Interindividual Differences in Perceptual and Neural Processing of Continuous Speech

Jonathan Venezia

INTERINDIVIDUAL DIFFERENCES IN PERCEPTUAL AND NEURAL PROCESSING OF CONTINUOUS SPEECH

Jonathan H. Venezia[1,2]
[1] VA Loma Linda Healthcare System, Loma Linda, CA, United States
[2] Loma Linda University, Loma Linda, CA, United States
jonathan.venezia@va.gov

ABSTRACT

Speech-in-noise performance varies considerably among listeners with normal or near-normal audiometric thresholds and among hearing-impaired listeners even when the signal is amplified. Together, these findings suggest the existence of "suprathreshold" differences in speech processing. I have recently developed a procedure called Auditory Bubbles, which is designed to capture such differences by generating listener-specific classification images in the spectrotemporal modulation domain. These classification images are essentially "internal maps" of the acoustic speech patterns that drive changes in an outcome measure. In this brief review, I describe classification images produced for two outcome measures: keyword recognition/intelligibility (perceptual classification images) and physiological activity measured with blood-oxygen-level-dependent fMRI (neural classification images). I summarize the results of two recently-completed studies in which perceptual classification images were generated for hearing-impaired listeners, showing respectively that: (1) the amount of global noise in a perceptual classification image predicts a given listener's threshold for recognizing distorted speech, even after accounting for differences in audibility; and (2) variation in the local patterns of a perceptual classification image predict a given listener's ability to recognize speech in the presence of a competing talker, even when speech is amplified to compensate for differences in audibility. Finally, I introduce unpublished work examining individual differences in neural classification images for a competing speech task within a heterogeneous sample of listeners with near-normal audiometric thresholds.

1. INTRODUCTION

The ability to recognize speech in background noise varies widely across individuals, both within and between groups of listeners defined by age or hearing status [1, 2]. Even among young (< 40 years) listeners with normal hearing, word recognition in background noise varies considerably between individuals (percent-correct range ~ 30% at a fixed signal-to-noise ratio), and these differences are stable across multiple testing sessions [3]. Within groups of older listeners with or without hearing loss, speech reception thresholds in modulated background noise vary up to 10 dB [2], and in hearing-impaired listeners for whom speech has been amplified to achieve similar levels of audibility, word recognition performance in background noise varies across the entire psychometric range (0-100% correct) [4]. A simple but influential model of speech reception in background noise attributes interindividual variation to two factors [5]: at low levels of background noise, differences in speech reception are attributable to differences in audibility plus a second factor referred to as 'distortion'; at high (suprathreshold) levels of background noise when speech is clearly audible, differences in speech reception are driven entirely by the distortion factor. Initially, distortion was taken to reflect changes in peripheral auditory encoding such as loss of temporal or spectral resolution, but more recent models imply that distortion can be understood as a "catch-all" category reflecting differences in peripheral and central auditory processing and the interaction of these systems with top-down cognitive processes [6].

Two broad approaches have been taken to characterize the contribution of suprathreshold distortion to individual differences in speech reception: a "correlation approach" in which various ancillary measures of auditory and cognitive processing are used to predict speech recognition in background noise [7], and a "simulation approach" in which the speech signal itself is acoustically distorted along different dimensions to selectively modulate the auditory-perceptual and/or cognitive systems involved in speech recognition [8, 9]. While both approaches have borne fruit, research investigating the different aspects of distortion has failed to pin down a particular mechanism or combination of mechanisms that is consistently able to explain individual differences in speech recognition [7]. Recently, Bernstein [10] has developed a promising framework based on the correlation approach that attributes differences in suprathreshold speech recognition to differences in the ability to process spectrotemporal modulations (STMs) in the range of temporal fluctuations (4-12 Hz) and frequencies (< 2 kHz) that predominate in speech. Indeed, STMs have proven useful in the characterization of complex acoustic patterns in speech and other natural sounds [11], and in the characterization of high-level (e.g., cortical) representations of those sounds in the auditory nervous system [12].

In this short review, I introduce a psychoacoustical technique based on the simulation approach called Auditory Bubbles, which is designed to capture differences in suprathreshold processing of continuous speech in terms of its component STMs. Briefly, acoustic distortion is applied to a set of speech signals (typically sentences) by filtering out randomly-selected segments of

each sentence's two-dimensional modulation power spectrum (MPS). A classification image (CImg) is produced by using the filter patterns to predict the effects of the resultant, idiosyncratic acoustic distortion on behavioral performance (e.g., keyword recognition). In recent work [13, 14], I have shown that listener-specific properties of these CImgs can explain individual differences in speech recognition. I have also extended Auditory Bubbles to the neurophysiological domain using functional magnetic resonance imaging (fMRI) [15], with an eye toward capturing how cortical speech processing networks may contribute to individual differences in speech recognition. Here, I summarize some of the key results from this work.

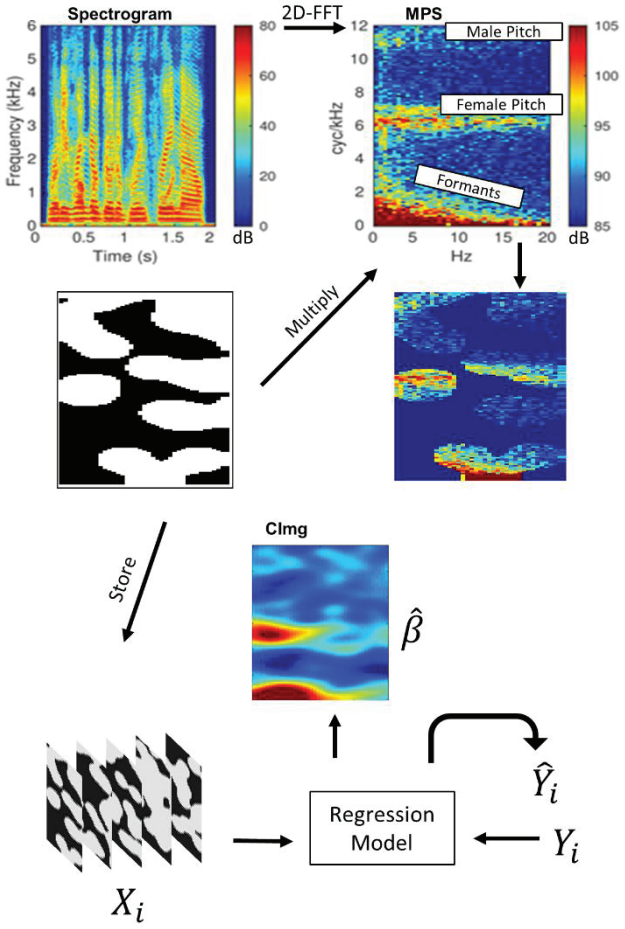

Figure 1. Auditory bubbles schematic: (1) obtain log-magnitude spectrogram with Gaussian windows (top left); (2) obtain log-magnitude MPS via two-dimensional Fourier transform of the spectrogram; (3) remove randomly-selected segments of the MPS using a multiplicative bubbles filter (mid left); (4) recover waveform from the filtered MPS (mid right) via inverse Fourier transform and spectrogram inversion [16]; (5) store bubbles filters, X_i, and behavioral responses, Y_i, across trials; (6) obtain CImg, $\hat{\beta}$, and predicted behavioral responses, $\hat{Y}_i$, via regression (bottom).

2. AUDITORY BUBBLES: TECHNIQUE

The Auditory Bubbles procedure receives its name from an analogous procedure developed in visual psychophysics [17]. In the visual procedure, an image (e.g., of a human face) is overlayed with an opaque noise masker. The masker is "pierced" with a number of transparent Gaussian apertures to allow partial information about the underlying image through to the observer. The observer is asked to do a simple behavioral task using this partial information (e.g., identify the gender of the face). From trial to trial, the location of the Gaussian apertures is varied at random. On each trial, the particular masker from that trial is stored along with a record of behavioral performance (e.g., correctness of gender identification judgments). At the end of the experiment, the noise maskers are sorted according to performance (correct vs. incorrect) and summed, similar to a spike-triggered average in unit neurophysiology. The result is a CImg: essentially a "heat map" showing the underlying visual features that contributed to successful performance of the task (e.g., those facial features required for gender identification).

In Auditory Bubbles (Fig. 1), the procedure is similar except that the bubbles "masker" is a multiplicative filter applied to the MPS of the speech signal. In the example shown in Fig. 1, the speech signal is a mixture of two sentences spoken by female and male talkers, respectively. The MPS, which decomposes the spectrogram in terms of its component STMs, plots coarse (fine) spectrotemporal patterns near (far from) the origin; temporal (spectral) patterns are plotted in terms of modulation rate in Hz (cyc/kHz) along the abscissa (ordinate). Speech energy clusters in three regions along the ordinate: vocal harmonics associated with the male talker (male pitch) are represented at the highest spectral modulation rates, vocal harmonics associated with the female talker (female pitch) are represented at mid spectral modulation rates, and the shared phonetic content of the two talkers (formants) is represented at the lowest spectral modulation rates. In our example, the task is the Coordinate Response Measure (CRM) [18], a matrix-style test in which the listener must identify a color-number combination spoken by one of the talkers (here the female) while ignoring a competing combination spoken by the other talker (here the male). An example sentence is "Ready baron go to green two now," where the target is "green two." Crucially, the task is performed only after the MPS of the two-talker mixture is filtered via bubbles, such that the listener must generate a response using only a subset of the STM patterns from the original signal. As in the visual procedure, the idea is that certain STM patterns – characterized from trial to trial by the shape of the bubbles filter – should reliably produce good task performance. A unique bubbles filter is applied on each trial and stored along with the behavioral response. As in the visual procedure, a CImg is obtained by comparing the bubbles filters from correct trials to those from incorrect trials. As shown in Fig. 1, for the CRM task the resultant CImg places most of the "perceptual weight" on the vocal harmonics of the target (female) talker and on the shared phonetic content of the two talkers ($\hat{\beta}$, warm colors).

In practice, the CImg is obtained by entering the bubbles filters as predictors (with each pixel in the filter representing a different feature) of trial-by-trial behavior in a regression model. A typical approach is to use the generalized linear model wherein the behavioral responses Y_i are taken to

be drawn from one of the exponential family of probability distributions, and the mean of the distribution, $E(Y) = \mu$, is modeled as:

$$\hat{u}_i = g^{-1}(X_i^T \beta), \qquad (1)$$

where g is the link function, X_i is the (vectorized) bubbles filters across trials, and β is a parameter vector of regression weights that must be estimated from the data. The estimated parameter vector, $\hat{\beta}$, is the CImg (when reshaped to two dimensions). In fact, since a high degree of collinearity exists among the X variables, some degree of regularization must be applied to β. Thus, the model must be tuned to optimize one or more regularizing hyperparameters using cross-validation [13, 14]. Moreover, a subset of the data is left aside to assess the model's ability to predict behavioral responses (i.e, generate $\hat{Y}_i$) on trials that were not used to tune the model. If, instead of behavioral responses, the goal is to model a continuous neurophysiological time series such as in fMRI, the regression model takes the form:

$$\hat{Y}_t = \sum_n^N \sum_\tau^T h_{n,\tau} * X_{n,t-\tau}, \qquad (2)$$

where the blood-oxygen level dependent (BOLD) signal across time, $\hat{Y}_t$, is modeled as the convolution of n predictors (pixels in the bubbles filter), X_n, with an n-dimensional kernel, h_n, over a restricted set of time lags, τ. Here, the CImg, h_n, is in fact a "classification volume" or a series of CImgs over unfolding over the lags in τ. In neurophysiology, this is referred to as a spectrotemporal receptive field (STRF). Again, cross-validation is applied to regularize h_n and provide independent data to test the model predictions. As such, the regression models in Eqs. 1 and 2 yield two crucial outputs: the CImg or STRF, and a set of unbiased model predictions from held-out trials that can be used to assess the model's ability to predict a listener's behavior or neural response on unseen data. A third crucial parameter in Auditory Bubbles is the number of bubbles applied to the speech signal on each trial. In general, fewer bubbles result in poorer performance (less acoustic information and therefore more distortion); in practice, the number of bubbles is calibrated individually to each listener using an adaptive procedure to target a pre-defined level of performance [19]. Threshold performance is defined as the average number of bubbles (or, alternatively, the proportion of total STM content remaining in the MPS) required to achieve this pre-defined performance level (lower = better performance).

3. AUDITORY BUBBLES: BEHAVIORAL RESULTS

In the first study to use Auditory Bubbles [19], 10 young, normal-hearing (YNH) listeners were assessed on keyword recognition for IEEE sentences [20] spoken by a single female talker in a quiet background. CImgs were generated for each listener from approximately 450 trials using the procedure described above (Eq. 1). The CImg pattern was highly reliable across listeners: maximal weight was placed on STMs < 2 cyc/kHz in the spectral modulation domain and 2-10 Hz in the temporal modulation domain. The temporal modulation peak was at 4 Hz, roughly equal to the peak of the third-octave spectrum for the speech envelope. To test whether CImg patterns were attributable entirely to the underlying envelope spectrum, a second set of CImgs were obtained from different listeners using IEEE sentences that had been 50% time-compressed, thus producing an octave shift in the envelope spectrum (peak at 8 Hz). In fact, the peak of these time-compressed CImgs did shift up along the temporal modulation axis, but only by one-third of an octave. This suggests that CImg weights reflect perception and not just the physical characteristics of the speech signal.

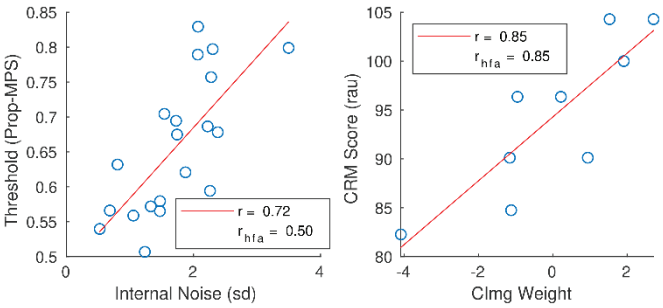

Figure 2. Left: internal noise estimated from CImgs in the IEEE bubbles task (standard normal units; abscissa) plotted against threshold performance on the task (average proportion STMs revealed to the listener at 50% correct; ordinate). Right: compositive weight derived from CImgs in the CRM bubbles task (arbitrary units; abscissa) plotted against CRM score on clean trials (rationalized arcsine units; ordinate). All listeners OHI. Adapted with permission from [13, 14].

With this as a proof of concept, a follow-up study used the same paradigm (without time compression) to obtain CImgs from groups of 10 YNH listeners, 19 older, normal-hearing (ONH) listeners, and 20 older, hearing-impaired (OHI) listeners [14]. The idea was to test whether CImgs would be different within or between groups, presumably reflecting individual differences in suprathreshold distortion of speech due to aging and/or hearing loss. Surprisingly, no such differences were observed between the groups. The mean CImg from each group faithfully replicated the pattern observed in the original study [19]. The CImg weights in the OHI group were found to be correlated significantly with high-frequency audiometric thresholds, but this could not be conclusively attributed to suprathreshold aspects of hearing loss. However, CImg regression-model predictions were also found to be poorer in OHI listeners compared to the other groups, suggesting a less reliable relation between bubbles-based acoustic distortion and behavior in these listeners. Indeed, the CImgs themselves were found to be more variable in OHI listeners, exhibiting different amounts of

idiosyncratic "distortion" of the broader CImg pattern in different listeners. These deviations were characterized in terms of a single distortion factor (D-NOISE). In Fig. 2 (left panel), D-NOISE is transformed into units of internal noise (standard normal) using a signal detection model and plotted against threshold performance (average proportion of STMs retained in the MPS) on the bubbles task for OHI listeners. Indeed, a strong correlation is observed ($r = 0.72$) and remains significant after accounting for variance explained by high-frequency audiometric thresholds (partial correlation: $r_{hfa} = 0.5$, $p < 0.03$). This suggests that CImgs are sensitive to an aspect of suprathreshold distortion in OHI listeners that reflects not *which* STM patterns are encoded, but how efficiently those patterns can be used by a given listener to support speech recognition.

One possible explanation for the lack of listener-group differences in mean CImg patterns is that speech recognition in quiet is not sufficiently challenging to produce such differences. Therefore, in a second follow-up study [13], groups of 10 ONH listeners and 10 OHI listeners were tested in Auditory Bubbles using the two-talker CRM task described in Section 2. Here, the ability to recognize speech in a background of competing speech was tested at a reasonably challenging signal-to-noise ratio (+ 3 dB). Crucially, the speech signals presented to OHI listeners were amplified using an individualized gain profile to compensate for differences in audibility. CImgs were obtained from 800 CRM trials for each listener. Additionally, each listener performed 50 "clean" trials without the addition of bubbles. This time, significant differences in mean CImg patterns were observed between the ONH and OHI groups. In particular, compared to ONH listeners, OHI listeners showed reduced weight on phonetic speech content modulated at 5-10 Hz, indicating less efficient encoding of these modulation rates, and negative weight on the competing (male) talker's harmonics, indicating relatively more competing-talker interference experienced by OHI listeners. In Figure 2 (right panel), a composite measure reflecting the overall CImg weight across these two regions (obtained via principal component analysis) is plotted against performance on clean CRM trials (rationalized arcsine units) for OHI listeners. A strong correlation was observed ($r = 0.85$) and did not diminish after accounting for variance explained by high-frequency audiometric thresholds (partial correlation: $r_{hfa} = 0.85$, $p < 0.01$). These findings confirm that CImg weights reflect not just idiosyncratic [14] but also systematic distortion of perceptual representations for continuous speech whose magnitude predicts the ability of an individual listener to recognize (undistorted) speech in background noise.

4. AUDITORY BUBBLES: FMRI RESULTS

In parallel with these behavioral studies, a study was undertaken to extend Auditory Bubbles into the neurophysiological domain with fMRI [15]. The goal was to characterize how neural CImgs (or STRFs) vary across regions of the auditory cortex. Ten YNH listeners performed the IEEE bubbles task while undergoing fMRI scanning. A STRF was estimated for each voxel and second-level (group-average) STRFs were obtained by calculating a t-score for each pixel in the STRF. An unsupervised learning procedure was then carried out to cluster all of the significantly tuned second-level STRFs (those with sufficiently large t-scores) according to their response patterns. In fact, four groups of STRFs were identified: a primary-like group that responded to vocal harmonics and phonetic speech content across a wide range of modulation rates, and high-, mid-, and low-modulation rate groups that responded to phonetic content modulated at approximately 8, 6, and 4 Hz, respectively. The primary-like group localized to the core auditory cortex and surrounds, while the high-modulation rate group localized to the posterior superior temporal plane, and the mid- and low-modulation rate groups localized to classic speech areas in the bilateral superior temporal gyrus (STG) and sulcus (STS). Importantly, this hierarchical organization of STRFs across the auditory cortex was highly reliable across individuals, as determined by applying the clustering algorithm to each listener's data individually and shuffling the cluster labels to maximize functional overlap with the group clusters. This suggests that STRFs can be reliably estimated using Auditory Bubbles at the level of individual listeners.

In my ongoing work (unpublished at the time of this writing), Auditory Bubbles is used with fMRI to estimate STRFs for a modified, three-alternative forced choice (3-AFC) version of the two-talker CRM bubbles task in listeners with near-normal audiometric thresholds. At present, data have been obtained from 25 Veterans of the United States Armed Forces (22 male, 3 female) whose average age is 49.4 years (SD = 9.2 years, range = 30-60 years), and who report varying degrees of subjective difficulty understanding speech in background noise. Each listener performs 400 trials in each of two conditions: the typical task, wherein the target (female) and competing (male) talkers utter different color-number combinations, which we term "Competing"; and a control task in which the two talkers utter the exact same sentence in perfect harmony (essentially replicating a single-talker scenario), which we term "Unison." STRFs are estimated at each cortical voxel in each condition (Eq. 2) and used to predict the brain response in held-out data. One crucial aim is to determine which cortical regions will yield significantly different STRF-based predictions for Unison vs. Competing. In fact, significant differences are observed in three cortical networks: a bilateral speech network in the STG, a bilateral sensorimotor network in the inferior frontal and parietal lobes, and a semantic memory network in the left superior frontal and inferior parietal lobe and bilateral midline regions. From each of these networks, a 'beta time series' was estimated for

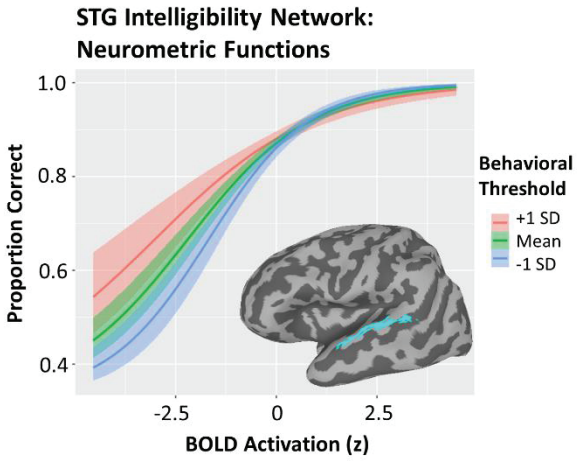

Figure 3. Neuometric functions in the STG network (inset). Normalized BOLD activation (abscissa) is plotted against proportion-correct performance on the 3-AFC CRM task (ordinate). Bold lines are posterior medians and shaded regions are 80% posterior credible intervals. Three hypothetical listeners are plotted: mean threshold performance (green), and 1 SD above (red) and below (blue) mean threshold performance.

each listener representing fluctuations in the overall BOLD amplitude from trial to trial. These were entered as predictors in a generalized linear model (Eq. 1) to explain trial-by-trial behavior in the 3-AFC task (correct vs. incorrect). The output of the model is a 'neurometric' function relating BOLD activation to performance across trials. Threshold task performance (number of bubbles required to achieve 67% correct) was entered as a mediating between-subjects predictor in the model. Although activation was significantly associated with behavior in the sensorimotor and STG networks, only in the STG network were neurometric functions significantly modulated by threshold performance across listeners. As shown in Fig. 3, neurometric functions tended to be steeper in those listeners who performed best on the task, suggesting that individual differences in STG STRFs may serve as a biomarker or cortical "readout" of distortions in the ascending auditory system. Alternatively, STG STRFs could be modulated by top-down systems differentially across listeners. Interestingly, STG STRFs tend to strongly resemble behavioral CImgs for speech recognition [15].

5. DISCUSSION

To summarize, Auditory Bubbles can be used to produce CImgs (STRFs) that reflect perceptual (neural) representations of STM patterns in continuous speech. These STMs are hypothesized to be crucial elements of speech whose processing and representation in the central auditory nervous system may be responsible for the significant interindividual differences observed for speech-in-noise recognition at suprathreshold levels. Indeed, I have shown that individual differences in CImg patterns ("suprathreshold distortion") can explain differences in recognition of distorted speech and speech in background noise. Moreover, cortical STRFs derived from Auditory Bubbles in fMRI reveal how speech is represented across multiple, hierarchical processing stages, each of which may be sensitive to suprathreshold distortion from different sources. Ongoing and future work will aim to leverage the sensitivity of Auditory Bubbles to improve our mechanistic understanding of individual differences in auditory processing related to aging, hearing loss, and brain injury.

One significant weakness of Auditory Bubbles is that the regression models at its core (Eqs. 1 and 2) are sensitive to any mechanism (auditory, cognitive, etc.) whose function may be correlated with the presence or absence of STMs in the speech signal. Thus, Auditory Bubbles alone cannot provide a complete mechanistic description of suprathreshold distortion. In my ongoing work, I collect a large battery of audiological, psychoacoustic, cognitive, and mental health measures to help drill down on the sources of individual differences in the CImgs and STRFs derived from Auditory Bubbles. However, in the long term, a more promising approach may be to combine process-specific models with Auditory Bubbles. For example, acoustic speech features could be "pre-processed" through an auditory model prior to the bubbles regression step, perhaps tuning parameters of the auditory model to best predict data from hearing impaired vs. normal hearing listeners [21]; additionally, STRFs could be explicitly modulated by listener-level variables during model fitting to improve brain predictions across listeners [22].

6. ACKNOWLEDGEMENTS

This work was supported by an American Speech-Language-Hearing Foundation New Investigators Research Grant to J.H.V. and by the U.S. Department of Veterans Affairs, Veterans Health Administration, Rehabilitation Research & Development Service Award IK2RX002702 to J.H.V. This material is the result of work supported with resources from and the use of facilities at the VA Loma Linda Healthcare System, Loma Linda, CA. The contents do not represent the views of the U.S. Department of Veterans Affairs or the U.S. government.

7. REFERENCES

[1] C. Fullgrabe, B. C. Moore, and M. A. Stone, "Age-group differences in speech identification despite matched audiometrically normal hearing: contributions from auditory temporal processing and cognition," *Front. Aging Neurosci.,* vol. 6, p. 347, 2014, doi: 10.3389/fnagi.2014.00347.

[2] E. L. J. George, A. A. Zekveld, S. E. Kramer, S. T. Goverts, J. M. Festen, and T. Houtgast, "Auditory and nonauditory factors affecting speech reception in noise by older listeners," *The Journal of the*

Acoustical Society of America, vol. 121, no. 4, pp. 2362-2375, 2007.

[3] K. M. Carbonell, "Reliability of individual differences in degraded speech perception," *The Journal of the Acoustical Society of America,* vol. 142, no. 5, pp. EL461-EL466, 2017.

[4] L. E. Humes, "Factors underlying the speech-recognition performance of elderly hearing-aid wearers," *The Journal of the Acoustical Society of America,* vol. 112, no. 3, pp. 1112-1132, 2002.

[5] R. Plomp, "A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired," *J. Speech Hear. Res.,* vol. 29, no. 2, pp. 146-154, 1986.

[6] M. K. Pichora-Fuller *et al.*, "Hearing Impairment and Cognitive Energy: The Framework for Understanding Effortful Listening (FUEL)," *Ear Hear.,* vol. 37 Suppl 1, pp. 5S-27S, Jul-Aug 2016, doi: 10.1097/AUD.0000000000000312.

[7] T. Houtgast and J. M. Festen, "On the auditory and cognitive functions that may explain an individual's elevation of the speech reception threshold in noise," *Int. J. Audiol.,* vol. 47, no. 6, pp. 287-295, 2008.

[8] R. Drullman, J. M. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *J. Acoust. Soc. Am.,* vol. 95, no. 5 Pt 1, pp. 2670-80, May 1994, doi: 10.1121/1.409836.

[9] M. Ter Keurs, J. M. Festen, and R. Plomp, "Effect of spectral envelope smearing on speech reception. II," *The Journal of the Acoustical Society of America,* vol. 93, no. 3, pp. 1547-1552, 1993.

[10] J. G. Bernstein, "Spectrotemporal modulation sensitivity as a predictor of speech intelligibility in noise for hearing-impaired listeners," *The Journal of the Acoustical Society of America,* vol. 139, no. 4, pp. 2121-2121, 2016.

[11] N. C. Singh and F. E. Theunissen, "Modulation spectra of natural sounds and ethological theories of auditory processing," *The Journal of the Acoustical Society of America,* vol. 114, p. 3394, 2003.

[12] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America,* vol. 118, no. 2, pp. 887-906, 2005.

[13] J. H. Venezia, M. R. Leek, and M. P. Lindeman, "Suprathreshold Differences in Competing Speech Perception in Older Listeners With Normal and Impaired Hearing," *J. Speech. Lang. Hear. Res.,* vol. 63, no. 7, pp. 2141-2161, Jul 20 2020, doi: 10.1044/2020_JSLHR-19-00324.

[14] J. H. Venezia, A.-G. Martin, G. Hickok, and V. M. Richards, "Identification of the Spectrotemporal Modulations That Support Speech Intelligibility in Hearing-Impaired and Normal-Hearing Listeners," *J. Speech. Lang. Hear. Res.,* vol. 62, no. 4, pp. 1051-1067, 2019.

[15] J. H. Venezia, S. M. Thurman, V. M. Richards, and G. Hickok, "Hierarchy of speech-driven spectrotemporal receptive fields in human auditory cortex," *Neuroimage,* vol. 186, pp. 647-666, Feb 1 2019, doi: 10.1016/j.neuroimage.2018.11.049.

[16] Z. Průša, P. Balazs, and P. L. Søndergaard, "A noniterative method for reconstruction of phase from STFT magnitude," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 25, no. 5, pp. 1154-1164, 2017.

[17] F. Gosselin and P. G. Schyns, "Bubbles: a technique to reveal the use of information in recognition tasks," *Vision Res.,* vol. 41, no. 17, pp. 2261-2271, 2001.

[18] R. S. Bolia, W. T. Nelson, M. A. Ericson, and B. D. Simpson, "A speech corpus for multitalker communications research," *J. Acoust. Soc. Am.,* vol. 107, no. 2, pp. 1065-6, Feb 2000, doi: 10.1121/1.428288.

[19] J. H. Venezia, G. Hickok, and V. M. Richards, "Auditory "bubbles": Efficient classification of the spectrotemporal modulations essential for speech intelligibility," *J. Acoust. Soc. Am.,* vol. 140, no. 2, p. 1072, Aug 2016, doi: 10.1121/1.4960544.

[20] I. S. o. S. Measurements, "IEEE Recommended Practices for Speech Quality Measurements," *IEEE Transactions on Audio and Electroacoustics,* vol. 17, pp. 227-246, 1969.

[21] E. Ponsot *et al.*, "Mechanisms of spectrotemporal modulation detection for normal-and hearing-impaired listeners," *bioRxiv,* 2020.

[22] S. V. David, "Incorporating behavioral and sensory context into spectro-temporal models of auditory encoding," *Hear. Res.,* vol. 360, pp. 107-123, 2018.