Proceedings of Meetings on Acoustics

Volume 19, 2013

<http://acousticalsociety.org/>

ICA 2013 Montreal

Montreal, Canada

2 - 7 June 2013

Speech Communication

Session 2pSCb: Speech Intelligibility (Poster Session)

2pSCb20. The role of across-frequency envelope processing for speech intelligibility

Alexandre Chabot-Leclerc*, Soren Jorgensen and Torsten Dau

***Corresponding author's address: Centre for Applied Hearing Research, Technical University of Denmark, Ørsted's plads, Bldg 352, Kongens Lyngby, 2800, DK-2800, Denmark, alech@elektro.dtu.dk**

Speech intelligibility models consist of a preprocessing part that transforms the stimuli into some internal (auditory) representation, and a decision metric that quantifies effects of transmission channel, speech interferers, and auditory processing on the speech intelligibility. Here, two recent speech intelligibility models, the spectro-temporal modulation index (STMI; Elhilali et al., 2003) and the speech-based envelope power spectrum model (sEPSM; Jørgensen and Dau, 2011) were evaluated in conditions of noisy speech subjected to reverberation, and to nonlinear distortions through either a phase jitter process or noise reduction via spectral subtraction. The contributions of the individual preprocessing stages in the models and the role of the decision metrics were analyzed in the different experimental conditions. It is demonstrated that an explicit across-frequency envelope processing stage, as assumed in the STMI, together with the metric based on the envelope power signal-to-noise ratio, as assumed in the sEPSM, are required to account for all three conditions. However, a simple across audio-frequency mechanism combined with a purely temporal modulation filterbank is assumed to be sufficient to describe the data, i.e., a joint two-dimensional modulation filterbank might not be required.

Published by the Acoustical Society of America through the American Institute of Physics

INTRODUCTION

The speech transmission index (STI; Houtgast *et al.*, 1980; Steeneken and Houtgast, 1980; IEC, 2003) was the first speech intelligibility model to consider the integrity of the temporal envelope fluctuations of a reference signal, quantified by the modulation transfer function (MTF) as the decision metric. The MTF measures the reduction of the envelope fluctuations of the reference signal, as a function of audio and modulation frequency, and captures effects of distortions on the envelope caused by reverberation and steady noises (Houtgast *et al.*, 1980; Steeneken and Houtgast, 1980; Houtgast and Steeneken, 1985). The STI, however, fails in conditions with nonlinear processing, such as envelope compression (Rhebergen and Versfeld, 2005), phase jitter and phase shifts (Elhilali *et al.*, 2003), or spectral subtraction (Ludvigsen *et al.*, 1993; Dubbelboer and Houtgast, 2007). To overcome this limitation, Payton and Braid (1999), and Goldsworthy and Greenberg (2004) introduced modifications of the STI, generally known as speech-based STI methods (sSTI). The main difference between the STI and sSTI methods was that speech was used as the reference signal, rather than amplitude modulated noise. Although the sSTI methods seemed promising, they were never evaluated explicitly using comparisons between measured and predicted speech intelligibility.

An alternative approach was considered by Elhilali *et al.* (2003) who proposed the spectro-temporal modulation index (STMI), which measured the integrity of the spectral and temporal modulation energy of a signal, in contrast to the STI, which considered the temporal modulations only. Conceptually, the STMI considers a two-dimensional (spectro-temporal) MTF at the output of a spectro-temporal modulation bandpass filterbank as the decision metric. Elhilali *et al.* (2003) defined two versions of the STMI: one version used a spectro-temporally modulated noise as the reference signal, denoted as a ripple and analogous to the temporally modulated noise of the STI, and the second version used clean speech as the reference signal, similar to the sSTI methods. The ripple-based and speech-based STMI were shown to be consistent with the STI in conditions with additive noise and reverberation. Furthermore, both STMI versions could account for the effects of phase jitter and phase shifts, two distortions to which the STI was insensitive. The key component in the STMI to account for the phase distortions was assumed to be the analysis of the modulations across the frequency axis, which was not included in the STI. However, the STMI is based on the MTF concept and should therefore have the same limitations as the STI when noisy speech is processed by spectral subtraction.

Recently, Jørgensen and Dau (2011) proposed the envelope power signal-to-noise ratio (SNR_{env}) as the decision metric, implemented as a part of the speech-based envelope power spectrum model (sEPSM). The SNR_{env} is estimated at the output of modulation-frequency selective filtering and was shown to account for the changes of intelligibility observed in conditions with additive noise, reverberation, and spectral subtraction. The key component allowing the SNR_{env} to account for spectral subtraction was the consideration of intrinsic modulations of the noise alone which are affected by this noise reduction processing and which are neglected in the MTF concept. However, sEPSM is likely to fail in conditions with distortions that affect the spectral structure (e.g. formant structure) of the speech while maintaining the temporal structure, such as in the case of a phase jitter distortion, since the sEPSM does not contain any across-frequency operation (besides a simple integration of information). Conceptually, the STMI and the sEPSM introduced different modifications to the STI. The STMI introduced the spectral modulation-frequency selectivity, which was essential in order to account for phase jitter effects, and kept the MTF-based decision metric. The sEPSM introduced the SNR_{env} decision metric, which was crucial to account for spectral subtraction, while keeping the purely temporal modulation-frequency selective process.

In the present study, two approaches are proposed; one integrates the two-dimensional filterbank of the STMI in the sEPSM (replacing the original 1-D filterbank), denoted here as

2D-sEPSM, and a second implements a simple across-frequency process by computing the across audio-frequency variance of activity at the output of the purely temporal (1-D) modulation filterbank of the sEPSM, denoted as sEPSM^X. The two approaches were evaluated in conditions of reverberation, phase jitter, and spectral subtraction. In particular, the focus was to analyze which components of the models were essential for successful predictions of speech intelligibility in the different conditions, in an attempt to better understand the essential auditory features underlying robust speech perception in noisy environments.

MODEL DESCRIPTION

Auditory Preprocessing

The processing structure of the modeling is illustrated in Figure 1. The first stage performs the peripheral filtering and transforms the acoustic signal into a two-dimensional representation of the neural activity as a function of time, t , and cochlear frequency, f . In the 2D-sEPSM, the acoustic signal is filtered using a bandpass filterbank consisting of 128 fourth-order gammatone filters equally spaced on a logarithmic scale between 90 Hz and 3.5 kHz (24 filters/octave over a 5.3 octave range). In the sEPSM^X, there are 22 filters with $1/3$ -octave spacing of the center frequencies, which cover the range from 63 Hz to 8 kHz. The higher number of filters in the 2D-sEPSM is required to have a sufficient spectral resolution for the spectro-temporal modulation processing. The envelope of the output of each gammatone filter is extracted using the Hilbert transform and short-time averaged in blocks of 8 ms. An absolute sensitivity threshold is included in both models, such that only peripheral filters with output energy above normal hearing threshold are considered.

The modulation filtering stage is specific to each model. The 2D-sEPSM implements a joint spectral and temporal modulation filtering using a bank of spectrally and temporally selective modulation filters (Chi *et al.*, 1999). The spectro-temporal modulation filterbank consists of third-octave wide octave-spaced filters; the center frequencies of the temporal modulation filters are ranged between 2 and 32 Hz, and the center frequencies of the spectral modulation filters are ranged between 0.25 and 8 cyc/oct. The output of the 2D modulation filtering stage is a four-dimensional representation, indexed by time, cochlear frequency, rate, ω , and scale, Ω (see Chi *et al.* (1999) and Chi *et al.* (1999)). The sEPSM^X considers a one-dimensional temporal modulation filterbank, consisting of a third-order lowpass filter in parallel with seven overlapping second-order bandpass filters. The cutoff frequency of the lowpass filter is 1 Hz and the bandpass filters have center frequencies from 2 to 128 Hz with octave spacing and a constant Q-factor of 1.

Decision Device

For the 2D-sEPSM, the cortical representation is first integrated over the sentence duration, leaving a three dimensional internal representation of the noisy sentence, or mixture, $\{T_{\text{mix}}(f, \omega, \Omega)\}$, and of the noise alone, $\{N(f, \omega, \Omega)\}$. The time-averaged $\{T_{\text{mix}}\}$ and $\{N\}$ internal representations are considered equivalent to the envelope power spectrum of the noisy speech, $P_{\text{env}, S+N}$, and of the noise alone, $P_{\text{env}, N}$, as defined in Jørgensen and Dau (2011). The SNR_{env} can then be expressed as the ratio of the difference of the mixture and the noise-alone (internal) representation and the noise-alone internal representation:

$$\text{SNR}_{\text{env}} = \frac{T_{\text{mix}} - N}{N}. \quad (1)$$

The SNR_{env} values are then integrated across peripheral and modulation filters using the “integration model” (Green and Swets, 1988). In the sEPSM^X, the long-term AC-coupled

FIGURE 1: Block diagram showing the overall structure of the modeling framework. The model consists of a gammatone bandpass filterbank followed by envelope extraction via Hilbert transformation, producing the auditory spectrogram. The modulation processing comprises either a 2D modulation filterbank (2D-sEPSM) or an across-frequency variance weighting after a 1D (temporal) modulation filterbank (sEPSM^X). The envelope signal-to-noise ratio, SNR_{env} , is calculated from the time-averaged cortical representation at the output of the modulation processing and combined across modulation filters and audio filters. The overall SNR_{env} is converted to a percentage of correctly recognized speech items using an ideal observer process.

envelope power is calculated from the temporal output of each modulation filter and normalized with the DC power of the unfiltered envelope. The SNR_{env} is then calculated from the envelope power of the noisy speech, $P_{\text{env},S+N}$, and the noise alone, $P_{\text{env},N}$, at the output of each modulation filter. The SNR_{env} values from all modulation filters are then integrated, assuming that the contribution from a given modulation filter, m , depends on the variance of the modulation-filter output across the audio-frequency filters:

$$\text{SNR}_{\text{env},m} = \sigma_{f_m} \cdot \text{SNR}_{\text{env},f_m}, \quad (2)$$

where σ_{f_m} is the across audio-filter variance, computed as the variance of the root-mean-square level of the outputs of all the modulation filters centered at f_m . σ_{f_m} is estimated in non-overlapping time-frames of 250 ms and averaged over the duration of a given speech token. Finally, the SNR_{env} contributions from all modulation and audio filters are combined using the integration model.

In both models, the combined SNR_{env} value is converted to the probability of correctly recognizing a given speech token using the concept of an ideal observer (Jørgensen and Dau, 2011).

METHOD

Model predictions were compared to data obtained using the Conversation Language Understanding Evaluation (CLUE) test, consisting of unique meaningful five-word sentences (Nielsen and Dau, 2009). The sentences were mixed with speech-shaped stationary noise having the same long-term average spectrum as the speech material. Six normal-hearing males participated in the experiment.

Reverberation

The noisy sentences were convolved with an impulse response corresponding to a particular reverberation time. The impulse responses were created using the ODEON room acoustic software version 10 (Christensen, 2009). The room simulated was identical to the one used by Jørgensen and Dau (2011): it was designed to have equal reverberation time (T_{30}) in the frequency range 63–8000 Hz. Five different values of T_{30} were used: 0, 0.4, 0.7, 1.3, and 2.3 s.

Spectral subtraction

The spectral subtraction was implemented using the scheme defined by Berouti *et al.* (1979). The processing was applied by subtracting an estimate of the noise power spectrum from the power spectrum of the noisy speech in short time windows. The amount of noise subtracted was defined by the over-subtraction factor, κ . The implementation was identical to the one in Jørgensen and Dau (2011). Six different over-subtraction factors were considered: 0, 0.5, 1, 2, 4 and 8, where $\kappa = 0$ corresponded to the reference condition with no spectral subtraction, but with signal reconstruction.

Phase Jitter

Phase-jitter distorted noisy speech was obtained by multiplying the noisy speech signal, $s(t)$, with a cosine with a random phase:

$$r(t) = \Re \left\{ s(t) e^{j\Theta(t)} \right\} = s(t) \cos(\Theta(t)), \quad (3)$$

where $\theta(t)$ is a random process uniformly distributed over $[0, 2\alpha\pi]$ ($0 < \alpha < 1$), and α is the parameter defining the amount of jitter (Elhilali *et al.*, 2003). The α -values used covered the range 0 to 1 in steps of 0.125. For $\alpha = \{0.5, 1\}$ the signal becomes a temporally modulated white noise.

Model Setup and Parameters

For the predictions, the model parameters were calibrated to the closest match between the predictions and the data in the conditions with SSN only. These parameters were then used for other experimental conditions. The speech material was the same as for obtaining the data. The sentences were down-sampled to 8192 Hz for the 2D-sEPSM and to 22050 Hz for the sEPSM^X to reduce computation time.

RESULTS

Figure 2 shows prediction and data for speech in stationary noise processed by the three distortions. The open squares represent the data and the filled symbols represent the model predictions. Figure 2a shows the Δ SRT as a function of the reverberation time. The mean SRT for the case without reverberation ($T_{30} = 0$) was -3.5 dB SNR. The vertical bars denote one standard deviation of the listener's average SRT. The measured Δ SRT increased with increasing reverberation time, reflecting a decrease in intelligibility, and all models accounted for the main characteristics in the data. The Pearson correlation coefficient between the 2D-sEPSM predictions (filled squares) and the data was 0.985 and the root-mean-square-error (RMSE) was 2.78 dB. The sEPSM^X (upward triangles) had a Pearson coefficient of 0.981 and an RMSE of 3.56 dB.

FIGURE 2: Measured (open symbols) and predicted (filled symbols) ΔSRTs obtained in the reverberation condition (panel A) and the spectral subtraction condition (panel B), and percentage of correctly understood words obtained in the phase-jitter condition (panel C). The vertical bars and the shaded area show one standard deviation. The reverberation and spectral subtraction data and the sEPSM predictions are replotted from Jørgensen and Dau (2011). Root-mean-square error (RMSE) and Pearson correlation coefficient between the data and both the 2D-sEPSM (filled squares) and sEPSM^x (filled upward triangles) are shown in the top two figures.

Figure 2b shows the results for the conditions with spectral subtraction as a function of the over-subtraction factor κ . The measured ΔSRT increased as the over-subtraction factor increases. The 2D-sEPSM predictions also increased with κ , however, the predicted ΔSRTs were higher than the data for all κ , except when $\kappa = 0.5$, where it was smaller. The Pearson correlation coefficient between the data and the 2D-sEPSM was 0.926 and the RMSE was 1.23 dB. The sEPSM^x predictions had a higher correlation to the data ($\rho = 0.981$) but a larger RMSE of 2.01 dB. The STMI (filled circles) clearly failed to account for spectral subtraction, predicting a decrease in ΔSRT in contrast to the data.

Figure 2c shows the results for the phase jitter experiment. The percentage of correctly understood words is shown as a function of the phase jitter parameter α . The shaded area represents one standard deviation. The intelligibility score showed a characteristic trend as a function of α , with 100 % intelligibility for α close to 0, a decreasing intelligibility down to 0 % for $\alpha = 0.5$, followed by a local maximum of about 65 % for $\alpha = 0.75$, and finally 0 % intelligibility for $\alpha = 1$. The 2D-sEPSM accounted for this trend, with 100 % intelligibility below $\alpha = 0.25$, minima at $\alpha = \{0.5, 1\}$, and a local maximum at $\alpha = 0.75$. However, the simulated intelligibility scores never reached values below 20.9 %. The predictions from the sEPSM^x covered the whole range of intelligibility scores between 0 and 100 %, in contrast to the 2D-sEPSM, but deviated clearly from the data for $\alpha = 0.75$. The sEPSM (downward triangles) was insensitive to the

effects of the phase jitter due to the absence of across-frequency processing. The predictions obtained with the STMI also showed the correct trend, but the values were bound between 12 and 89 % intelligibility. In essence, all models except for the (original) sEPSM were sensitive to the effects of the phase jitter due to their respective across-frequency process.

DISCUSSION

All considered models could, to some extent, account for the effects of reverberation on the intelligibility of noisy speech. In contrast, only the models considering the decision metric based on the SNR_{env} could account for the detrimental effect of spectral subtraction. The MTF-like metric of the STMI could not account for the spectral subtraction data because it does not consider the effects of the nonlinear processing on the noise modulations alone. This is consistent with the results from Jørgensen and Dau (2011) and supports the SNR_{env} as an appropriate metric for speech intelligibility prediction.

The experiment considering phase jitter distortion demonstrated that the sEPSM proposed by Jørgensen and Dau (2011) is limited in this condition, which affected mainly the spectral structure of speech. However, predictions from the two modified versions considered here, the 2D-sEPSM and the sEPSM^X, suggested that the limitation was due to the lack of an across (audio-) frequency mechanism. The across-frequency mechanism of the 2D-sEPSM was included in the 2D spectro-temporal modulation frequency selective stage inspired by the STMI. While this approach may contain all relevant information, it may also be based on an overly complex internal representation. The sEPSM^X applied a one-dimensional (temporal only) modulation-frequency selective processing (as in the original sEPSM), but included an across-frequency mechanism in the integration stage of the model. While both approaches produced similar outcomes in the conditions considered in the present study, the conceptual simplicity may favor the sEPSM^X over the 2D-sEPSM.

CONCLUSION

Two modified versions of the speech-based envelope power spectrum model, differing only in their across audio-frequency processing, accounted well for the effects of reverberation, spectral subtraction and phase jitter on speech intelligibility. The results support the hypothesis that the SNR_{env} is a key component in a model of speech intelligibility prediction. Moreover, it was demonstrated that a purely temporal modulation frequency selective process, combined with an across audio-frequency weighting, was sufficient to account for the effects of the phase jitter distortion. A joint two-dimensional modulation filterbank might thus not be required.

ACKNOWLEDGMENTS

This research was supported by the National Science and Engineering Research Council of Canada (NSERC).

REFERENCES

- Berouti, M., Schwartz, R., and Makhoul, J. (1979). "Enhancement of speech corrupted by acoustic noise", *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '79*. 208–211.
- Chi, T., Gao, Y., Guyton, M. C., Ru, P., and Shamma, S. (1999). "Spectro-temporal modulation transfer functions and speech intelligibility.", *J Acoust Soc Am* **106**, 2719–32.

- Christensen, C. L. (2009). *Odeon Room Acoustics Program, Version 10.1, User Manual, Industrial, Auditorium and Combined Editions*, ODEON A/S, Kgs. Lyngby, Denmark, URL <http://www.odeon.dk/pdf/OdeonManual10.pdf>, last viewed 02/12/2011.
- Dubbelboer, F. and Houtgast, T. (2007). "A detailed study on the effects of noise on speech intelligibility.", *J Acoust Soc Am* **122**, 2865–71.
- Elhilali, M., Chi, T., and Shamma, S. A. (2003). "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility", *Speech Communication* **41**, 331–348.
- Goldsworthy, R. L. and Greenberg, J. E. (2004). "Analysis of speech-based speech transmission index methods with implications for nonlinear operations", *J Acoust Soc Am* **116**, 3679–3689.
- Houtgast, T. and Steeneken, H. (1985). "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria", *J Acoust Soc Am* **77**, 1069.
- Houtgast, T., Steeneken, H. J. M., and Plomp, R. (1980). "Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics", *Acustica* **46**, 60–72.
- IEC (2003). *60268-16-2003 Sound system equipment — Part 16: Objective rating of speech intelligibility by speech transmission index*, 3rd edition (International Electrotechnical Commission, Geneva, Switzerland).
- Jørgensen, S. and Dau, T. (2011). "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing", *J Acoust Soc Am* **130**, 1475–1487.
- Ludvigsen, C., Elberling, C., and Keidser, G. (1993). "Evaluation of a noise reduction method—comparison between observed scores and scores predicted from STI.", *Scandinavian Audiology. Supplementum* **38**, 50–5.
- Nielsen, J. B. and Dau, T. (2009). "Development of a Danish speech intelligibility test.", *International journal of audiology* **48**, 729–41.
- Payton, K. L. and Braida, L. D. (1999). "A method to determine the speech transmission index from speech waveforms.", *J Acoust Soc Am* **106**, 3637–48.
- Rhebergen, K. S. and Versfeld, N. J. (2005). "A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners", *J Acoust Soc Am* **117**, 2181.
- Steeneken, H. J. and Houtgast, T. (1980). "A physical method for measuring speech-transmission quality.", *J Acoust Soc Am* **67**, 318–26.