Interspeech 2018
2-6 September 2018, Hyderabad

Who are you listening to? Towards a dynamic measure of auditory attention to speech-on-speech

Moïra-Phoebé Huet[1,2], *Christophe Micheyl*[1,3], *Etienne Gaudrain*[1,4], *Etienne Parizet*[2]

[1]Université de Lyon, CNRS UMR5292, Inserm U1028, Lyon Neuroscience Research Center, Lyon, France
[2]Université de Lyon, Institut National des Sciences Appliquées de Lyon, Laboratoire Vibrations Acoustique, F-69621 Villeurbanne, France
[3]Starkey, Créteil, France
[4]University of Groningen, University Medical Center Groningen, Department of Otorhinolaryngology, Groningen, Netherlands

moira-phoebe.huet@inserm.fr

Abstract

When studying speech-on-speech perception, even when participants are explicitly instructed to focus selectively on a single voice, they can spuriously find themselves listening to the wrong voice. These paradigms generally do not allow to infer, retrospectively, which of the speakers was listened to, at different times during presentation. The present study sought to develop a psychophysical test paradigm, and a set of speech stimuli to that purpose. In this paradigm, after listening to two simultaneous stories, the participant had to identify, among a set of words, those that were present in the target story. Target and masker stories were presented dichotically or diotically. F0 and vocal-tract length were manipulated in order to parametrically vary the distance between the target and masker voices. Consistent with the hypothesis that correct-identification performance for target words depends on selective attention, performance decreases with the distance between the target and masker voices. These results indicate that the paradigm and stimuli described here can be used to infer which voice a participant is listening to in concurrent-speech listening experiments.

Index Terms: speech-on-speech, auditory attention, method, voice

1. Introduction

Sixty-five years after Colin Cherry articulated the *cocktail party* problem [1], how the human ear and brain solve this problem are still an active topic of research. Notably, a series of studies performed during the past decade have identified neural correlates of auditory selective attention in concurrent-speech listening tasks [2, 3, 4].

In general, in speech-on-speech tasks, the listener is asked to focus on one of two simultaneously presented voices, for example, a female voice heard on the right side (the target), while trying to ignore a male voice on the left side (the masker). In most cases, for the data analysis, it was assumed that the participants listen unwaveringly the "target" voice. There are number of circumstances that can render this hypothesis less plausible. First as introspection and experience suggest, auditory selective attention is not infallible, and controlling one's attentional focus for several seconds, let alone minutes, can be quite challenging. Even with the best intentions, the participant's attention can occasionally be attracted toward a non-target element in the auditory scene. Second, when the voices are similar, they can be confused for one another, leading the participant to listen to the masker while believing it is the target. To be able to better identify neural correlates of auditory selective attention when target and masker can be confused, one would need to assess what the participant was actually attending to at different moments during the listening task.

The present study sought to validate a behavioral measure of listeners' attention in a concurrent-speech task, which can be used to infer which voice participants are actually listening at in such a task. For this purpose, a new speech material was created. Many corpora of sentences (e.g., CRM, Matrix, HINT, etc.) and shorter stimuli (e.g., syllables, phonemes, etc.) were used in previous speech-on-speech studies [5] but they are not representative of natural speech in the context of ecological communication. Moreover, closed set corpora where possible responses are presented before the trial do not constrain participants to listen to the entirety of the stimuli for a successful speech identification. An open set response, wherein participants need to listen to the whole stimulus, is therefore needed to achieve more realism. To sum up, the open set corpus created and used in this study aimed to be longer and more ecological than those used in previous studies. It is noteworthy than stimuli used in the decoding attention studies [2, 3, 4] last 60 seconds but they do not allow to infer which voice participants are listening to.

According to the literature, voice characteristics and spatialisation are two main cues to segregate two talkers [6]. Therefore, we used these two factors to control the difficulty of the task. Two stimuli presentation configurations were used in this study: a dichotic and a diotic presentations. In the dichotic presentation mode, the target and masker signals do not physically overlap, as one is presented to one ear, and the other to the other ear. In the diotic presentation mode, however, the target and masker signals physically overlap and, in that case, listeners can only rely on other perceptual cues to separate them. Voice characteristics are also important to discriminate and segregate two talkers. In practice, it is not easy to determine how different two voices are. Therefore, we created the masker voices from the target voices and manipulated two voice parameters: the voice pitch (F0) and the vocal-tract length (VTL) [7, 8].

In conclusion, this study presents the creation and the validation of a new set of stimuli and a behavioral measure to infer which voice participants are listening to at different time points during the presentation. Participants performance will be

10.21437/Interspeech.2018-2053

presented and discussed in two main conditions: two different stimuli presentation configurations and three different masker voices.

2. Method

2.1. Participants

Twenty-two participants, aged between 20 to 32, took part in this study. All participants were French native speakers and had audiometric thresholds $\leq$ 20 dB HL at audiometric test frequencies between 250 Hz and 8 kHz.

2.2. Procedure

Each trial began with the word "attention" uttered by the target voice since listeners can use priming voice information to identify and attend a target message [9]. After listening to two simultaneous stories — a target and a masker — the participants had to find, among a set of 9 words, those present in the target story. The set of words was composed of 3 words from the target story, 3 words from the masker story and 3 extraneous words.

The entire procedure consisted of a short training followed by 12 blocks. Between blocks, subjects could take a break and resume the experiment when they wished. Each block was composed of a list of story pairs presented in a random order. Within a block, the same voice condition was used. Moreover, six of the blocks were presented diotically and the six other blocks were presented dichotically. For each presentation, 2 blocks had the JND voice as the masker voice, 2 other blocks had the intermediate voice and the last 2 blocks had the male voice. Characteristics of these voices are described in the next section. The presentation mode (diotic or dichotic) and the masker voice (JND, Intermediate or Male) were randomly assigned to each block.

Data collection lasted 60 to 90 min, and the entire procedure was completed in a single session.

2.3. Stimuli

2.3.1. Material Content

The stimuli consisted of short stories extracted from the French Audiobook *Le Charme discret de l'intestin* [The Inside Story of Our Body's Most Underrated Organ] [10]. A total of 528 stories were selected according to several criteria, namely, the duration ($11 - 18$ seconds) and the number of words ($22 - 55$). The stories had to make sense and be engaging, so mostly anecdotes and fun facts were selected.

The words that the subjects were asked to find in the stories were also carefully selected. Firstly, three words per story were selected at different times: a word at the beginning of the story, a word in the middle and a word at the end. However the very first and last words of the stories were excluded from the selection in order to reduce the recency and primacy effects. Secondly, the selected words could only appear once in the story to avoid repetition. Thirdly, based on a lexical database [11], the words that were too rare or too frequent were also excluded from the selection. At the end, the language frequency distribution of the selected words had a mean of 7.45 per million of occurrences in the French language and 95% of values range between 0.10 and 528.76 per million of occurrences.

Then, 264 pairs of stories, each composed of a target story and a mask story, were created according to their duration such that the words chosen for the target story did not appear in the masker story and vice versa. Three extraneous words were also assigned to each pair so that all 9 words would be visually presented together.

This material was validated in a preliminary online study. Two hundred and nine volunteers managed to identify the words of the stories presented in isolation condition (without masker), and obtained an average score of 89%. This led to removing 87 stories that yielded poorer scores (participants failed to find a target word from the story). At this stage, 167 "good" pairs of stories remained in the set of material.

Finally, a subset of 12 lists of 12 story pairs were selected from the 167. Once again, if two pairs of stories contained the same words, they could not belong to the same list.

In the end, the material consisted of 12 cleaned lists of story pairs counterbalanced in duration. All lists had a similar total duration ($\mu = 175.08$ s, $\sigma = 0.9$) and within each list, there were equally short and long stories.

2.3.2. Voice Manipulation

The audiobook was recorded by a female speaker. The original female voice was used to produce the target voice. Without any parameter change, the orginal voice was analysed and resynthesed using the STRAIGHT software implemented in MATLAB [12]. To produce the maskers, the voice parameters were also manipulated through analysis-resynthesis. Two voice characteristics, voice pitch (F0) and vocal-tract length (VTL), were manipulated with STRAIGHT.

Voice	ΔF0	ΔVTL	Total distance
Male	8	3.04	8.56
JND	1.6	0.61	1.71
Intermediate	4.8	1.82	5.13

Table 1: *Distance, in semitones, from the original voice for each voice condition. The total distance between two voices is calculated as* $\sqrt{\Delta F0^2 + \Delta VTL^2}$.

First, a credible male voice was synthesized from the original voice by adjusting these two parameters in order to obtain the direction for the voice changes. Then, from this voice, three voices were created based on literature. A change of 8 semitones (st) in F0 and 3.04 st in VTL were enough to create a "Male" voice [8] (see Table 1). For the second voice, the parameters were adjusted on the basis of the just-noticeable-differences (JNDs) for F0 and VTL [13]. This voice is thus a female voice that is barely distinguishable from the original speaker. The JND voice has a total distance of 1.71 semitones (st) along the female-male axis. A third voice was synthesized to be equidistant from the two previous voices.

3. Results

3.1. General description

Figure 1 shows the averaged performance in percentage of correctly identified target words to the total number of words presented in each condition. Chance performance was 33%. All subjects demonstrated high scores, well above chance, for each and all condition. All conditions were at ceiling except for the diotic presentation with the just-noticeable-difference voice.

A two-way repeated measures ANOVA was used to analyse the *logit* transformed scores and indicated a significant effect of voice [$F(1, 21) = 80.56, p < 0.001, \eta_g^2 = 0.32$] as well as a significant effect of presentation mode [$F(2, 42) =$

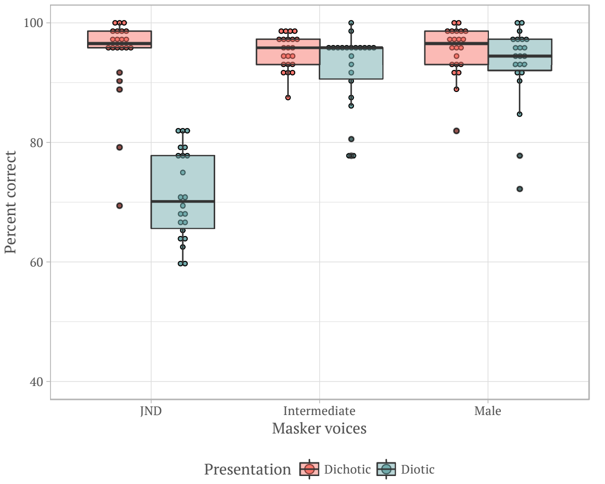

Figure 1: *Percentage of correct responses for each voice in dichotic presentation (in red) and in diotic presentation (in green). The points represent the scores for every participant in each condition. The hinges of the boxplot represent the first and the third quartile. The middle of the boxplot is the median. The length of the whiskers is 1.5 interquartile range.*

$58.66, p < 0.001, \eta_g^2 = 0.29$]. The interaction was also significant [$F(2, 42) = 46.62, p < 0.001, \eta_g^2 = 0.3$]. Post-hoc analysis with a holm correction showed that when stimuli were presented diotically, participants performed worse when the masker was the JND voice than when the masker was the intermediate voice ($t(30) = -10, p < 0.001$) or the male voice ($t(30) = -10, p < 0.001$). On the other hand, results with the intermediate voice and the male voice were not significantly different from each other.

A generalized linear mixed model (gLMM) was also fitted to the binary (correct/incorrect) scores for every words. The models were implemented in R using the lme4 package [14]. Before conducting a model selection procedure, a random intercept per subject was included in the initial model. This initial model do not have fixed effect and can be written in lme4 syntax:

$$response \sim 1 + (1|subject).$$

Then, a model selection procedure was conducted: a fixed factor was included (e.g., voice, presentation) to a new model. If the comparison between the initial model and the new model, based on a chi-square test with the log-likelihood difference, was significant, then the new model with the added fixed factor was kept and the procedure was repeated. At the end of the procedure, the final model was:

$$response \sim voice{*}(presentation{+}position){+}(1|subject).$$

The fixed effects of the final model were the voice, the presentation and the position of the word (beginning, middle or end of the story), as well as the interaction between the voice and the presentation and the interaction between the voice and the position of the word in the story. The afex package [15] was used to calculate the p-values for all fixed effects by comparing the final model to restricted models. In restricted models, the parameters corresponding to the fixed effect estimated is fixed to 0. On the basis of the final model, 5 restricted models were estimated (3 fixed effects and 2 interactions). Based on the likelihood ratio tests, the voice [$\chi^2(2) = 133.34, p < 0.001$], the presentation [$\chi^2(1) = 141.84, p < .001$], the word position [$\chi^2(2) = 106.69, p < 0.001$], the interaction between the voice and the presentation [$\chi^2(2) = 87.67, p < .001$] and the interaction between the voice and word position [$\chi^2(4) = 24.01, p < .001$] were significant. Figure 2 illustrates the percentage of correct response for each position of the words in the story for each voice.

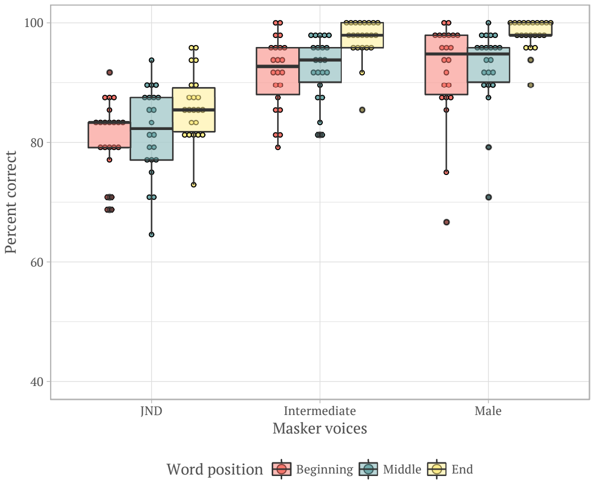

Figure 2: *Percentage of correct responses for each voice and for each word.*

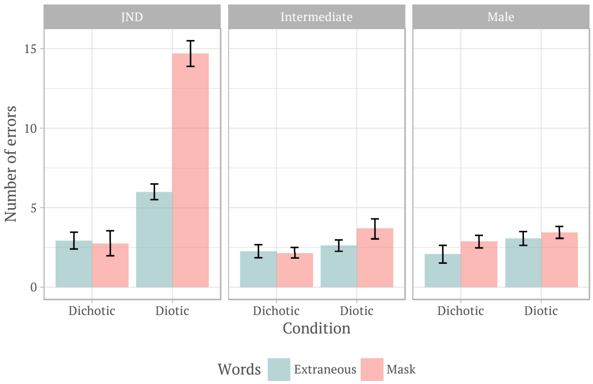

Figure 3: *Error distribution for each condition for the mask word (in red) and the extraneous word (in green). The bars represent the mean of errors across all the participants for each condition and the error bars represent the standard error of the mean.*

3.2. Error analysis

Results from the previous section showed that participants have poorer results for a particular condition: the JND voice with a diotic presentation. Analysing the participants errors is important because it gives information on the nature of these errors. Did the participant made a mistake and chose the wrong word because he was not able to listen to the target voice or because he was actually listening to the masker voice?

Figure 3 illustrates how the errors are distributed for each condition. A gLMM was conducted to the binary (mask/extraneous) scores for every mistake with the six conditions for fixed factor. Results showed that the diotic presentation with the JND voice was different from the 5 other conditions (see Table 2). Moreover, subjects made, in the di-

Presentation	Voice	z value	p value
Diotic	Intermediate	-3.28	0.001
	Male	-2.89	0.004
Dichotic	JND	-3.65	< 0.001
	Intermediate	-3.02	0.003
	Male	-2.10	0.036

Table 2: *Comparison between the diotic JND voice condition and the other conditions.*

otic JND voice condition, 2.4 more mistakes with the mask word than with the extraneous word which is above chance [$t(20) = 10, p < .001$]. These results indicate that in more difficult condition such as in the diotic presentation with a JND voice, participants were listening to the masker voice instead of the target voice for at least some part of the stories.

4. Discussion

The finding of higher performance for dichotic than for diotic listening conditions is consistent with the literature, which indicate benefits of spatial separation in concurrent speech-listening tasks [16]. The finding of increasing performance with increasing F0 and vocal-tract size distance between the target and masker voices is also in line with the results of previous studies [7, 8], which showed that participants' performance decreased when the competing voices were similar.

Controlling parametrically the distance between the target voice and the masker voice has been already done previously with some studies involving short stimuli, such as syllables and sentences [17, 7]. Our results show that this method can be extended to continuous speech-on-speech. The present study shows than more ecological stimuli can be used to infer which voice the participant is listening to even if the task potentially requires a bigger cognitive load (e.g., listening carefully for more than 12 seconds, remembering the words, etc.). Crucially, these longer stimuli also allow us to track the time course of the voice that participants were listening to, at three key time points of the story (the beginning, the middle and the end).

One limitation of the present study, which it shares with all or most other studies of auditory selective attention, relates to the difficulty of separating voice segregation and attention and hence determining the reason why participants listen to the masker voice.

To sum up, the pattern of results is generally consistent with the hypothesis that performance in the task depends on voice characteristics and spatialisation of the stimuli. Accordingly, the test method and stimuli used in this study appear to provide a tool for researchers to infer when and which voice is listened by a participant. This could be particularly useful in the context of neurophysiological studies of auditory selective attention to speech; for instance, taking into account information as to when the listener is listening the target, the masker, or neither could improve temporal response functions and bolster correlations between acoustic and neural speech envelopes.

5. Acknowledgements

The authors would like to thank Laurie Bazin, Lina Bourgeois, Lucie Vallet and Alice Watkowski for their help in running the experiment. This work was conducted in the framework of the LabEx CeLyA ("Centre Lyonnais d'Acoustique", ANR-10-LABX-60).

6. References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, 2012.

[3] N. Ding and J. Z. Simon, "Neural coding of continuous speech in auditory cortex during monaural and dichotic listening," *Journal of Neurophysiology*, vol. 107, no. 1, pp. 78–89, 2012.

[4] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, 2015.

[5] A. MacPherson and M. A. Akeroyd, "Variations in the slope of the psychometric functions for speech intelligibility: A systematic survey," *Trends in Hearing*, vol. 18, p. 2331216514537722, 2014.

[6] A. W. Bronkhorst, "The cocktail-party problem revisited: early processing and selection of multi-talker speech," *Attention, Perception, & Psychophysics*, vol. 77, no. 5, pp. 1465–1487, 2015.

[7] C. J. Darwin, D. S. Brungart, and B. D. Simpson, "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2913–2922, 2003.

[8] D. Başkent and E. Gaudrain, "Musician advantage for speech-on-speech perception," *The Journal of the Acoustical Society of America*, vol. 139, no. 3, pp. EL51–EL56, 2016.

[9] K. S. Helfer and R. L. Freyman, "Lexical and indexical cues in masking by competing speech," *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 447–456, 2009.

[10] G. Enders, *Le Charme discret de l'intestin [Audiobook]*, 2016. [Online]. Available: http://www.audiolib.fr/livre-audio/le-charme-discret-de-lintestin-9782367621029

[11] B. New, C. Pallier, L. Ferrand, and R. Matos, "Une base de données lexicales du français contemporain sur internet: LEXIQUE™//A lexical database for contemporary french: LEXIQUE™," *L'année psychologique*, vol. 101, no. 3, pp. 447–462, 2001.

[12] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[13] E. Gaudrain and D. Başkent, "Factors limiting vocal-tract length discrimination in cochlear implant simulations," *The Journal of the Acoustical Society of America*, vol. 137, no. 3, pp. 1298–1308, 2015.

[14] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *arXiv preprint arXiv:1406.5823*, 2014.

[15] H. Singmann, B. Bolker, J. Westfall, and F. Aust, *afex: Analysis of Factorial Experiments*, 2016. [Online]. Available: https://CRAN.R-project.org/package=afex

[16] D. Broadbent, "The role of auditory localization in attention and memory span." *Journal of Experimental Psychology*, vol. 47, no. 3, pp. 191–196, 1954.

[17] M. D. Vestergaard, D. T. Ives, and R. D. Patterson, "The advantage of spatial and vocal characteristics in the recognition of competing speech," *Proceedings of the International Symposium on Auditory and Audiological Research*, vol. 2, pp. 535–544, 2009.