The Interaction of Envelope and Temporal Fine Structure Cues in Speech Perception

Rasha A. Ibrahim and Ian C. Bruce

Department of Electrical & Computer Engineering
McMaster University, Hamilton, ON, Canada
ibruce@ieee.org

ABSTRACT

The role of temporal fine structure (TFS) in speech perception has been the subject of much recent investigation and debate. Lorenzi et al. (2006) utilized a vocoder scheme aimed at flattening the envelope in each frequency band, producing "TFS speech". They reported that normal-hearing listeners could learn over several sessions to understand TFS speech quite well, while perception by hearing-impaired listeners remained poor. However, speech envelope cues could be partly reconstructed by passing TFS speech through the narrow-band filters of the normal cochlea, bringing into question the interpretation of these results.

Smith et al. (2002) created an alternative form of processing that uses a pair of vocoders to produce "auditory chimaeras" having the TFS of one signal and the envelope of another. This has the potential to at least partially mask any envelope reconstructed by cochlear filtering of the TFS signal. However, because of the different experimental paradigms utilized, it is not possible to directly compare the results of the two studies.

In our study, we compared perception of TFS-only speech with that of speech-noise chimaeras in a group of 5 normal hearing subjects. NU-6 words were processed to generate five types of chimaeras: i) speech envelope + white Gaussian noise (WGN) TFS; ii) speech envelope + matched-noise (MN) TFS; iii) speech TFS + WGN envelope; iv) speech TFS + MN envelope; and v) TFS-only speech. For each chimaera type, we utilized 50 words for 7 different cases of the number of vocoder filters used, giving a total of 1750 words tested. An ANOVA on phoneme recognition scores shows large significant effects of chimaera type and number of filters ($p \ll 0.001$ in both cases). Predictions with a model of cochlear frequency filtering and cortical modulation filtering indicate that envelope reconstruction can partially but not fully explain the intelligibility of speech-TFS chimaeras.

I INTRODUCTION

- The last decade has seen substantial efforts to delineate the contributions of envelope (ENV) and temporal fine structure (TFS) cues to speech perception.
- Studies such as those of Shannon et al. (1995) and Smith et al. (2002) have indicated the dominance of envelope cues. In the latter investigation, "auditory chimaeras" were created by using vocoders to mix the envelope of one signal (within each frequency band) with the TFS of another signal (within the same frequency band), as depicted in Fig. 1. For a large number of narrow frequency bands, the envelope cues provided the greatest speech understanding.

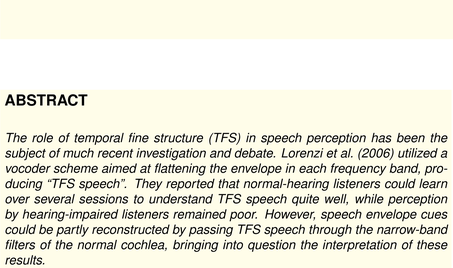

Figure 1: *Auditory chimaera processing of Smith et al. (2002).*

- However, recent reports have argued that normal-hearing subjects, in contrast to hearing-impaired listeners, can learn to utilize TFS speech cues if forced to (Lorenzi et al., 2006; Hopkins and Moore, 2007; Hopkins et al., 2008).
- A complicating factor is that when a TFS signal is passed through a narrow-band filter, some reconstruction of the envelope cues may occur (e.g., Zeng et al., 2004; Kale and Heinz, 2010), as illustrated in Fig. 2. The filtering of the normal cochlea may provide such narrowband filtering, and consequently envelope reconstruction in the discharge patterns of the auditory nerve.

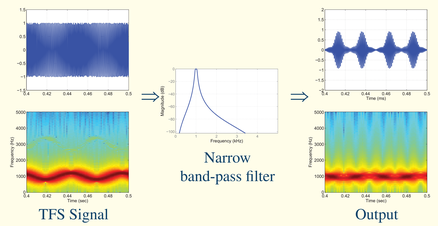

Figure 2: *Envelope reconstruction from a TFS signal passing through a narrow-band filter.*

- The "auditory chimaera" processing scheme of Smith et al. (2002) has the potential to reduce the effects of envelope reconstruction, because of the confounding envelope introduced by the second signal.
- However, for the case of a "noise" envelope mixed with a speech TFS, Smith et al. (2002) used a "noise" matched in spectrum to the individual speech stimulus, and Paliwal and Wojcicki (2008) have shown that such signals can contain useable speech cues.
- Consequently, the goals of this study were to:
  1. determine directly how the matched-noise signal in auditory chimaeras effects speech intelligibility, by comparing speech perception results for matched-noise chimaeras with results for two alternative chimaera schemes: a) a white Gaussian noise (WGN) signal instead of a matched noise for either the ENV or the TFS signal, and b) a flat envelope for speech-TFS signals, as used by Lorenzi et al. (2006); and
  2. estimate the amount of envelope reconstruction for all these chimaera types, by utilizing a computational model of the auditory periphery combined with a model of cortical spectro-temporal envelope analysis.

II SPEECH PERCEPTION EXPERIMENT

A Methods

- 5 chimaeras types were tested:
  - i. Speech-ENV+WGN-TFS
  - ii. Speech-ENV+Matched-Noise-TFS
  - iii. Speech-TFS+WGN-ENV
  - iv. Speech-TFS+Matched-Noise-ENV
  - v. TFS-only
- The number of frequency bands in the vocoders was 1, 2, 3, 6, 8, 16, or 32.
- 50 NU-6 words were randomly chosen (from the set of 200 NU-6 words; Tillman and Carhart, 1966) for each vocoder with a given number of filters, giving 350 words per chimaera type and a total of 1750 stimulus presentations.
- Results were obtained from 5 normal hearing subjects aged 18–21 who were all native speakers of North American English.
- Each subject completed 5 sessions, with a different chimaera type being tested in each session. The order of the sessions was randomized for each subject. The word presentation within each session was also randomized.
- All signals were generated with a high-quality "Turtle Beach - Audio Advantage Micro" PC sound card at a sampling rate of 44100 Hz. The signals are calibrated through a "B & K 2260 Investigator" sound meter/audiometer (artificial ear type 4152) to present the speech at ~ 65 dB SPL. The sound is presented binaurally to the subjects via a Yamaha HTR-6150 amplifier and Sennheiser HDA 200 headphones while seated in a quiet room.
- Responses by the subjects were typed by the experiment operator, in addition to the voice response being digitally recorded. The responses were later scored to give a percentage correct for phonemes, vowel, consonants and the entire word. Only phoneme perception data are reported here. The experiment operators and the scorer were all native English speakers.
- No training or feedback was provided.

B Results

- The results of a 3-way ANOVA on the phoneme scores of the main effects of *Chimaera Type*, *No. of Filters* and *Subject No.* plus two-factor interactions are shown in Table I.
- All three factors are statistically significance, but the chimaera type and number of filters are much stronger factors than the subject number.
- The interactions between the number of filters & chimaera type and between subject number & chimaera type are significant, but the interaction between the subject number & number of filters is not.

Table I: *3-way ANOVA on Phoneme Perception Data*

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
Chimaera Type	60.83	4	15.2074	262.18	0
No. of Filters	72.44	6	12.073	208.14	0
Subject No.	0.83	4	0.2073	3.57	0.0064
No. of Filters×Chimaera Type	380.46	24	15.8527	273.31	0
Subject×Chimaera Type	3.85	16	0.2405	4.15	0
Subject×No. of Filters	1.8	24	0.0748	1.29	0.1553
Error	502.95	8671	0.058		
Total	1023.15	8749			

- Phoneme scores are plotted in Fig. 3. An effect of using a "matched" noise signal is clearly seen—it tends to improve perception for the speech-ENV chimaera and degrade intelligibility for the speech-TFS chimaera.
- A continuing decrease in intelligibility for all 3 speech-TFS chimaera schemes (right panel) is seen to occur with increasing vocoder filter numbers, i.e., with decreasing vocoder filter bandwidth.

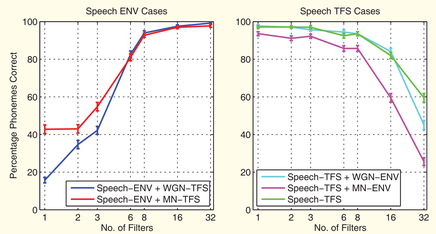

Figure 3: *Phoneme perception scores from listening experiment. Error bars show ± 1 SEM.*

III MODEL PREDICTIONS

A Model of cochlear filtering

- The auditory periphery model of Zilany and Bruce (2006, 2007b), shown in Fig. 4, was utilized to evaluate the effects of cochlear filtering on envelope reconstruction.

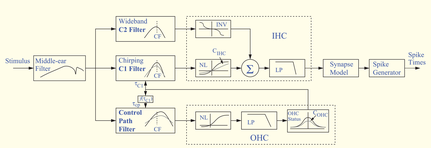

Figure 4: *Zilany and Bruce (2006, 2007b) auditory periphery model.*

- The cochlear tuning of the model was modified to match estimates from humans (Shera et al., 2002), as described in Ibrahim and Bruce (2010).
- Simultaneous outputs (discharge rates averaged over 8 ms with 50% overlap) from 128 AN fibers, CFs ranging from 0.18 to 7.04 kHz spaced logarithmically, make up the AN "neurogram", as shown in Fig. 5.

B Speech Intelligibility Predictor

- A cortical model of speech processing (Elhilali et al., 2003) analyzes the AN neurogram to estimate the spectral and temporal modulation content, as shown in Fig. 5. It is implemented by a bank of modulation-selective filters ranging from slow to fast rates (2 to 32 Hz) temporally and narrow to broad (0.25 to 8 cyc/oct) scales spectrally.

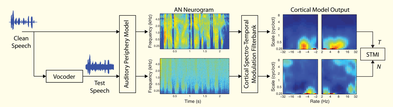

Figure 5: *Schematic of the Spectro-Temporal Modulation Index (STMI) speech-intelligibility predictor computation. The clean and chimaera speech signals are given as inputs to the auditory periphery model, and the spectral and temporal modulations in the auditory nerve responses are then analyzed by the cortical models filters to compute the STMI.*

- After analyzing the two-dimensional (time and frequency) AN neurogram with the modulation filter banks, the cortical output is a four-dimensional (time, frequency, rate and scale) complex-valued representation.
- Once the cortical output of the test stimulus, N, and the template, T, for that stimulus are computed, the STMI can be calculated as (Elhilali et al., 2003):

$$\mathrm{STMI} = 1 - \frac{\| T - N \|^2}{\| T \|^2}, \qquad (1)$$

where $\| \cdot \|$ indicates the 2-norm of the corresponding signal.

- Following Zilany and Bruce (2007a), the template has been chosen as the output of the normal model to the unprocessed stimulus at 65 dB SPL (conversational speech level) in quiet.
- The STMI takes values between 0 and 1, with higher values predicting better speech intelligibility. In practice, the STMI has a lower limit of $\simeq 0.13$ for the speech material tested in this study.
- Due to large time bins in the AN neurogram and the slow temporal modulation rates for cortical filters, all TFS cues are filtered out in our STMI results, and consequently *the STMI predictions are based on direct and reconstructed envelope cues only*.

C Results

- STMI predictions are shown in Fig. 6. The general trend of improving intelligibility with a larger number of filters for the speech-ENV chimaeras and the reverse for the speech-TFS chimaeras is observed.

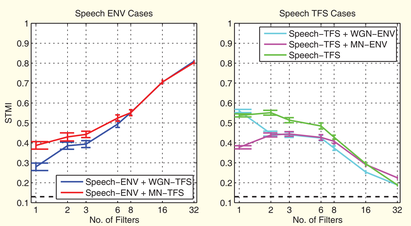

Figure 6: *Model predictions of phoneme perception. Error bars show ± 1 SEM.*

- A mapping of STMI predictions to speech experiment results, plotted in Fig. 7, shows that envelope reconstruction can partially but not fully explain the phoneme perception for the speech-TFS chimaeras.

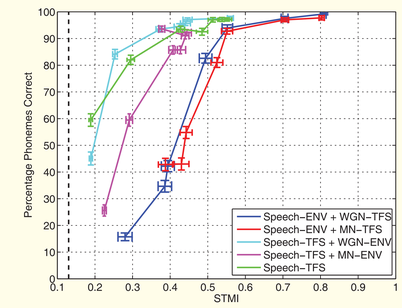

Figure 7: *Mapping of STMI to experiment results. Error bars show ± 1 SEM.*

IV CONCLUSIONS AND FUTURE WORKS

- Envelope and TFS cues do interact in speech perception, as indicated by the effects of using "matched noise" signals in auditory chimaeras.
- Envelope reconstruction can partially but not completely explain the intelligibility of speech-TFS chimaeras.
- These results motivate the inclusion of TFS cues in the neural predictor of speech intelligibility.

V ACKNOWLEDGMENTS

We thank Laurel Carney & Hubert de Bruin for advice on the experiment design, Malcolm Pilgrim & Timothy Zeyl for assistance with running the experiment, Sue Becker for use of her amplifier and headphones, Dan Bosnyak & Dave Thompson for assistance with the acoustic calibration, and the subjects for their participation.

This research was supported by NSERC (Discovery Grant #261736), and the human experiments were approved by the McMaster Research Ethics Board (#2010 051).

REFERENCES

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

Hopkins, K. and Moore, B. C. J. (2007). "Moderate cochlear hearing loss leads to a reduced ability to use temporal fine structure information," J Acoust Soc Am **122**, 1055–1068.

Hopkins, K., Moore, B. C. J., and Stone, M. A. (2008). "Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech," J Acoust Soc Am **123**, 1140–1153.

Ibrahim, R. A. and Bruce, I. C. (2010). "Effects of peripheral tuning on the auditory nerve's representation of speech envelope and temporal fine structure cues," in *The Neurophysiological Bases of Auditory Perception*, edited by E. A. Lopez-Poveda, A. R. Palmer, and R. Meddis (Springer, New York), Chap. 40, pp. 429–438.

Kale, S. and Heinz, M. G. (2010). "Envelope coding in auditory nerve fibers following noise-induced hearing loss," J. Assoc. Res. Otolaryngol. **11**, 657–673.

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., and Moore, B. C. J. (2006). "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," Proc. Natl. Acad. Sci. U.S.A. **103**, 18866–18869.

Paliwal, K. and Wojcicki, K. (2008). "Effect of analysis window duration on speech intelligibility," IEEE Signal Processing Letters **15**, 785–788.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). "Speech recognition with primarily temporal cues," Science **270**, 303–304.

Shera, C. A., Guinan, J. J., and Oxenham, A. J. (2002). "Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements," Proc. Natl. Acad. Sci. U.S.A. **99**, 3318–3323.

Smith, Z. M., Delgutte, B., and Oxenham, A. J. (2002). "Chimaeric sounds reveal dichotomies in auditory perception," Nature **416**, 87–90.

Tillman, T. W. and Carhart, R. (1966), "An expanded test for speech discrimination utilizing CNC monosyllabic words. Northwestern University Auditory Test No. 6," Tech. Report. SAM-TR-66-55, (USAF School of Aerospace Medicine, Brooks Air Force Base, Texas).

Zeng, F.-G., Nie, K., Liu, S., Stickney, G., Rio, E. D., Kong, Y.-Y., and Chen, H. (2004). "On the dichotomy in auditory perception between temporal envelope and fine structure cues," J. Acoust. Soc. Am. **116**, 1351–1354.

Zilany, M. S. A. and Bruce, I. C. (2006). "Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery," J. Acoust. Soc. Am. **120**, 1446–1466.

Zilany, M. S. A. and Bruce, I. C. (2007a). "Predictions of speech intelligibility with a model of the normal and impaired auditory-periphery," in *Proceedings of 3rd International IEEE EMBS Conference on Neural Engineering* (IEEE, Piscataway, NJ).

Zilany, M. S. A. and Bruce, I. C. (2007b). "Representation of the vowel /ɛ/ in normal and impaired auditory nerve fibers: Model predictions of responses in cats," J. Acoust. Soc. Am. **122**, 402–417.