourmel, M. (2007). *Validation du PAV2L: évaluation de la Perception auditive verbale et de la lecture labiale de l'adulte devenu sourd* (undergraduate dissertation). University of Lille 2, Lille.

Van Zyl, M., & Hanekom, J. J. (2013). Perception of vowels and prosody by cochlear implant recipients in noise. *Journal of Communication Disorders*, *46*(5), 449-464.

Welby, P. (2007). The role of early fundamental frequency rises and elbows in French word segmentation. *Speech Communication*, *49*(1), 28-48.

Wilson, B. S., & Dorman, M. F. (2008). Cochlear implants: a remarkable past and a brilliant future. *Hearing Research*, *242*(1), 3-21.