In the development of modern telecommunication systems, it is common practice to use computer models in the place of human listeners for the evaluation of transmission quality. These models usually consider aspects of speech quality, while little or no attention is paid to speech intelligibility in adverse conditions. This project investigates if a speech intelligibility model can be applied to predict the intelligibility of speech transmitted over mobile phones. The Danish Dantale II speech material was mixed with three different kinds of background noise and transmitted through four different mobile phones and recorded at the receiver end. Speech intelligibility of the recordings was assessed for six normal-hearing listeners and the intelligibility score was compared to the predictions of the model. The results showed a good correspondence between measured data and predictions in conditions with speech-shaped noise and traffic noise, but the model failed with pub noise, where speech intelligibility was overestimated. The results serve as a proof-of-concept while additional investigations are required to clarify the limitations of the prediction model.
About the publication
This publication and possible comments and discussions can be downloaded from www.soundtechnology.dk.

The content of this publication reflects the authors’ point of view and not necessarily the view of the Danish Sound Technology Network as such.

Copyright of the publication belongs to the authors. Possible agreements between the authors might regulate the copyright in detail.

About the network
Danish Sound Technology Network is an innovation network funded by the Danish Agency for Science, Technology and Innovation. The Network is hosted by the Technical University of Denmark and is headed by Director, Associate Professor, PhD Jan Larsen.

The vision of the network is that Denmark is a leading country with regards to sound technology in terms of knowledge, research and education. Danish sound technology will be the epitome of high quality in products and services, as well as in physical rooms and social contexts.

Network membership is free of charge and open for all. Registration at www.soundtechnology.dk

This publication is the result of an innovation project, an instrument to strengthen the cooperation between knowledge institutions and private companies. The primary goal is to promote innovation by combining accessible/existing research and technologies with creative uses in order to facilitate the creation of new products, services or experiences. Innovation projects are mainly short term feasibility studies conducted on a pre-competitive level.
This publication is the result of an innovation project entitled “Prediction of Speech intelligibility in modern telecommunication”. The project was financed by the Danish Sound Technology Network through a grant from the Danish Agency for Science, Technology and Innovation. The project was conducted in the period 15. Apr. 2012 – 14. Oct. 2012 and managed by Technical University of Denmark (DTU), project managers were Søren Jørgensen and Torsten Dau. Additional project participants were: Jens Cubick (DTU), Wookeun Song (Brüel & Kjær) and Morten Løve Jepsen (Widex).

INTRODUCTION

BACKGROUND

Traditionally, mobile phones have been evaluated mainly in terms of speech quality. In many cases the quality is assessed experimentally using subjective judgments based on a categorical scale (ITU-T P.800 1996). The judgments are averaged over a panel of listeners and referred to as the mean opinion score (MOS).

Even though the assessment of speech quality is more commonly used in modern telecommunication, speech intelligibility is equally important for characterizing the performance of the devices. This is especially the case at low signal-to noise ratios (SNRs), where speech quality is typically judged to be poor. Measures of speech intelligibility may therefore provide an additional descriptor of the performance of the transmission system in such situations.

Several objective speech intelligibility prediction metrics have been suggested, starting with the articulation index (AI) and its successor, the speech intelligibility index (SII, ANSI S3.5 - 1997). While these metrics were designed to account for the reduced bandwidth of early telecommunication systems, they cannot be applied straight away in situations where noisy speech has been processed by nonlinear noise reduction, which is standard in modern digital mobile phones, because the definitions of the target and masker signals are no longer clear following the non-linear processing (Loizou and Ma, 2011).

Recently, the speech-based envelope power spectrum model (sEPSM; Jørgensen & Dau 2011), which is based on the signal processing of the human auditory system, has been shown to account for some conditions with nonlinear noise reduction.

OBJECTIVE

The aim of this project was to apply the sEPSM for speech intelligibility prediction in mobile phones, using an existing measurement setup available at Brüel & Kjær. The project is a proof of concept, i.e., whether this model can be used as a reliable predictor of mobile phone intelligibility such that it can be used to differentiate between phones and potentially different settings of the same phone.

IMPACT/EFFECT

At present, there is no reliable predictor of speech intelligibility in mobile phone transmission systems. If the sEPSM would turn out to provide a reliable prediction of speech intelligibility in this context, it might decrease the need for expensive and time-consuming listening tests in the mobile phone industry and
therefore could become a powerful tool in product development, e.g. when a new algorithm needs to be evaluated.

METHODS AND RESULTS

THEORY

In mobile communication situations, the transmitted speech signal is often compromised by the presence of background noise at the sending end, e.g. when the talker is situated in a car or in a pub. Even if situated in a quiet room, the listener can have difficulties understanding the transmitted speech. The speech intelligibility is quantified by the percentage of speech tokens that the listener can understand, and the speech reception threshold (SRT₅₀) is the SNR at which the listener understands 50% of the target speech. The sEPSM was developed to predict the percentage of correctly understood speech tokens in noisy situations. The model is based on knowledge about the human hearing, and mimics, to some extent, the effective signal processing of the auditory system. It relies on the finding that the relatively slow amplitude fluctuations in the envelope of speech, the so-called modulations, play a crucial role for speech intelligibility in adverse conditions (e.g., Drullman, 1995). The sEPSM estimates the signal-to-noise ratio in the envelope domain (SNRₚₑᵥ), i.e., the relative power of the speech and noise envelope fluctuations at the output of an auditory model. The greater the envelope fluctuations of the noise, the lower the SNRₑᵥ and the more speech information will be masked, resulting in a poor speech intelligibility predicted by the model.

Unlike speech quality measures that require access to the clean speech signal, the sEPSM estimates the SNRₑᵥ from the noisy speech and the noise alone. This could be an advantage in practical applications, because it would allow for the prediction of intelligibility also in situations, where the clean signal is not available, e.g. when the receiver is in a noisy environment. However, it is assumed in the model, that the noise alone and the noise-part of the noisy speech are identical. It is, therefore, critical to have a good estimate of the noise alone, in order to ensure valid simulations.

Details on the model stages can be found in Jørgensen and Dau (2011).

METHOD

Rationale

A Matlab-implementation of the sEPSM is evaluated for predicting intelligibility of speech transmitted by mobile phones by comparing predictions to psychoacoustic data obtained for normal-hearing listeners with four different mobile phones. In addition, the sEPSM are compared to predictions from the standardized SII-calculation procedure.

Listeners

Six native Danish normal-hearing listeners participated in the study and were paid an hourly wage by DTU.

Speech material

The Dantale II speech corpus (Wagener et al. 2003) was used for measuring the speech intelligibility. It consists of 160 five-word sentences with an identical syntactical structure (name + verb + numeral + adjective + object). Unlike in the original Dantale II test procedure, the SNR was kept constant within each sentence list to reduce the number of necessary recordings compared to an adaptive procedure.
Stimuli

Sound samples were recorded in an IEC-standardized listening room (IEC 268-13, 1985) with the mobile phone attached to a B&K Head and torso simulator 4128-D, with a handset positioned in a setting according to (ETSI EG 202 396-1 2008, cf. Figure 1). Figure 1 shows a schematic of the recording setup, Figure 2 and Figure 3 show photos of the live setup. Four loudspeakers played the background noise with delays between 0 and 29 ms for each channel for de-correlation. The mobile phone was connected to a network simulator through a locally established mobile phone network and the electrical output signal, as it was send by a given phone, was recorded using Matlab. Individual recordings of all 160 sentences mixed with the three different noise types at the four SNRs for each phone were obtained. In addition, the signal from a B&K ¼-inch microphone positioned close to the microphone of the mobile phone was recorded (cf. Figure 3).

![Figure 1: Recording setup for the speech test stimuli according to (ETSI EG 202 396-1 2008). The signals from the HATS'ear-microphones were not used here. Drawing courtesy of Brüel & Kjær A/S.]

![Figure 2: Photo of recording-setup.]

![Figure 3: Photo of mobile phone mounting on HATS.]

Apparatus and procedure
The psychoacoustic measurements were conducted in a double-walled listening booth at the Centre for Applied Hearing Research, DTU. Both the listeners and the experimenter were seated inside the booth, where the sentences were presented to the test subject’s right ear via Sennheiser HD 650 headphones. The headphones were equalized to have a flat frequency response at the ear reference point and the stimuli were filtered with the modified intermediate reference system (IRS) “receive” transfer function (ITU-T P.830 1996). The task of the listeners was to repeat as many words of the presented sentences as possible and to guess, if unsure. The experimenter then scored the responses on a computer screen hidden from the test subject.

**Conditions**

Four mobile phones from three different manufacturers were considered, here denoted by A, B, C, and D, in three types of background noise: Dantale II Speech shaped noise (SSN) (Wagener et al. 2003), Traffic Crossroads, and Pub noise from the noise database provided in (ETSI EG 202 396-1 2008). Moreover, two reference conditions were considered: the broadband reference microphone signal (Ref) and the reference microphone signal filtered with the modified IRS bandpass filter (Ref BP), simulating the reduced bandwidth of the transmission channel. Four SNRs were tested in all conditions, where the SNRs were determined in a pilot study to cover the range from 40%-90% intelligibility for a given condition. All conditions and SNRs were tested twice for every listener. In total, 2 x 48 (3 phones x 3 noises x 4 SNRs + 3 additional SSN conditions x 4 SNRs) lists of 10 sentences were used for each listener. Additional 24 lists were used for training, resulting in a total training and test time of approx. six hours per listener, divided into three sessions of two hours.

**Simulations**

Model simulations were performed using 30 of the recorded noisy sentences for each condition. The speech and noise alone at the output of the mobile phone were estimated from the noisy speech mixture using the method presented by Hagerman and Olofsson (2004). Assuming two input signals:

\[
a_{in} = s + n \quad \text{and} \quad b_{in} = s - n
\]

where \(s\) and \(n\) denote the speech and noise signals at the input. The output signals can be estimated as:

\[
s'_{out} + \frac{1}{2}E_1 = \frac{1}{2}(a_{out} + b_{out}) \quad \text{and} \quad n'_{out} + \frac{1}{2}E_2 = \frac{1}{2}(a_{out} - b_{out})
\]

Assuming that the error terms \(E_1\) and \(E_2\) are small, \(s'_{out}\) and \(n'_{out}\) denote the estimated speech and noise alone at the output of the mobile phones. The error terms can be estimated using the methods described in Hagerman and Olofsson (2004). Here, they are neglected the output signals are estimated as:

\[
s'_{out} = \frac{1}{2}(a_{out} + b_{out}) \quad \text{and} \quad n'_{out} = \frac{1}{2}(a_{out} - b_{out})
\]

For the SII simulations, the Matlab-implementation of the SII-calculation procedure available at [http://sii.to/html/programs.html](http://sii.to/html/programs.html) was applied. A separate simulation for each sentence was performed for all conditions, and the resulting SII was averaged across the 30 sentences. The SII procedure requires the determination of a transfer function from SII to measured intelligibility score.
The transfer function used here was taken from Amlani et al. (2002) and is of the form:

$$P_c = 1 - 10^{-\left(\text{SII} + K\right)/Q}$$

where $K$ and $Q$ are constants to be fitted. The best fitting values of $K$ and $Q$ were found via minimum-mean-square fitting to the psychoacoustic data plotted as a function of the corresponding SII. Figure 4 shows the psychoacoustic data for the two reference conditions in SSN as a function of the corresponding SII (triangles), together with the best-fitting transfer function (dashed line). In addition, the circles show the data obtained with the A, C, and B in all noise conditions as a function of the corresponding SII, together with the best fitting transfer function (solid line). The triangles and circles fall in two distinct clusters, with two different best-fitting transfer functions. If the reference (dashed) transfer function was used here, the SII would fail in all conditions except the reference conditions for which the transfer function was fitted, demonstrating an inherent limitation of the SII. Since the aim of this project was to predict the intelligibility of mobile phones, the transfer function fitted to the mobile phone data (solid line) was used. This means that the SII cannot be expected to account for the reference conditions. However, ideally, a predictive model should be calibrated to a reference condition without any processing and should be able to predict the intelligibility in other conditions using the same set of parameters.

To be consistent, both the parameters of the sEPSM and SII were adjusted to account optimally for the conditions with the phones A, C, and B in all three noise-types using one set of parameters, i.e. all model parameters were fixed for all conditions.

**Postprocessing of the data**

A psychometric function with two parameters, the slope ($S_{50}$) and the 50-\% point ($SRT_{50}$), of the form (Wagener et al, 2003):

$$pc(SNR) = \left[1 + e^{-S_{50}(SNR-SRT_{50})}\right]^{-1}$$

was fitted to the results of each individual listener in a given condition, and the parameters were averaged across the listeners to obtain an average psychometric function for all listeners in that condition. Similarly, a psychometric function of this form was fitted to the predictions from the speech intelligibility models. In the following, all comparisons of measured data and model predictions were based on the psychometric functions.
RESULTS

Psychoacoustic data

Figure 5 shows the average percentage of correct responses from the individual listeners across the two trials as a function of the SNR, for the Ref (left panel) and the Ref BP (middle panel) conditions in SSN. The solid black line indicates the average psychometric function, representing the measured data. In the Ref-condition, the six listeners were fairly consistent, illustrated by the small vertical spread of the data. When the bandpass filter was applied, the variability across the listeners increased (note that listener GAM was accidently not measured in the Ref BP-condition), demonstrating that the bandpass filtering alone affected the variability across the listeners. The right panel of Figure 5 shows the individual data and average psychometric function with mobile phone A in SSN; here, the variation across the listener’s responses was even greater than in the Ref BP-condition, especially at low SNRs. This was the case for most conditions with mobile phones.

Figure 6 shows the average psychometric functions obtained from the measured data in conditions with SSN (left), Pub (middle), and Traffic (right) noise. The shaded areas represent plus and minus one standard error of the estimated $S_{50}$ and $SRT_{50}$. One subject was excluded from the following statistical analysis because it was not possible to determine a psychometric function for his results in the phone C-conditions. A one-way analysis of variance with subject-SRT$_{50}$ and mobile phone as entries revealed that the phones were significantly different in the SSN-conditions ($p<0.001$), while no significant difference of SRT$_{50}$ across phones was found for the other noise-types. Thus, the SSN was best suited to distinguish between the mobile phones. A multiple comparison analysis with Bonferroni correction revealed that phone D was significantly different (on a 5%-level) from phones A and C, and that phone B was significantly different from C.

There was a trend towards the SRT$_{50}$ in the Ref BP-condition (green) being higher than the broadband condition (turquoise), but this was not statistically significant. A two-way ANOVA with noise-type (SSN, Pub, and Traffic) and mobile phone (A, B, and C) as factors and subjects as repetitions showed a significant effect of noise-type and mobile phone, but no significant interaction. The effect of phone was generated by the SSN conditions, while the effect of noise was dominated by the phones A and C.
Comparison of model predictions and psychoacoustic data

Figure 7 shows the SRT$_{50}$ determined from the measured psychometric functions (open squares), as a function of the different mobile phone conditions. The error-bars denote the standard error across the listener’s individual SRT$_{50}$. For the SSN-conditions (left panel), the rank-order was such that phone D provided the best intelligibility, followed by B, C, and A. This pattern of results was also found for the Pub (middle panel) and Traffic (right panel) noises, however, the data for the three phones were not significantly different. The root-mean-square errors (RMSE) between the predictions (sEPSM and SII) and the measured data are indicated in the top left corner of each panel. The sEPSM follows the trend of the data for the SSN conditions, except for a bias towards higher SRTs, and phone D where it fails. In comparison, the SII accounts for the phones A, B and D, but fails for phone C and the reference conditions. The RMSE of the sEPSM (2.3 dB) is lower than that of the SII (5.9 dB). This trend was also seen for the Traffic-noise where the sEPSM provided the best predictions in terms of RMSE (1.6 dB), although with a bias towards lower SRTs, and the SII failed for to account for the data.

In the conditions with Pub-noise, the sEPSM failed (RMSE = 6.3 dB), while the SII predictions were in close agreement with the data (RMSE = 1.5 dB).

The three panels of Figure 8 re-plot the measured and predicted SRTs as a function of the noise type for phone A (left panel), B (middle panel), and C (right panel). For all phones, the Traffic noise led to the lowest SRT$_{50}$ (best intelligibility) while the Pub-noise led to the highest SRT$_{50}$ (worst intelligibility). When the results are plotted in this way, the SII predictions showed the lowest RMSE for all phones. However, the higher
RMSE of the sEPSM predictions are dominated by the Pub-conditions, for which the sEPSM systematically underestimated the SRT$_{50}$.

![Figure 8: SRT$_{50}$ as a function of the noise-condition for phone A (left), B (middle) and, C (right).](image)

**CONCLUSION**

The psychoacoustic data showed that it was possible to measure a difference in intelligibility between the four mobile phones, where the SSN-conditions provided the lowest variability across listeners and thereby the most reliable data. Moreover, the sEPSM accounted for the trends in the data in the SSN and Traffic-conditions, while it failed in the Pub-conditions. The sEPSM outperformed the SII model in terms of root-mean-squared error (RMSE) when comparing phones for a given noise-type, except for the Pub-noise. However, if the aim is to predict the effect of different noises, for a given phone, the SII provided the smallest RMSE. Additional model analysis is required to clarify why the models fail in some conditions.

It is concluded that the SSN provides the highest sensitivity to differentiate between phones, while also leading to the most reliable model predictions. It can therefore be recommended to use SSN when differentiating between mobile phones or processing algorithms within the same phone. This is supported by a recent study (Wong et al., 2012), which concludes that the SSN is a good predictor of the intelligibility in more realistic background noises. Thus, from the developer’s point of view, it may be sufficient to consider the SSN noise only, where the prediction models are most reliable.

The results of this project serve as a proof of concept that the sEPSM can be applied to predict intelligibility of mobile phones. However, no finished products or services resulted from this project, and additional development is required to have a complete platform for evaluating intelligibility of speech transmitted by mobile phones.

**REFERENCES**


ETSI EG 202 396-1 (2008), ‘Speech processing, transmission and quality aspects (STQ); Speech quality performance in the presence of background noise; Part 1: Background noise simulation technique and background noise database’.


