INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Speech reductions cause a de-weighting of secondary acoustic cues.

Léo Varnet[1], *Fanny Meunier*[2], *Michel Hoen*[3]

[1] Institut d'Étude de la Cognition, Laboratoire des Systèmes Perceptifs (LSP), CNRS UMR 8248, École Normale Supérieure (ENS), Paris, France.
[2] Laboratoire Base, Corpus, Langage (BCL), CNRS UMR 7320, Nice Sophia Antipolis, France.
[3] Oticon Medical, 2720 Chemin Saint Bernard, Vallauris, France.

Abstract

The ability of the auditory system to change the perceptual weighting of acoustic cues when faced with degraded speech has long been evidenced. However, the exact changes that occur remain mostly unknown. Here, we proposed to use the Auditory Classification Image (ACI) methodology to reveal the acoustic cues used in natural speech comprehension and in reduced (i.e. noise-vocoded or re-synthesized) speech comprehension. The results show that in the latter case the auditory system updates its listening strategy by de-weighting secondary acoustic cues. Indeed, these are often weaker and thus more easily erased in adverse listening conditions. Furthermore our data suggests that this de-weighting does not directly depend on the actual reliability of the cues, but rather on the expected change in informativeness.

Index Terms: Auditory Classification Image (ACI), acoustic cues, phoneme categorization, speech-in-noise, noise-vocoded speech, synthetic speech.

1. Introduction

The identification of phonemes or syllables depends on multiple, redundant acoustic cues. In general, it is assumed that no unique acoustic cue is absolutely mandatory for correct perception, and that some cues are sufficient. The analysis of confusion matrices however suggests that all different cues are not equally robust to noise and degradations [1], [2]. On the perceptual side, listeners attempting to categorize speech signals can selectively attend to one dimension more than the other (i.e., allowing it a greater "weight") [3]. These observations raise the question of the extent to which we adapt cue-weighting in response to signal reductions.

The ability of the auditory system to change the perceptual weighting of acoustic cues has been evidenced in several psychoacoustics studies. It has been shown that primary cues in a given listening situation can become secondary cues in other cases. When categorizing frequency-modulated sinusoids [4] or complex tones [5] between two categories differing along two parameters, listeners are able, in some cases, to adapt their use of the cues depending on their relative reliability, even during the course of the experiment. A comparable result has been observed during speech-in-noise perception: while Voice Onset Time plays a major role in the perception of voicing in clear speech, the addition of noise makes this cue ambiguous. The listener then relies more on secondary cues such as f0 and formant transitions [6]. A similar re-weighting of the f0 cue is also observed when the cues are artificially made equally informative (i.e., the tokens are chosen to be at the same perceptual distance along two dimensions), in some cases [7]. The spectral distribution of masking noise [8], or artificial manipulations of the cues [9], [10] can also be a cause of weighting changes. In all, these experiments suggest that the auditory system is able to adapt its weighting strategy to selectively attend to the most reliable cues in a particular acoustic context. To our knowledge however, no study has directly explored the consequences of speech reductions on cue weighting strategies.

It is well established that noise-vocoded speech, an artificial manipulation resulting in a dramatic loss of spectral details, can remain highly intelligible after a short training [11], [12], suggesting that listeners are able to switch to the use of temporal cues in this case. Up to now, however, no study has directly tackle the exact changes in the weighting strategy induced by noise-vocoding. An increased understanding of the processes by which normal-hearing listeners understand noise-vocoded speech may have important applications for users of hearing aids or cochlear implants.

Synthetic speech is another type of speech reduction abundantly used in psychoacoustical studies as a "simplification" of the speech signal [13], [14]. It thus provides a means of controlling for the multiple covariant features in stimuli. On the other hand, it has not been properly ascertained that such synthetic speech is processed in the same way as natural speech by the auditory system, which could reduce the scope of synthetic speech studies [15].

The recent development of Auditory Classification Images (ACI), a new psychoacoustical tool for studying speech perception, has made it possible to precisely identify the acoustic cues involved during a phoneme categorization task [16]–[20]. In a typical ACI experiment, listeners perform a large number of phoneme categorizations in noise. The exact distribution of noise at a given trial is then used to predict the corresponding response of the participant. The statistical model involves the estimation of a decision template, the ACI. This template can be seen as psychoacoustical maps showing greater weights in regions which have greater influence on the participants' responses. Therefore, it has been informally presented as providing a direct visualization of the acoustic cues used by the listener.

Two previous ACI experiments conducted on normal-hearing participants have demonstrated the involvement of multiple acoustic cues during phonetic categorizations:

1) In the first application of this method, Varnet et al. used two /aba/ and /ada/ recordings as targets [17]. The resulting ACIs for the three listeners confirmed the role of the F2 onset in the /ba/-/da/ categorization, but also indicated the influence

http://dx.doi.org/10.21437/Interspeech.2016-343

of an additional anticipatory cue located on the F2 offset of the first syllable. In the following, this experiment will be referred to as “BD2013”.

2) A second study on /da/-/ga/ categorization in context /al-/ or /aʁ-/ yielded similar findings [18]. The ACIs estimated on a group of 16 listeners revealed that the main cue in this task was the height of F2 and F3 onsets, as suggested by previous investigations. However additional clusters of weights were found in the low frequency bands, indicating that the F1 onset also played a role in the categorization. Anticipatory cues were also found in the first syllable. In the following, this experiment will be referred to as “DG2015”.

In the present paper, we asked whether the secondary cues observed in these tasks are also used when the listener has to cope with reduced speech. To answer that question, we replicated the previous experiments using two artificial manipulations of stimuli: 22-bands noise-vocoding and re-synthesized speech. Two experiments are described here, “BD2016” and “DG2016”, following the same scenario as in BD2013 and DG2015 with noise-vocoded or re-synthesized stimuli, respectively.

2. Materials & Methods

The experimental setup for the two experiments was identical as that used in BD2013 and DG2015.

Participants were seated in front of a video monitor in a quiet room, wearing Sennheiser’s HD 448 headphones. On a given trial, they were presented with one of the possible target words at random, embedded in white noise, and then asked to select the word they heard via a textual MATLAB interface. No feedback was provided during the main test. Overall, each participant performed a total of 10.000 trials, for a total duration of 3 to 4 hours. For this reason, the listening test was broken into 4 days of experiment, each divided into 3 sessions lasting for about 15 min.

The SNR was automatically updated from trial to trial by means of an adaptive method to target a constant correct response rate (75% for /ba/-/da/ categorization experiments, 79% for /da/-/ga/ categorization experiments).

An additional practice session of ~10 trials was provided at the beginning of the experiment where the listener was given feedback on their responses.

/ba/-/da/ categorization experiment

One normal-hearing participant (LV, coauthor of this paper) from BD2013 took part in this experiment.

In BD2013, Targets were two disyllabic non-word speech sounds composed of a recording of /a/, identical for the two targets, followed by a recording of /ba/ or a recording of /da/ (equalized in duration and RMS normalized).

In experiment BD2016, noise-vocoded versions of the same targets were used. Noise-vocoded stimuli were created by dividing the frequency axis into 22 frequency bands logarithmically spaced. The amplitude envelope in each frequency band was applied to band-limited noise. The modulated noises were then summed to constitute the final sound. By construction, the first syllable is also identical between these two stimuli. Noise-vocoding on 22 channels is considered to be quite understandable without any learning phase.

The original and noise-vocoded stimuli are shown in Figure 1 (upper panel).

/da/-/ga/ categorization experiment

In DG2015, targets were 4 natural recordings of /alda/, /alga/, /aʁda/ and /aʁga/ (equalized in duration and RMS normalized). The task was to categorize the last syllable of the stimulus as /da/ or /ga/, independently of the preceding consonantal context.

The 4 stimuli were re-synthesized for experiment DG2016. Formant onsets and offsets, syllable durations and f0 movements were close to the original ones, but formant trajectories were made linear. Furthermore, in the synthetic stimuli the two syllables were acoustically independent (e.g. /al/ was identical in /alda/ and /alga/).

Formant trajectories in the original and synthetic speech stimuli are represented by lines in Figure 2.

10 participants took part to the experiment DG2016. For comparison purposes, only the results of the first 10 participants of DG2015 are considered here.

ACI derivation

Several methods have been developed for deriving ACIs. All data presented in this paper have been analyzed using the latest version of the algorithm [19].

During the experiment, the exact distribution of noise, target presented and participant's response are collected for each trial. By linking the time-frequency representation of stimuli (here, the cochleogram [21]) with the corresponding answer of the listener, via a Generalized Linear Model (GLM), a psychoacoustical map indicating the contribution of each spectrotemporal bin to the decision can be obtained. To limit the amount of over-fitting, the ACI is estimated via a Maximum A Posteriori algorithm with a smoothness prior. This method is very similar to that used for the estimation of Receptive Fields of sensory neurons from their firing rate [22]. For a more complete description of the algorithm, the reader is referred to [18], [19].

The obtained ACI is a linear approximation of the strategy used by the participant. It shows clusters of high positive or negative weights in regions where the presence of energy biases one response over the other. For instance, in Figure 1, red clusters correspond to regions favoring response “ba”, whereas blue clusters correspond to “da” regions. On the contrary, weights close to zero are time-frequency regions where noise has no marked influence on percept formation. Groups of neighboring positive and negative clusters mark the positions of acoustic cues. In the following, the ACIs are z-scored and all weights with $|Z|<2$ are plotted in gray for legibility.

3. Results

For comparison purposes, the SNR was varied during the experiment in order to maintain a constant percentage of correct answers across participants and conditions. The SNR levels and performances are given in Table 1.

	BD2013	BD2016	DG2015	DG2016
% correct	70.5	73.4	78.8 ± 0.4	76.8 ± 2.0
Mean SNR (dB)	-13.1	-13.0	-11.8 ± 0.7	3.3 ± 4.9

Table 1. Percentage of correct answers and mean SNR over all trials for the 4 experiments (mean and standard deviation over all participants the group experiments).

The two panels of Figure 1 show the two ACIs obtained for the same participant performing the /ba/-/da/ categorization task, with natural (A.) or noise-vocoded (B.) stimuli. Only one acoustic cue (composed of two positive and two negative clusters) is clearly identifiable in Figure 1B., around 0.2 s and 1500 Hz. It is also visible in Figure 1A., as well as two other cues: one located at the same frequency position but in the first syllable (around 0.05 s and 1500 Hz), and one low-frequency cue at the same time position (around 0.2 s and 800 Hz).

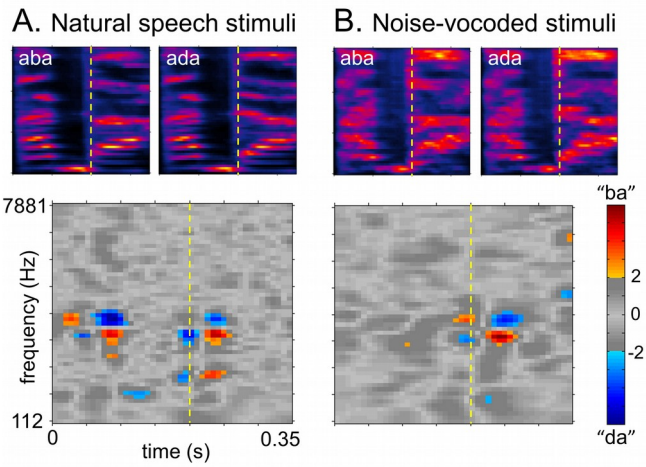

Figure 1: *Stimuli and ACI for participant LV in the two /ba/-/da/ categorization experiments. A. Experiment with natural speech stimuli (BD2013). B. Experiment with noise-vocoded stimuli (BD2016). ACIs are rendered as Z-score maps (colored pixels refer to |Z|>2). A dotted line indicates 0.2 s, i.e. approximately the beginning of the second syllable.*

Similarly, Figure 2 shows two ACIs obtained by averaging the ACIs of 10 participants performing the /da/-/ga/ categorization experiment, with natural (A.) or re-synthesized (B.) stimuli. For clarity, the formant trajectories of the targets are superimposed. Figure 2A and 2B both present a central cue on the F2 and F3 onsets in the second syllable (around 2000 Hz and 0.4 s). Additional cues are seen on Figure 2A. only, on the F1 onset of the second syllable and on the F1 and F2 in the first syllable.

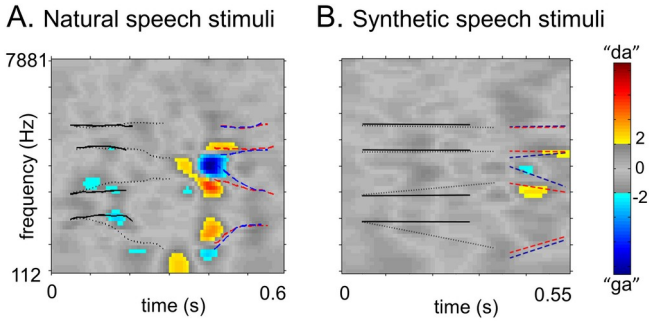

Figure 2: *Stimuli and mean ACI over 10 participants in the two /da/-/ga/ categorization experiments. A. Experiment with natural speech stimuli (DG2015). B. Experiment with synthetic speech stimuli (DG2016). Format as in Figure 1. Lines indicate formant positions in the stimuli (black dotted line: "al", black solid line: "ar", red line: "da", blue line: "ga").*

4. Discussion

In this paper, two previous ACI experiments (BD2013 and DG2015) with natural speech stimuli were replicated using reduced stimuli (BD2016 and DG2016).

The ACIs obtained in the original experiments (Figure 1A. and Figure 2A.) confirmed the acoustic cues identified by previous synthetic speech studies on these tasks (e.g. [23], [24]): the F2 onset serves as primary cue for the /ba/-/da/ categorization, whereas both F2 and F3 onsets are used in the /da/-/ga/ categorization (suggesting that the distance between the two is the primary cue in this case).

Interestingly however, the ACI method also revealed the presence of additional acoustic cues in these two experiments, the categorization of a speech signal is not performed on the basis of a single acoustic cue; rather, the covariance of multiple acoustical features allows for multiple cue extraction and combination, yielding more robust categorization. In both BD2013 and DG2015, we observed two types of secondary cues: anticipatory cues, located in the syllable preceding the target, and probably reflecting the extraction of coarticulatory information; and low-frequency cues on the F1 onset of the target syllable. It is important to note that these cues are visible on the ACI even in cases where they are not actually present in the stimuli. For example, the ACI in BD2013 shows an anticipatory cue on the F2 onset of the first syllable, although the targets do not contain any relevant information for the task, as they both begin with the same recording of /a/. A tentative explanation is that the auditory system, used to extract anticipatory information on the F2 offset of the syllable preceding /b/ or /d/, is highly sensitive to the variations introduced by noise on this region, resulting in large positive and negative weights on the ACI [20].

In the present study we investigated how artificial reduction of the speech signals affect the participants' listening strategy and the weighting of secondary cues. Two types of reductions were employed: noise-vocoding, reducing the number of spectral channels used to encode the signal, and re-synthesized speech, a simplification of the spectrotemporal details of the targets (such as exact formant trajectories). Re-synthesis appears to be more deleterious than 22-band noise vocoding, as it results in a much larger SNR threshold increase

(Table 1). However it should be noted that the participant enrolled in BD2013 and BD2016 is very used to ACI experiments, unlike those enrolled in DG2015 and DG2016.

The corresponding ACIs are shown in Figure 1B. and 2B., respectively. In both cases the primary cues for the task are preserved (although slightly delayed relatively to formant onsets in DG2016), confirming that participants are still processing the stimuli as speech signals. However the secondary cues are given less weights, comparatively, and do not reach the $|Z|>2$ threshold on the ACIs. These observations suggest that, when they have to cope with reduced speech signal, listeners selectively focus on the main cues by ignoring the secondary cues.

Previous experiments on cue weighting adaptation with speech and non-speech stimuli have proposed that the auditory system modulates its cue-weighting strategy to target the most reliable (i.e. less variable, less degraded) cues [4]–[6], [9]. This may account for changes observed in the listening strategy between DG2015 and DG2016. Indeed, while secondary cues may carry some coarticulatory information in the former case, in the latter they are likely to be uninformative, due to the re-synthesis. The listener is able to adapt to this loss of information by de-weighting the secondary cues.

Our results cannot be completely explained by this interpretation, however. As noted earlier, the coarticulatory information in both /ba/-/da/ categorization experiments is irrelevant for the task, by construction of the stimuli. Still, this cue is extracted in BD2013 but not in BD2016. This demonstrates that the weighting of one cue can be changed without varying its reliability, at least for this participant and this type of noise. One plausible explanation would be that the weighting strategy does not only depend on the objective, but also on the predicted, informativeness of the cues, according to a contextual model of intelligibility. The auditory system would generate a mental map of cue informativeness, depending on the listening conditions.

ACIs in Figure 2 are averaged over 10 participants. The absence of secondary cues in DG2016 could therefore be a result of an increase of the inter-individual variability, as often observed in degraded speech experiments. No participant show clear secondary cues in DG2016, but the individual ACIs also appear to be noisy, compared to DG2015. This may explain, at least partly, the decrease in the weighting of the primary cue. An other possible explanation would be that the more distant formant trajectories in the synthetic speech stimuli have resulted in broader (and hence weaker) clusters of weights.

For comparison purposes we have chosen here to equalize the performances across conditions by adapting the noise levels as a function of participants' responses. As a consequence, SNRs differed by 0,1 dB between BD2013 and BD2016, and by 15 dB between DG2015 and DG2016 (see Table 1). To put it another way, the difficulty of the task primarily results from the addition of noise in BD2013 and DG2015, but from the speech reduction in BD2016 and DG2016. This is obviously an inherent constraint of this type of studies. Nevertheless it is unlikely that the diminution of the level of noise in BD2016 and DG2016 has caused the observed reduction in the weighting of secondary cues.

The results of BD2016 remain to be replicated with more participants. Still, they already give an important insight on the processes by which users of hearing aids or cochlear implants adapt to their new input. While it was hypothesized that speech comprehension with a reduced number of channels implied switching form spectral cues to temporal cues [11], the ACI shows that, for 22-band noise-vocoded speech, the listener relies on the same primary acoustic cue as for natural speech. Synthetic speech has been used by speech researchers for decades [13], [14], and is still widely used today, in particular in neuroimaging studies [25], [26]. One major assumption underlying such studies is that reduced speech is processed in the same way as natural speech by the auditory system. However, the absence of secondary cues on the ACI in DG2016 calls into question this assumption.

5. Conclusion

To summarize, when it has to cope with reduced speech such as noise-vocoded speech or "rough" synthetic speech, the auditory system is able to update its listening strategy by de-weighting secondary acoustic cues and focusing on primary acoustic cues. Indeed, secondary cues are often weaker and thus more easily erased in adverse listening conditions. Furthermore our data suggests that this de-weighting does not directly depend on the actual reliability of the cues, but rather on the expected change in informativeness.

6. Acknowledgment

The authors would like to thank G. Trollé for running part of the DG2015 and DG2016 experiments. This research was partially supported by a European Research Council grant for the SpiN project (No. 209234) attributed to FM and by a public grant overseen by the French National Research Agency as part of the "Investissements d'Avenir" program (reference: ANR-10-LABX-0083).

7. References

[1] L. Varnet, J. Meyer, M. Hoen, and F. Meunier, "Phoneme resistance during speech-in-speech comprehension," in *Proceeding of Interspeech 2012*, 2012.

[2] J. R. Benkí, "Analysis of English nonsense syllable recognition in noise," *Phonetica*, vol. 60, no. 2, pp. 129–157, 2003.

[3] A. Lotto and L. Holt, "Psychology of auditory perception," *Wiley Interdiscip. Rev. Cogn. Sci.*, vol. 2, no. 5, pp. 479–489, 2011.

[4] L. L. Holt and A. J. Lotto, "Cue weighting in auditory categorization: implications for first and second language acquisition," *J. Acoust. Soc. Am.*, vol. 119, no. 5 Pt 1, pp. 3059–3071, 2006.

[5] M. Scharinger, B. Herrmann, T. Nierhaus, and J. Obleser, "Simultaneous EEG-fMRI brain signatures of auditory cue utilization," *Front. Neurosci.*, 2014.

[6] W. Serniclaes and Y. Arrouas, "Perception des traits phonétiques dans le bruit," *Verbum*, no. 2, pp. 131–144, 1995.

[7] A. L. Francis, N. Kaganovich, and C. Driscoll-Huber, "Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English," *J. Acoust. Soc. Am.*, vol. 124, no. 2, pp. 1234–1251, 2008.

[8] M. S. Régnier and J. B. Allen, "A method to identify noise-robust perceptual features: application for consonant /t/," *J. Acoust. Soc. Am.*, vol. 123, no. 5, pp. 2801–2814, 2008.

[9] A. L. Francis, K. Baldwin, and H. C. Nusbaum, "Effects of training on attention to acoustic cues," *Percept. Psychophys.*, vol. 62, no. 8, pp. 1668–1680, 2000.

[10] F. Li and J. B. Allen, "Manipulation of consonants in natural speech," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 19, no. 3, pp. 496–504, 2011.

[11] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, “Speech Recognition with Primarily Temporal Cues,” *Science*, vol. 270, no. 5234, pp. 303–304, 1995.

[12] L. Xu, C. S. Thompson, and B. E. Pfingst, “Relative contributions of spectral and temporal cues for phoneme recognition,” *J. Acoust. Soc. Am.*, vol. 117, no. 5, pp. 3255–3267, 2005.

[13] A. M. Liberman, H. D. Harris, H. S. Hoffman, and B. C. Griffith, “The discrimination of speech sounds within and across phoneme boundaries,” *J. Exp. Psychol.*, vol. 54, no. 5, pp. 358–368, 1957.

[14] P. Delattre, A. M. Liberman, F. S. Cooper, and L. J. Gerstman, “An experimental study of the acoustic determinants of vowel color; observations on one- and two-formant vowels synthesized from spectrographic patterns,” *Word*, vol. 8, no. 3, pp. 195–210, 1952.

[15] F. Li, A. Menon, and J. B. Allen, “A psychoacoustic method to find the perceptual cues of stop consonants in natural speech,” *J. Acoust. Soc. Am.*, vol. 127, no. 4, pp. 2599–2610, 2010.

[16] L. Varnet, K. Knoblauch, F. Meunier, and M. Hoen, “Show me what you listen to! Auditory classification images can reveal the processing of fine acoustic cues during speech categorization.,” in *Proceeding of Interspeech 2013*, 2013, pp. 3167–3171.

[17] L. Varnet, K. Knoblauch, F. Meunier, and M. Hoen, “Using auditory classification images for the identification of fine acoustic cues used in speech perception.,” *Front. Hum. Neurosci.*, vol. 7:, p. 865, 2013.

[18] L. Varnet, K. Knoblauch, W. Serniclaes, F. Meunier, and M. Hoen, “A Psychophysical Imaging Method Evidencing Auditory Cue Extraction during Speech Perception: A Group Analysis of Auditory Classification Images,” *PLoS ONE*, vol. 10, no. 3, p. e0118009, 2015.

[19] L. Varnet, T. Wang, C. Peter, F. Meunier, and M. Hoen, “How musical expertise shapes speech perception: evidence from auditory classification images,” *Sci. Rep.*, vol. 5, p. 14489, 2015.

[20] L. Varnet, “Identification of acoustic cues involved in degraded speech comprehension,” Theses, Université Claude Bernard - Lyon I, 2015.

[21] M. Slaney and R. F. Lyon, *Lyon’s cochlear model*. Apple Computer, Advanced Technology Group, 1988.

[22] M. C.-K. Wu, S. V. David, and J. L. Gallant, “Complete functional characterization of sensory neurons by system identification,” *Annu. Rev. Neurosci.*, vol. 29, pp. 477–505, 2006.

[23] N. Viswanathan, J. S. Magnuson, and C. A. Fowler, “Compensation for coarticulation: disentangling auditory and gestural theories of perception of coarticulatory effects in speech,” *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 36, no. 4, pp. 1005–1015, 2010.

[24] A. M. Liberman, P. C. Delattre, F. S. Cooper, and L. J. Gerstman, “The role of consonant-vowel transitions in the perception of the stop and nasal consonants,” *Psychol. Monogr. Gen. Appl.*, vol. 68, no. 8, pp. 1–13, 1954.

[25] M. W. Noordenbos, E. Segers, W. Serniclaes, and L. Verhoeven, “Neural evidence of the allophonic mode of speech perception in adults with dyslexia,” *Clin. Neurophysiol. Off. J. Int. Fed. Clin. Neurophysiol.*, vol. 124, no. 6, pp. 1151–1162, 2013.

[26] E. F. Chang, J. W. Rieger, K. Johnson, M. S. Berger, N. M. Barbaro, and R. T. Knight, “Categorical speech representation in human superior temporal gyrus,” *Nat. Neurosci.*, vol. 13, no. 11, pp. 1428–1432, 2010.