Auditory features in consonant perception - a modeling perspective

Johannes Zaar[1] and *Torsten Dau*[2]

[1,2] Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark

jzaar@elektro.dtu.dk, tdau@elektro.dtu.dk

Index Terms: consonant perception, microscopic speech perception modeling, spectral features, modulation features

1. Introduction

Speech perception is often studied from a *macroscopic* perspective by presenting meaningful speech stimuli in a range of acoustic conditions to a panel of listeners and evaluating the speech intelligibility in terms of the percentage of correct responses (e.g. [1]). Speech intelligibility measured this way reflects a complex interaction between the acoustic cues available to the listeners and the syntactic structure, the semantic predictability and the lexical content of the speech stimuli. This makes it difficult to tease apart these influencing factors when trying to investigate the preservation/restoration of the speech signal's acoustic cues (e.g. in acoustic transmission channels and/or through an impaired ear with or without hearing aid). Taking a *microscopic* perspective, this problem can be overcome using nonsense syllables (e.g., consonant-vowel combinations, CVs), thus excluding syntactic, semantic, and lexical effects. Many related studies have focused on the perception of consonants in steady-state noise at various signal-to-noise ratios (SNRs) and evaluated the responses in terms of consonant recognition and consonant confusions [2,3,4,5]. Several studies [5,6,7] demonstrated a large perceptual variability across different speech tokens of the same type (e.g., different utterances of /ba/). The results of a recent study [8] by the authors of the current study confirmed the large speech-token induced variability of consonant-in-noise perception and additionally showed that even different realizations of white Gaussian masking noise mixed with identical speech tokens had a perceptual effect.

Several vastly different modeling approaches have been proposed for predicting consonant perception. Li et *al.* [9,10] related the SNR in experimentally determined time-frequency regions to consonant recognition rates. Gallun and Souza [11] demonstrated that the correlation of long-term modulation power representations of their noise-vocoded stimuli was related to the observed consonant confusions. Jürgens and Brand [12] used a modulation-frequency selective auditory model in combination with a template-matching speech recognizer that showed convincing recognition predictions while the confusion predictions were inconclusive.

The data from [8] allow for a detailed evaluation of model predictions obtained with the corresponding experimental stimuli. The current study investigated the predictive power of various auditory-inspired internal representations (IRs) in combination with a template-matching back end. The perceptual variability across speech tokens was taken into account using a token-by-token analysis of the correlation between model predictions and perceptual data.

2. Stimuli and data set

The speech tokens from [8] consist of the 15 CVs (/bi, di, fi, gi, hi, ji, ki, li, mi ni, pi, si, ʃi, ti, vi/), each spoken three times by a male and a female talker (i.e., six speech tokens per CV). The talkers were native speakers of Danish. Eight young normal-hearing native Danish listeners were presented with the 90 different speech tokens mixed with white noise at SNRs of 12, 6, 0, -6, -12, and -15 dB (3 presentations per speech token and SNR). The listeners had to select the consonant they heard. The occurrences of responses obtained with each speech token at each SNR were pooled across listeners and converted to response probabilities. As a result, response rate patterns (consonant recognition and confusion rates) were obtained for each speech token and SNR.

3. Model components

3.1. Internal representations (IRs)

Six different model IRs were considered as front ends:

st-AS: The *auditory spectrogram* was defined as the Hilbert envelopes at the outputs of 25 fourth-order gammatone filters with center frequencies between 63 Hz and 16 kHz. The envelopes were low-pass filtered using a second-order Butterworth filter with a cut-off frequency of 8 Hz. The power of the subband envelopes was calculated in 4-ms time frames.

st-MS: The *modulation spectrogram* was obtained using the same gammatone filterbank as described above, followed by Hilbert envelope extraction and a first-order low-pass filter with a cut-off frequency of 150 Hz. Each subband envelope was then passed through a modulation filterbank consisting of a third-order low-pass filter with a cut-off frequency of 2 Hz and five second-order band-pass filters with center frequencies of 4, 8, 16, 32, and 64 Hz. The power at the outputs of the modulation filters was calculated in 4-ms time frames.

st-MS_{ac}: The *ac-coupled modulation spectrogram* was calculated from the modulation spectrogram by normalizing the power at the output of each modulation filter to the DC power of the respective subband envelope.

lt-EP: The long-term *excitation pattern* was calculated as the long-term power at the output of the gammatone filterbank described above.

lt-MS: The long-term *modulation spectrum* was obtained by integrating the modulation spectrogram over time.

lt-MS_{ac}: The long-term *ac-coupled modulation spectrum* was obtained by integrating the ac-coupled modulation spectrogram over time.

All IRs were converted to dB.

3.2. Template matching

A template-matching back end was used to predict the perceptual data. The templates were generated from the 90 speech tokens used in the experiment (15 CVs X 2 talkers X 3 recordings). Two sets of templates were considered:

1) Clean speech tokens, each mixed 10 times with random white noise at 12 dB SNR ("Noise").
2) Clean speech tokens, each 10 times time-scale modified (TSM) using a standard TSM algorithm [13] with the stretching factor α randomly drawn from a uniform distribution on the open interval [0.5, 1.5], and then mixed with random white noise at 12 dB SNR ("TSM & Noise").

The experimental stimuli were considered as test signals and compared to all *talker-specific* templates. The correct response alternative was represented only by the 10 templates generated from the test speech token, assuming a priori knowledge about the speech content in the test signal. The other response alternatives were represented by all available talker-specific templates.

The test signal and the templates were passed through the front end under investigation and the Euclidean distances between the test signal IR and the template IRs were obtained. In the case of the time-dependent IRs (st-AS, st-MS, and st-MS_{ac}), a standard dynamic time warping (DTW) algorithm [14] was applied to compensate for temporal misalignment. The Euclidean distances were converted to response rates by counting the number of times a template representing a given CV showed the minimum distance to the test signal, considering all possible combinations of templates. The number of times minima occurred for each phonetic category of templates was divided by the number of considered template combinations, yielding the predicted response rates.

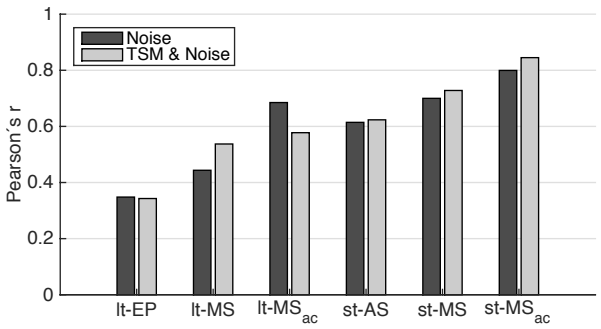

Figure 1: Average correlation between data and predictions obtained with the considered IRs (left to right) and template sets (black and gray bars).

4. Results and discussion

The predictions obtained with the different model configurations were evaluated by comparing them to the perceptual data from [8]. Pearson's correlation coefficients between model predictions and data were calculated for each speech-token specific combination of across-SNR average response rate patterns. Fig. 1 shows the average correlation between predictions and data as a function of the considered model IRs (along the abscissa) for the two template sets (black bars: Noise; gray bars: TSM & Noise). The correlation results indicate that the short-term modulation-based IRs yielded the by far largest predictive power, in particular the ac-coupled modulation spectrogram (st-MS_{ac}). Time-scale modification of the templates mostly yielded a correlation benefit (gray versus black bars). The highest correlation of $r = 0.85$ was obtained with st-MS_{ac} using the TSM & Noise template set (rightmost gray bar in Fig. 1). The results suggest that (1) modulation-frequency selective internal representations are well suited for predicting consonant perception in noise, (2) an ac-coupled modulation-domain representation that only describes deviations from the long-term spectrum is particularly suitable, and (3) time-scale modification of a limited set of template tokens may help improve the model predictions.

5. Acknowledgements

This research was funded with support from the European Commission under Contract No. FP7-PEOPLE- 2011-290000.

6. References

[1] M. Nilsson, S. D. Soli, and J. A. Sullivan, "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," *The Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1085–1099, 1994.

[2] G. A. Miller and P. E. Nicely, "An analysis of perceptual confusions among some English consonants," *The Journal of the Acoustical Society of America*, vol. 27, no. 2, pp. 338–352, 1955.

[3] M. D. Wang and R. C. Bilger, "Consonant confusions in noise: A study of perceptual features," *The Journal of the Acoustical Society of America*, vol. 54, no. 5, pp. 1248–1266, 1973.

[4] S. A. Phatak and J. B. Allen, "Consonant and vowel confusions in speech-weighted noise," *The Journal of the Acoustical Society of America*, vol. 121, no. 4, pp. 2312–2326, 2007.

[5] S. A. Phatak, A. Lovitt, and J. B. Allen, "Consonant confusions in white noise," *The Journal of the Acoustical Society of America*, vol. 124, no. 2, pp. 1220–1233, 2008.

[6] R. Singh and J. B. Allen, "The influence of stop consonants' perceptual features on the articulation index model," *The Journal of the Acoustical Society of America*, vol. 131, no. 4, pp. 3051–3068, 2012.

[7] J. C. Toscano, J. B. Allen, "Across- and within-consonant errors for isolated syllables in noise," *Journal of Speech Language and Hearing Research*, vol. 57, pp. 2293–2307, 2014.

[8] J. Zaar and T. Dau, "Sources of variability in consonant perception of normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol. 138, no. 3, pp. 1253–1267, 2015.

[9] F. Li, A. Menon, and J. B. Allen, "A psychoacoustic method to find the perceptual cues of stop consonants in natural speech," *The Journal of the Acoustical Society of America*, vol. 127, no. 4, pp. 2599–2610, 2010.

[10] F. Li, A. Trevino, A. Menon, and J. B. Allen, "A psychoacoustic method for studying the necessary and sufficient perceptual cues of American English fricative consonants in noise," *The Journal of the Acoustical Society of America*, vol. 132, no. 4, pp. 2663–2675, 2012.

[11] F. Gallun and P. Souza, "Exploring the role of the modulation spectrum in phoneme recognition," *Ear and Hearing*, vol. 29, no. 5, pp. 800–813, 2008.

[12] T. Jürgens and T. Brand, "Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model," *The Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2635–2648, 2009.

[13] J. Driedger and M. Müller, "TSM toolbox: Matlab implementations of time-scale modification algorithms," *Proc. Int. Conference on Digital Audio Effects (DaFx-14),* 2014.

[14] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 26, no. 1, pp. 43–49, 1978.