INFORMATION-BEARING COMPONENTS OF SPEECH INTELLIGIBILITY UNDER BABBLE-NOISE AND BANDLIMITING DISTORTIONS

Mounya Elhilali[*]

[*]Dept. of Electrical & Computer Engineering
Johns Hopkins University, Baltimore

Shihab Shamma[†]

[†]Dept. of Electrical & Computer Engineering
University of Maryland, College Park

ABSTRACT

Performance of speech technologies can benefit greatly from a deeper appreciation of the nature of the information-bearing features in continuous speech. To explore these features, we focus here on the role of the spectral and temporal modulations in maintaining the intelligibility of speech as it becomes severely degraded by low-pass filtering and additive babble noise. These modulations are estimated using a biological model of auditory processing which approximates the representation of sound in the cortex. Intelligibility of the noisy speech is computed directly from this model via the Spectro-Temporal Modulation Index (STMI) [1], and the validity of this metric is confirmed by a detailed comparison with results of psychoacoustic tests. Our analysis reveals quantitatively why certain types of noise are more disruptive to speech intelligibility than others (e.g., babble vs. white noise). It also highlights the important contribution of *both* spectral and temporal modulations in accurately predicting the intelligibility of speech under adverse conditions.

Index Terms— Speech intelligibility, STMI, auditory system, bandlimited speech, babble noise

1. INTRODUCTION

A ubiquitous operation at the front-end of all speech technologies (recognition systems, coding schemes, hearing prostheses, etc) is the extraction of acoustic elements that are most informative about speech for the task at hand. However, this process poses a real challenge to all these systems due to the inherent high redundancy in continuous speech, making it difficult to delineate which components of the acoustic signal carry what information about its attributes. This challenge is made even more strenuous when dealing with speech signals in the presence of background noise and channel distortions.

Much inspiration about these information-bearing features of speech can be gained from knowledge of how the brain processes and perceives sounds. In the current study, we use a biologically-inspired model of auditory processing to investigate the contribution of different spectro-temporal elements of speech in determining its intelligibility. Within this context, we explore how gracefully do these features degrade under conditions of background babble noise and bandlimiting distortions, and how informative are these degradations about the actual intelligibility of speech.

This work is supported by contract with the Southwest Research Institute. We are grateful to Ms. Diana Strickland and Dr. Brian Zook for extensive discussions and supplying the speech intelligibility data.

2. APPROACH

Speech in its journey from the eardrum to the cortex undergoes profound transformations from a simple one dimensional pressure waveform to an elaborate multidimensional representation; thereby robustly encoding specific acoustic features in different nuclei in the auditory system [2]. To better investigate the contribution of each of these acoustic elements to speech perception, we use a biologically-inspired model which mimics the signal processing taking place along the pathway from the periphery all the way to the cortex [3]. In earlier work, we showed the fidelity of this model in encoding speech features, and proposed a metric (STMI, Spectro-Temporal Modulation Index) which proved to be a robust predictor of intelligibility under a variety of noise conditions including severe and nonlinear distortions [1], as well as an effective framework for developing strategies for directional microphone modes in hearing-aid circuits [4].

Briefly, the model starts with an early stage, where the acoustic signal is transformed into an 'auditory spectrogram' - a time-frequency representation that is the end-result of frequency analysis in the cochlea, followed by edge detection and temporal smoothing. A subsequent central stage analyzes the modulation profile of the signal along the spectral and temporal dimensions. This analysis is performed via a bank of 'modulation selective filters', mimicking the array of feature selective filters (called Spectro-Temporal Receptive Fields, STRF) observed at the level of the primary auditory cortex (A1) [2]. A1 contains a large variety of STRFs, each 'tuned' to a particular pattern of spectral peaks, temporal rates, and tonotopic frequencies. In the model, an ordered bank of such multi-resolution filters tuned to a range of bandwidths and dynamic rates provides a unique characterization of sounds, one that is sensitive to the spectral shape and temporal dynamics over the entire stimulus (Fig. 1). Mathematically, this 'cortical' mapping is performed via a two-dimensional wavelet analysis of the spectrogram (see [1, 3] for further details).

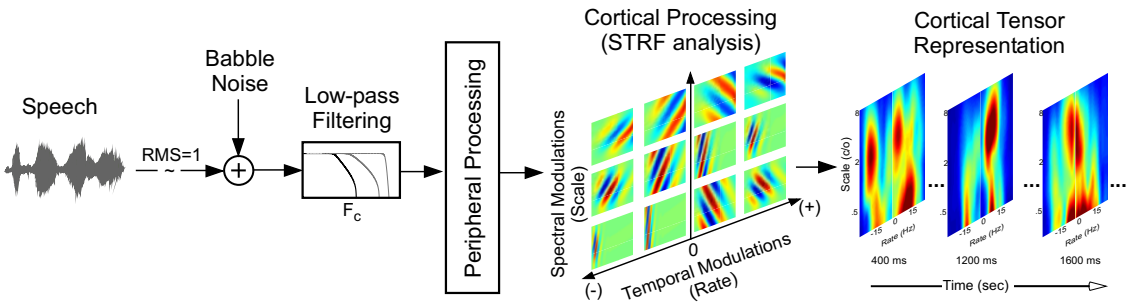

Fig. 1. Schematic of channel distortion followed by model of auditory processing. Each speech signal is contaminated with a constant level of babble noise, then low-pass filtered at varying cutoff frequencies F_c. The 'noisy' signal is then processed via a model of the auditory periphery, followed by a cortical analysis through an array of modulation-selective "filters", (STRFs).

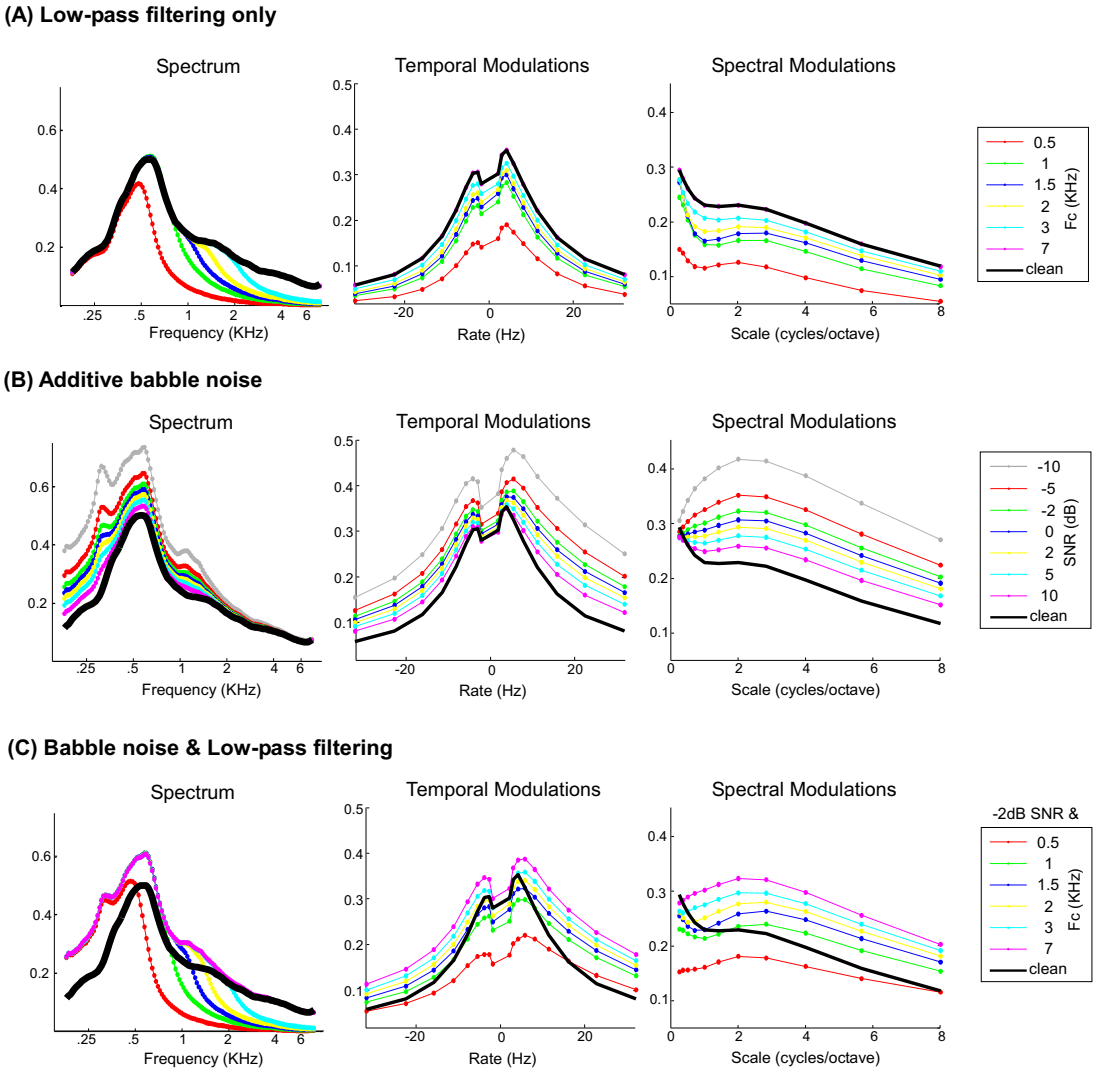

Fig. 2. Distortion of the frequency spectrum (right column), temporal modulations (middle) and spectral modulations (left) under conditions of low-pass filtering, babble noise and both combined. Each curve represents the average profile of 25 sentences.

3. EXPERIMENTAL METHODS

Speech material: 25 utterances were extracted from the TIMIT speech corpus, and consisted of spoken sentences by male and female speakers across different American dialect regions (sampled at 16 KHz with 16-bit resolution). A masking babble noise was added to each sentence (at -2 dB), simulating the effect of a crowded cafeteria or a cocktail party. A further bandlimiting distortion was applied by passing the noisy signal through a low-pass, Chebyshev II, 10^{th} order filter with 96 dB stopband attenuation. Six cutoff frequencies were tested: 0.5, 1, 1.5, 2, 3 and 7 KHz.

Psychophysical procedure: Data was provided by the Southwest Research Institute based on speech intelligibility tests conducted on thirteen subjects under the same bandlimiting and babble noise conditions mentioned above. Each test included a total of eight nonsense sentences of a male speaker, each constructed from five randomly chosen monosyllabic words. Each word consisted of three phonemes with approximately balanced presentations of vowels and consonants. A count of correct phonemes reported was then averaged across all listeners and all test material for each condition.

Model analysis: Each speech waveform was normalized to an RMS of one, processed through the distortion channel, then analyzed through the auditory model (as in [1]) (Fig. 1). The output was then contrasted (via a linear L2 distance) with a 'clean' template constructed by averaging the output of about 60 seconds of clean speech, yielding an STMI value between 0 and 1. Due to the inevitable mismatch between the tested utterance and the 'generic' templates, we mapped the STMI values through a sigmoidal nonlinearity as discussed in [1]. We then converted the STMI values into percentage scores; and derived the angle (i.e. inverse tangent) of the slope that linearly fits the model vs. subject scores. We confirmed the statistical significance of this correspondence by performing a bootstrap procedure, which consisted of randomly choosing 13 STMI values (out of the 25 computed for each condition), and correlating them with the behavioral scores from each subject (1000 iterations). The across subject slopes were then combined using circular statistics to yield an angular mean. Confidence measures were derived from the bootstrap statistics.

4. DEGRADATION OF SPEECH FEATURES

The auditory representation obtained through the model allows us to explore the contribution of different spectro-temporal elements of speech as viewed through the cortical array of modulation filters. The resulting feature tensor spans 3 dimensions of tonotopic frequency, spectral and temporal modulations. We investigated the effect of distortions on the 'marginals' of these dimensions in order to gain some insight into the components of speech that are more prone to degradation due to babble noise and bandlimiting effects.

Figure 2A shows the effect of low-pass filtering on the average frequency spectrum, temporal as well as spectral modulation profiles of 25 speech utterances. Overall, the gradual decrease in bandlimiting cutoff appears to attenuate all spectro-temporal modulations equally; having effectively a scaling effect on these distributions. The effect on the spectrum is obviously limiting the spectral bandwidth preserved in the signal. In contrast, adding babble noise with no low-pass filtering drastically affects the slow spectral modulations (below 1 cycle/octave); effectively reshaping the overall spectral envelope of the signal (Fig. 2B). No such effect was observed with additive white noise [1], providing further evidence for the more detrimental interference of babble over flat noise with the intelligibility of speech. Fig. 2B(right) reveals that babble distortions are more localized to the slow modulation range, while the fast modulations ($>$1 cycle/octave) appear merely scaled by a constant factor with increasing levels of noise. A similar scaling effect is also manifested over the entire range of temporal modulations. Note that the increase in energy (relative to the clean signal shown with a thick black line) is due to the addition of the noise to the clean signal. Finally, combining both additive babble noise (at -2 dB SNR) with bandlimiting (Fig.2C) maintains the scaling effect on the temporal modulations (middle panel), and the expected scaling and limitation in bandwidth along the tonotopic frequency dimension (leftmost panel). In contrast, the spectral modulation (i.e. scale) dimension appears to be severely reshaped at the low-end for low-pass cutoff frequencies $>$2 KHz.

5. CONTRIBUTION OF DIFFERENT SPEECH DIMENSIONS TO INTELLIGIBILITY

5.1. Behavioral intelligibility and the STMI

Figure 3A depicts the results of listeners' performance under all 6 conditions tested. The mean recognition rates (averaged across test sentences and subjects) reveal a smooth increase of scores over the range between 500 and 7000 Hz. The gray circles below the curve represent average scores per subject and reveal the variability in responses across listeners. This result complements previous findings about the contribution of different frequency bands to intelligibility, though it is worth noting that previous work has mostly focused on measurements in quiet [5]. As noted by Pollack back in 1948 [6], the interaction between frequency bands and speech intelligibility is different in quiet vs. noisy conditions. The curve in Fig 3A indicates a roughly linear increase in phoneme recognition with logarithmic frequency up to 2 KHz, qualitatively conforming with the findings of Pollack [6] under white noise conditions; though the effects do not closely match due to the different nature of the background [7].

Next, we use the STMI metric based on all cortical dimensions (frequency, rate and scale) to estimate a predicted score for each test condition. Figure 3B shows the strong agreement between the model's predictions and behavioral results, yielding an R-square value of 0.96. At the inset of Fig. 3B, we confirm the statistical significance of this correspondence

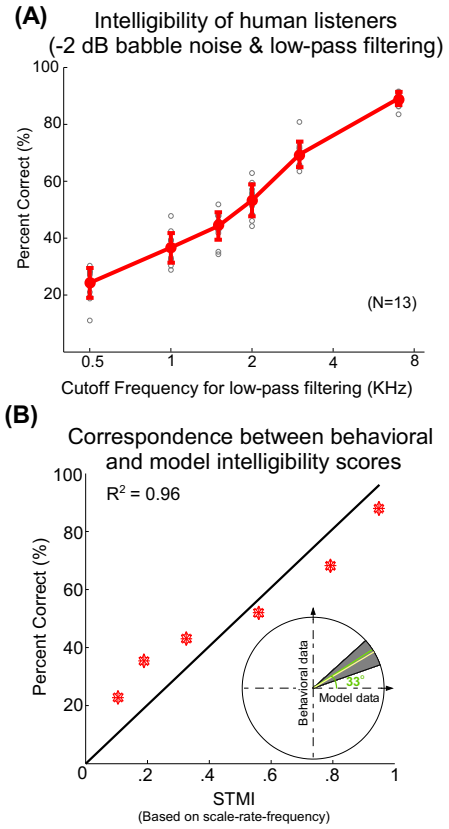

Fig. 3. Listeners performance for different low-pass cutoffs in a perceptual test with nonsense sentences, and the correspondence between the behavioral data and the model's STMI predictions for each condition.

between model and behavioral data through a bootstrap procedure. We quantified the slope (converted into an angle) that linearly maps the model and human scores on a per subject basis. The bootstrap analysis confirms that the two data sets are strongly correlated, with an original mean angle of 33°, a bootstrap mean of 30.8°, and the 5^{th} and 95^{th} confidence intervals, falling well within the upper right quadrant.

5.2. How informative are the different dimensions about intelligibility?

Finally, we extend this analysis by using only a subset of the dimensions present in the cortical feature tensor in order to perform similar model predictions. The goal of this exercise is to explore how closely does each of these dimensions match the listeners' scores. The results indicate that using scale, rate, frequency, rate-frequency, scale-frequency, rate-scale as well as all 3 combined (i.e. the STMI) yield an R-square matching to the behavior data of 0.32, 0.59, 0.61, 0.63, 0.82, 0.59 and 0.96 respectively. These numbers are very revealing as to the amount of variance that one can capture using each of these dimensions.

We particularly focus on the rate-frequency model (R-square: 0.63) compared to the STMI. Temporal modulations (i.e. rate) have been classically attributed a major role in the perception of speech [8], and constitute the backbone of many successful intelligibility metrics, such as the STI [9]. However, our analysis indicates that they only offer a partial view of the information in the speech signal, making them blind to the distortions of the overall spectral structure of the signal, hence limiting their applicability to adverse conditions.

6. CONCLUSIONS

Using a biologically-inspired model of auditory processing, we explored the contribution of different spectro-temporal elements of speech to its intelligibility under conditions of low-pass filtering and additive babble noise. Our results were validated against behavioral data from listeners tested using nonsense words. The analysis highlights the role of temporal and spectral modulations inherent to speech in its intelligibility. It also suggests that favoring certain dimensions over others (e.g., temporal over spectral modulations in the classic STI measurements) can lead to significant loss of information about the original speech, potentially yielding erroneous conclusions regarding the effects of a wide range of spectral distortions.

7. REFERENCES

[1] M. Elhilali, T. Chi, and S. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," *Speech Comm.*, vol. 41, pp. 331–348, 2003.

[2] J. Eggermont, "Between sound and perception: reviewing the search for a neural code," *Hear. Res.*, vol. 157, pp. 1–42, 2001.

[3] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *JASA*, vol. 118, no. 2, pp. 887–906, 2005.

[4] K. Grant, M. Elhilali, S. Shamma, B. Walden, R. Surr, M. Cord, and V. Summers, "An objective measure for selecting microphone modes in omni/dir hearing-aid circuits," *Ear and Hearing*, in press.

[5] R. Silipo, S. Greenberg, and T. Arai, "Temporal constraints on speech intelligibility as deduced from exceedingly sparse spectral representations," in *Eurospeech'99*, pp. 2687–2690.

[6] I. Pollack, "Effects of high-pass and low-pass filtering on the intelligibility of speech in noise," *JASA*, vol. 20, no. 3, pp. 259–266, 1948.

[7] J. Sperry, T. Wiley, and M. Chial, "Word recognition performance in various background competitors," *J Am Acad Audiol*, vol. 8, no. 2, pp. 71–80, 1997.

[8] R. Drullman, J. Festen, and R. Plomp, "Effect of envelope smearing on speech perception," *JASA*, vol. 95, no. 2, pp. 1053–1064, 1994.

[9] T. Houtgast and H. Steeneken, "A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria," *JASA*, vol. 77, no. 3, pp. 1069–1077, 1985.