Lavandier, M., and Culling, J. F. (**2010**). "Prediction of binaural speech intelligibility against noise in rooms" J. Acoust. Soc. Am. **127**, 387-399.

Lavandier, M., Culling, J. F., and Jelfs, S. (**2010**). "Prediction of reverberant speech intelligibility against multiple noise interferers in rooms: Binaural useful-to-detrimental ratios (A)" J. Acoust. Soc. Am. **128**, 2361.

Lochner, J. and Burger, J. (**1964**). "The influence of reflections on auditorium acoustics" J. Sound Vib. **1**, 426-454.

Peissig, J. and Kollmeier, B. (**1997**). "Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners" J. Acoust. Soc. Am. **101**, 1660-1670.

Rennies, J., Brand, T., and Kollmeier, B. (**2011**). "Prediction of the inuence of reverberation on binaural speech intelligibility in noise and in quiet" J. Acoust. Soc. Am. (in press).

Steeneken, H. J. M. and Houtgast, T. (**1980**). "A physical method for measuring speech-transmission quality" J. Acoust. Soc. Am. **67**, 318-326.

van Wijngaarden, S. J., and Drullman, R. (**2008**). "Binaural intelligibility prediction based on the speech transmission index" J. Acoust. Soc. Am. **123**, 4514-4523.

vom Hövel, H. (**1984**). "Zur Bedeutung der Übertragungseigenschaften des Außenohres sowie binauralen Hörsystems bei gestörter Sprachübertragung" (On the importance of transmission properties of the outer ear and the binaural auditory system for disturbed speech transmission), doctoral dissertation, RWTH Aachen.

Wagener, K., Brand, T., and Kollmeier, B. (**1999**). "Entwicklung und Evaluation eines Satztests für die deutsche Sprache II: Optimierung des Oldenburger Satztests (Development and evaluation of a German sentence test II: Optimization of the Oldenburg sentence test)" Z. Audiol. **38**, 44-56.

Warzybok, A., Rennies, J., Brand, T., Doclo, S. and Kollmeier, B. (**2011**). "Effects of spatial and temporal integration of early reflections on speech intelligibility" J. Acoust. Soc. Am. (submitted).

Predicting speech intelligibility in adverse conditions: evaluation of the speech-based envelope power spectrum model

SØREN JØRGENSEN AND TORSTEN DAU

Centre for Applied Hearing Research, Technical University of Denmark, DK-2800 Lyngby, Denmark

The speech-based envelope power spectrum model (sEPSM) [Jørgensen and Dau (2011). J. Acoust. Soc. Am., **130** (3), 1475–1487] estimates the envelope signal-to-noise ratio ($\mathrm{SNR}_{\mathrm{env}}$) of distorted speech and accurately describes the speech recognition thresholds (SRT) for normal-hearing listeners in conditions with additive noise, reverberation, and nonlinear processing by spectral subtraction. The latter represents a condition where the standardized speech intelligibility index and speech transmission index fail. However, the sEPSM is limited to stationary interferers due to the fact that predictions are based on the long-term $\mathrm{SNR}_{\mathrm{env}}$. As an attempt to extent the model to deal with fluctuating interferers, a short-time version of the sEPSM is presented. The $\mathrm{SNR}_{\mathrm{env}}$ of a speech sample is estimated from a combination of $\mathrm{SNR}_{\mathrm{env}}$-values calculated in short time frames. The model is evaluated in adverse conditions by comparing predictions to measured data from [Kjems *et al.* (2009). J. Acoust. Soc. Am. **126** (3), 1415-1426] where speech is mixed with four different interferers, including speech-shaped noise, bottle noise, car noise, and cafe noise. The model accounts well for the differences in intelligibility observed for the different interferers. None of the standardized models successfully describe these data.

INTRODUCTION

Models of speech intelligibility can be very useful as tools for investigating which features of the physical speech signal are crucial for understanding the speech in a noisy background. Moreover, an accurate prediction metric is of great relevance in practical applications such as hearing-aid and telecommunication development. Current intelligibility metrics include the articulation index (AI) and its successor the speech intelligibility index (SII). SII-based metrics estimate the effective amount of audible speech information in a number of frequency bands, from the long-term frequency spectra of speech and noise. The audible information is weighted by an empirically determined importance function, describing the relative importance of the individual frequency bands to intelligibility. This approach can predict the intelligibility of speech subjected to low-pass and high-pass filtering and the effects of different stationary noise backgrounds (Kryter, 1962). However, the SII-metric is based on frequency information only, and cannot be successfully applied to conditions with reverberation. As an alternative, the speech transmission index (STI)

Proceedings of ISAAR 2011: Speech perception and auditory disorders. 3rd International Symposium on Auditory and Audiological Research. August 2011, Nyborg, Denmark. Edited by T. Dau, M. L. Jepsen, J. Cristensen-Dalsgaard, and T. Poulsen. ISBN 87-990013-3-0. EAN 9788799001330. The Danavox Jubilee Foundation, 2012.

estimates the integrity of the long-term temporal modulation content of speech. This approach makes it possible to account for room coloration such as reverberation, making this metric very useful for evaluating room acoustics in terms of speech intelligibility. However, both the SII and STI metrics are limited to predicting effects of stationary and linear distortions; they typically come short when noisy speech is processed by noise-reduction algorithms such as spectral subtraction, (Ludvigsen *et al.,* 1993; Dubbelboer and Houtgast, 2007). One hypothesis for the shortcomings is that the metrics do not include the effect of the noise-reduction processing on the noise-part of the noisy speech (Dubbelboer and Houtgast, 2007, 2008). In line with this hypothesis, Jørgensen and Dau (2011) presented a new metric denoted the envelope signal-to-noise ratio (SNR_{env}). This metric quantifies the ratio between the useful speech envelope power and the intrinsic noise envelope power within the noisy speech signal. The SNR_{env} therefore captures the changes to the noise envelope modulations induced by the noise-reduction processing, which is not included in the SII or STI. The SNR_{env} is determined using the speech-based envelope power spectrum model (sEPSM) where the key component is modulation-frequency selective processing of the speech envelope. Here, key aspects of the sEPSM are presented and the model is evaluated in adverse conditions, including stationary and fluctuating interferers as well as linear and non-linear distortions.

MODEL DESCRIPTION

The processing structure of the sEPSM is illustrated in Fig. 1A. The first stage is a bandpass filterbank comprised of 22 gammatone filters with ERB bandwidth and one-third octave spacing, covering the range from 63 Hz to 8 kHz. The temporal envelope of each filter output is extracted via Hilbert-transformation and in turn analyzed by a modulation bandpass filterbank. The long-term integrated ac-coupled envelope power is then calculated from the output of each modulation filter. For each modulation channel, the SNR_{env} is calculated from the envelope power of noisy speech (P_{S+N}) and noise alone (P_N):

$$SNR_{env} = \frac{P_{S+N} - P_N}{P_N} \qquad \text{(Eq. 1)}$$

The resulting envelope-SNR values are combined across modulation filters and across gammatone filters using an integration model from Green and Swets (1988). An absolute sensitivity threshold is included such that only gammatone channels that are excited above the absolute hearing threshold are processed further in the model. The overall SNR_{env} is converted to the percentage of correctly recognized speech items using the concept of a statistically "ideal observer". The ideal observer-stage contains two parameters that reflect the response set-size and the redundancy of a given speech material (see Jørgensen and Dau (2011) for details).

The scheme for predicting intelligibility of processed noisy speech is shown in Fig. 1B. Noisy speech and noise alone (assumed available separately) are passed through some transmission channel under test, such as a room with reverberation, and the stimuli are analyzed by the sEPSM. Here, the noise alone represents an estimate of the intrinsic noise within the noisy speech. Figure 1C illustrates the resulting effect of the transmission channel on the SNR_{env} (top panel) and on the corresponding predicted percent correct (bottom panel) as a function of the input SNR. By comparing predictions with and without the transmission channel in the signal path, the change in intelligibility can be estimated. For instance, the change in speech recognition threshold, ΔSRT is estimated from the corresponding shift (in terms of the input SNR) at the 50 % point of the predicted psychometric functions.

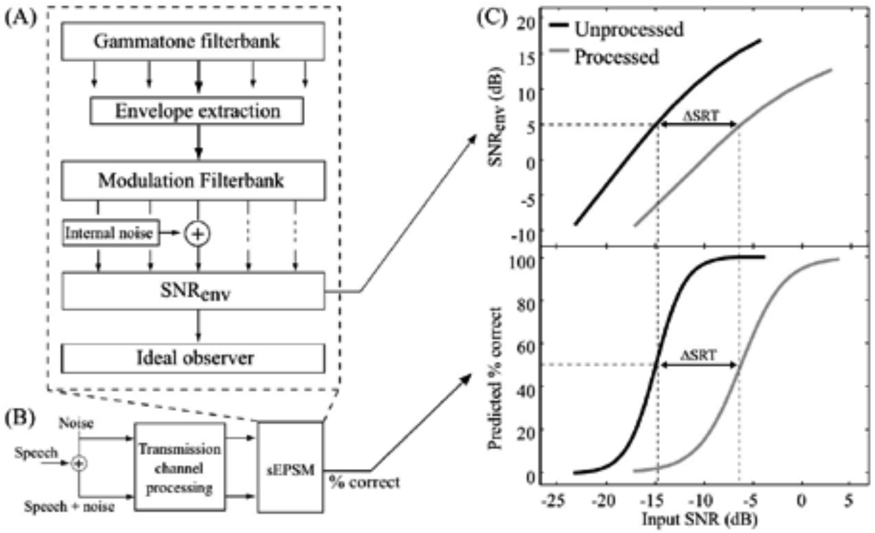

Fig. 1: (A) Block-diagram of the sEPSM processing structure. **(B)** Scheme for predicting speech intelligibility using the sEPSM. **(C)** SNR_{env} as a function of the input SNR (top panel) and the corresponding predicted percentage of correct responses (bottom panel).

PREDICTING INTELLIGIBLITY OF PROCESSED NOISY SPEECH

Model predictions were compared to intelligibility data of processed noisy speech by measuring the speech recognition thresholds (SRT) corresponding to 50% correctly understood sentences from the CLUE test (Nielsen and Dau, 2009). In one experiment, sentences were mixed with a speech-shaped noise and convolved with simulated room impulse responses having reverberation times corresponding to T_{30} = 0, 0.4, 0.7, 1.3 and 2.3 seconds. In a second experiment, the noisy sentences were processed by a spectral subtraction algorithm defined by Berouti *et al.* (1979):

$$\hat{S}(f) = \sqrt{U_{S+N}(f) - \alpha \hat{U}_N(f)} \qquad \text{(Eq. 2)}$$

$\hat{S}(f)$ denotes the estimated clean-speech magnitude spectrum, $\hat{U}_N(f)$ is an estimate of the noise power spectrum, $U_{S+N}(f)$ is the power spectrum of the noisy speech and α denotes the over-subtraction factor which controls the amount of subtraction. The experimental parameter was α, taking the values: 0, 0.5, 1, 2, 4 or 8.

Results

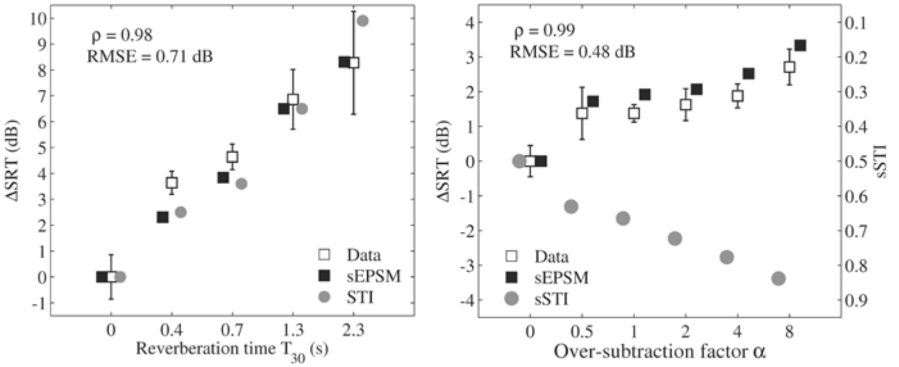

Fig. 2: Left: change in SRT as a function of the reverberation time. **Right**: change in SRT as a function of the over-subtraction factor α. The linear correlation coefficient (ρ) and root-mean-square error (RMSE) between the data end the sEPSM predictions are indicated on each panel. STI predictions are indicated as closed gray circles.

Figure 2 (left panel) shows results from experiment one with ΔSRT as a function of the reverberation time. The open squares represent data averaged across six listeners where the SRT in the reference condition (T_{30} = 0) was found at an SNR of -3 dB, consistent with data from Nielsen and Dau (2009). The vertical bars indicate +/- one standard deviation of the listeners' mean SRT and amount to 0.9 dB on average. The SRT increases with increasing degree of reverberation consistent with the data by Duquesnoy and Plomp (1980). Predictions from the sEPSM (closed squares) and STI (closed circles) also show an increase of SRT with increasing reverberation time, in good agreement with the data. Both metrics appear to capture the effect of reverberation on intelligibility of noisy speech.

Figure 2 (right panel) shows results from the experiment with noisy speech processed by spectral subtraction. Here, the ΔSRT averaged across four normal-hearing listeners is increased for all α > 0, reflecting a reduced speech intelligibility compared to the reference condition without spectral subtraction (α = 0). Such reduction in intelligibility is consistent with data from Ludvigsen *et al.* (1993).

The filled squares represent predictions by the sEPSM, showing an increase of ΔSRT which agrees well with the measured data. In contrast, the corresponding speech-based STI (indicated on the right ordinate) is increased in all conditions of spectral subtraction, compared to the reference condition, predicting an increase in speech intelligibility. The STI thus fails to account for the measured data.

Even though the two models are consistent in predicting effects of reverberation, they completely disagree in the case of spectral subtraction processing, with only the sEPSM being in line with the data. The critical difference between the STI and the SNR$_{\mathrm{env}}$ metric used in the sEPSM is that the SNR$_{\mathrm{env}}$ captures the effect of the spectral subtraction processing on the noise modulations, quantified by an increased noise envelope power, which is neglected in the STI. In the two cases studied here, the SNR$_{\mathrm{env}}$ metric appears to be a more general predictor of intelligibility than the STI.

PREDICTING INTELLIGIBLITY IN FLUCTUATING NOISES

The fact that the SNR$_{\mathrm{env}}$ is calculated from the long-term integrated envelope power leads to specific limitations in the abilities of the sEPSM to predict speech intelligibility. An amplitude modulated noise typically has a larger long-term envelope power compared to a stationary noise with the same audio-frequency domain SNR. This leads to a smaller SNR$_{\mathrm{env}}$ for modulated noise compared to stationary noise and the sEPSM would predict a lower intelligibility in modulated noise backgrounds. This contrasts the well known phenomenon of "speech masking release", referring to the increased intelligibility of speech presented in a fluctuating noise compared to a stationary noise with the same long-term SNR (e.g., Festen and plomp, 1990). Typically, speech masking release is explained by the listeners ability to "listen in the dips" of the masker.

Here, it is hypothesized that speech masking release can be explained by an increase of SNR$_{\mathrm{env}}$ during the time periods where the masker's amplitudes are low. This hypothesis is investigated by modifying the sEPSM to estimate the envelope SNR in short time frames. Specifically, the temporal outputs from the modulation filterbank are segmented in 10-ms frames with square windows. For each segment, *i*, and modulation filter, the ac-coupled envelope power of noisy speech and noise alone is calculated and inserted in Eq. (1), yielding the SNR$_{\mathrm{env},i}$ of that particular segment and modulation filter. Integrating SNR$_{\mathrm{env},i}$ -values across modulation and audio filters gives an overall SNR$_{\mathrm{env},i}$ for each temporal segment. The SNR$_{\mathrm{env}}$ of a given sentence is taken as the average SNR$_{\mathrm{env},i}$ across all segments of that sentence. Apart from the segmentation of SNR$_{\mathrm{env}}$, the signal-processing of model is the same as previously described.

Results

Predictions from the short-term sEPSM are compared to data collected by Kjems *et al.* (2009) on DANTALE II-sentences presented in four different noise backgrounds: Bottle noise, Car noise, Cafe noise, and Speech-shaped noise (SSN). The Cafe noise and the SSN have the same long-term frequency spectra, but differ in their temporal characteristics, with the café noise being highly modulated with time. Figure 3 (left panel) shows psychometric functions (solid lines) estimated from measured data and corresponding sEPSM predictions (closed symbols connected by dashed lines). In addition, predictions from the long-term sEPSM for the Cafe noise are shown. There is a good qualitative correspondence between the predictions from the short-term sEPSM and the experimentally determined psychometric functions for all noise types, both in terms of horizontal placement and slope. In contrast, the long-term sEPSM clearly fails for the Cafe noise. It is noted that the ideal-observer parameters were calibrated to the SSN condition, after which, the parameters were fixed and only the noise changed. Figure 3 (right panel) shows a quantitative comparison between the predicted (closed squares) and measured (open squares) SRTs for the four interferers. The short-term sEPSM accounts for the masking release of the fluctuating Cafe noise, although it is slightly overestimated.

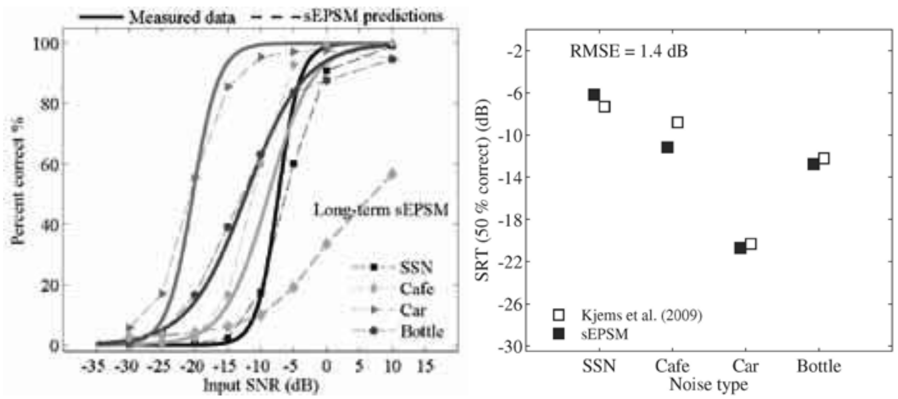

Fig. 3: Left: Psychometric functions (solid lines) estimated from measured data by Kjems *et al.* (2009) and corresponding predictions by the short-term sEPSM (connected symbols) for speech presented in four different noise backgrounds (four shades of gray). Predictions from the long-term sEPSM are shown with a label on the curve. **Right**: SRTs estimated from the measured data (open squares) and predictions by the sEPSM (closed squares) as a function the noise type.

MODEL ANALYSIS

It is investigated how the prediction of speech masking release is reflected in the internal representation of the sEPSM. The top-left panel of Figure 4 shows an example of the temporal waveform of speech mixed with a stationary noise (black) together with the noise alone (gray). These are the stimuli that are input to the sEPSM, although the predictions in Figure 3 are based on an average across 50 different sentences. Similarly, the top-right panel of Figure 4 shows the situation with an amplitude modulated noise. The corresponding segmental $\mathrm{SNR}_{\mathrm{env}}$ is shown in the bottom panels of Fig 4. Comparing the left and right panels, it appears that the $\mathrm{SNR}_{\mathrm{env}}$ is increased during the periods where the amplitude of the modulated noise is low, i.e. in the period between 0.2 and 0.4 s and around 0.8 s. This leads to an increased mean $\mathrm{SNR}_{\mathrm{env}}$ across the whole speech sample which in turn leads to an increase in predicted intelligibility. Masking release is thus predicted by the model due to a time-local increase of the short-term $\mathrm{SNR}_{\mathrm{env}}$ during the dips of the masking noise.

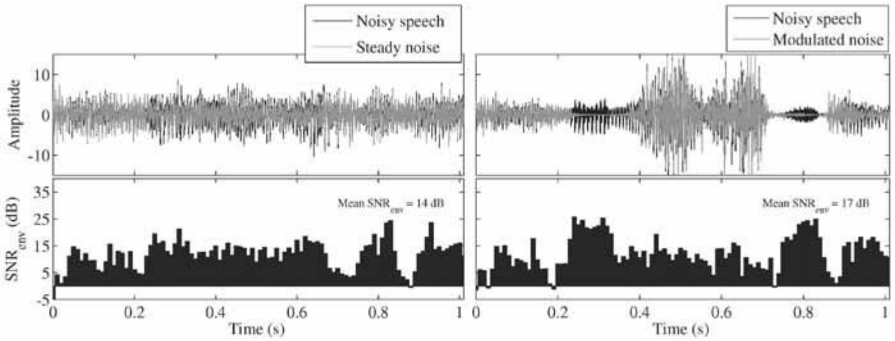

Fig. 4: Left: Temporal waveform (top panel) of a sentence mixed with a stationary noise (black) together with the noise alone (gray) and the corresponding $\mathrm{SNR}_{\mathrm{env}}$ (bottom panel). **Right**: The same situation as the left panel but with speech mixed with a modulated noise. Comparing the right and left panel, it appears that the $\mathrm{SNR}_{\mathrm{env}}$ is increased during the dips of the modulated masker, i.e. around 0.3 and 0.8 seconds.

DISCUSSION AND CONCLUSION

The sEPSM could accurately predict the change in intelligibility of noisy reverberant speech, similar to the classical STI metric. In addition, the sEPSM predicted data for noisy speech processed by a spectral subtraction algorithm where the STI failed completely. The gain over STI is the $\mathrm{SNR}_{\mathrm{env}}$-metric that includes the effect of the processing on the intrinsic noise envelope power, which increases after spectral subtraction, leading to a decrease of $\mathrm{SNR}_{\mathrm{env}}$ and thus to a decreased predicted intelligibility. However, the sEPSM has shortcomings in conditions of fluctuating maskers, since predictions are based on the long-term envelope power. A solution to this is a short-term version that estimates the $\mathrm{SNR}_{\mathrm{env}}$ in short time frames. The short-term sEPSM could accurately predict the psychometric functions (percent correct versus SNR) for speech presented in four different noises, including a highly fluctuating Cafe noise. Neither the STI nor the SII are able to do this (Christiansen *et al.*, 2010). A model analysis showed that the masking release predicted by the model, in case of the fluctuating noise, was caused by an increased $\mathrm{SNR}_{\mathrm{env}}$ in the dips of the masker. The increase therefore occurs at higher modulation frequencies than the masker fluctuation frequency. The short-term calculation of $\mathrm{SNR}_{\mathrm{env}}$ may, however, change the models ability to accurately capture changes to slow modulations, e.g. induced by spectral subtraction. It is therefore possible that the short-term sEPSM will not predict the same as the long-term version in the conditions shown in Fig 3. This could indicate that different timescales are necessary to account for the short-term and long-term effects.

It is an ongoing research topic whether speech masking release is dominated by speech envelope information or temporal fine structure (TFS) information. The sEPSM relies only on envelope information. Nevertheless, it predicts the masking release observed for the fluctuating Cafe noise. To the extent that sEPSM correctly models the auditory system, this suggests that envelope cues are more important for masking release than TFS, at least for these particular speech and noise combinations. This is in line with recent behavioral findings that TFS information may not be the key to speech masking release. Rather, it may facilitate the segregation of masker and target based on differences in fundamental frequency.

REFERENCES

Berouti, M., Schwartz, R., and Makhoul, J. (**1979**). “Enhancement of speech corrupted by acoustic noise” ICASSP **4**, 208-211.

Christiansen C., Pedersen, M. S., Dau,T. (**2010**). “Prediction of speech intelligibility based on an auditory preprocessing model” Speech. Commun.,**52**, 678–692.

Dubbelboer, F., and Houtgast, T. (**2007**). “A detailed study on the effects of noise on speech intelligibility” J. Acoust. Soc. Am. **122**, 2865-2871.

Dubbelboer, F., and Houtgast, T. (**2008**). “The concept of signal-to-noise ratio in the modulation domain and speech intelligibility” J. Acoust. Soc. Am. **124**, 3937-3946.

Duquesnoy, A. J., and Plomp, R. (**1980**). “Effect of reverberation and noise on the intelligibility of sentences in cases of presbyacusis” J. Acoust. Soc. Am. **68**, 537-544.

Festen, J. M., and Plomp, R. (**2011**). “Effects of fluctuating noise and interfering speech on the speech reception threshold for impaired and normal hearing” J. Acoust. Soc. Am., **88** (4), 1725–1736.

Green, D. M. and Swets, J. A. (**1988**). *Signal Detection Theory and Psychophysics* (Peninsula Publishing, Los Altos California), 238-239.

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** (3), 1475–1487.

Kjems, U., Boldt, J. B., Pedersen, M. S., Lunner, T., and Wang, D., (**2009**), “Role of mask pattern in intelligibility of ideal binary-masked noisy speech” J. Acoust. Soc. Am. **126** (3), 1415–1426.

Kryter, K. D., (**1962**), “Validation of the Articulation Index” J. Acoust. Soc. Am. **34** (11), 1698–1702.

Ludvigsen, C., Elberling, C., and Keidser, G. (**1993**). “Evaluation of a noise reduction method-comparison between observed scores and scores predicted from STI” Scand. Audiol. Suppl. 38 **22**, 50-55.

Nielsen, J. B. and Dau, T. (**2009**). “Development of a Danish speech intelligibility test” Int. J. Audiol. **48**, 729-741.

Ordinal models of audiovisual speech perception

TOBIAS S. ANDERSEN

Informatics and Mathematical Modelling, Technical University of Denmark, 2800 Lyngby, Denmark

Audiovisual information is integrated in speech perception. One manifestation of this is the McGurk illusion in which watching the articulating face alters the auditory phonetic percept. Understanding this phenomenon fully requires a computational model with predictive power. Here, we describe ordinal models that can account for the McGurk illusion. We compare this type of models to the Fuzzy Logical Model of Perception (FLMP) in which the response categories are not ordered. While the FLMP generally fit the data better than the ordinal model it also employs more free parameters in complex experiments when the number of response categories are high as it is for speech perception in general. Testing the predictive power of the models using a form of cross-validation we found that ordinal models perform better than the FLMP. Based on these findings we suggest that ordinal models generally have greater predictive power because they are constrained by a priori information about the adjacency of phonetic categories.

INTRODUCTION

Speech perception in face-to-face conversation is based not only on hearing the acoustic speech signal but also on lip-reading. Observers tend to integrate audiovisual information across the sensory modalities without being aware of it. In the natural, ecological valid situation where the voice and lip-movements are congruent this facilitates speech perception (Sumby and Pollack, 1954). When an incongruent voice is dubbed onto a video of a talking head observers may perceive a fusion type McGurk illusion in which the perceived phoneme differs both from that mediated by the voice and that mediated by the face (MacDonald and McGurk, 1978; McGurk and MacDonald, 1976). The typical example of this fusion type McGurk illusion is when a voice saying /ba/ is dubbed onto a face saying /ga/ causing observers to hear /da/. Other types of McGurk illusions include combination illusions in which the observer hears both the phoneme mediated by the voice and the phoneme mediated by the face. An example of a fusion illusion is when a voice saying /da/ is dubbed onto a face saying /ba/ which observers tend to hear as /bda/. Visual dominance illusions is another type of McGurk illusions in which observers hear the phoneme mediated by the lip-movements rather than that mediated by the voice.

Because the influence of vision on hearing in speech perception is so profound understanding how it works may give us fundamental cues to how speech

Proceedings of ISAAR 2011: Speech perception and auditory disorders. 3rd International Symposium on Auditory and Audiological Research. August 2011, Nyborg, Denmark. Edited by T. Dau, M. L. Jepsen, J. Cristensen-Dalsgaard, and T. Poulsen. ISBN 87-990013-3-0. EAN 9788799001330. The Danavox Jubilee Foundation, 2012.