Predicting Automatic Speech Recognition Performance over Communication Channels from Instrumental Speech Quality and Intelligibility Scores

Laura Fernández Gallardo[1], *Sebastian Möller*[1], *John Beerends*[2]

[1]Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin, Germany
[2]TNO, The Hague, Netherlands

(laura.fernandezgallardo|sebastian.moeller)@tu-berlin.de, john.beerends@tno.nl

Abstract

The performance of automatic speech recognition based on coded-decoded speech heavily depends on the quality of the transmitted signals, determined by channel impairments. This paper examines relationships between speech recognition performance and measurements of speech quality and intelligibility over transmission channels. Different to previous studies, the effects of super-wideband transmissions are analyzed and compared to those of wideband and narrowband channels. Furthermore, intelligibility scores, gathered by conducting a listening test based on logatomes, are also considered for the prediction of automatic speech recognition results. The modern instrumental measurement techniques POLQA and POLQA-based intelligibility have been respectively applied to estimate the quality and the intelligibility of transmitted speech. Based on our results, polynomial models are proposed that permit the prediction of speech recognition accuracy from the subjective and instrumental measures, involving a number of channel distortions in the three bandwidths. This approach can save the costs of performing automatic speech recognition experiments and can be seen as a first step towards a useful tool for communication channel designers.

Index Terms: automatic speech recognition, speech intelligibility, instrumental speech quality, communication channels

1. Introduction

An increasing number of commercial products and services incorporate automatic speech recognition (ASR) transmitting coded speech signals to a server with strong processing capabilities. A disadvantage of remote ASR based on transmitted speech, however, is the distortion introduced by communication channels, which may severely affect the recognition performance.

Speech transmission channels are designed considering the quality of transmitted signals essentially in order to meet the users' needs and expectations. Although the ASR accuracy is not yet systematically taken into account in the communication channel design process, we believe that ASR can be an additional criterion for choosing among transmission channel elements. In this respect, this paper presents correspondences between speech quality and ASR rates over channels of different bandwidths and using codecs operating at different bitrates. Since the ASR performance can be thought of being strongly connected to the intelligibility aspect of the signal quality, our work also examines the correspondences between speech intelligibility and ASR accuracies. An intelligibility test based on a closed set of vowel-consonant-vowel logatomes with eight alternatives was conducted to obtain subjective scores. Moreover, the POLQA-intelligibility model was applied to compute objective intelligibility scores [1]. The employment of the prediction models for ASR performance proposed in this work can be an alternative to performing ASR experiments, advantageous when no resources are available to train acoustic and language models.

Previous investigations have also studied the relationships between ASR accuracy and transmitted speech quality. The study in [2] compared the subjective MOS and the ASR performance for the AMR-NB codec, showing a general detriment of word recognition with decreasing bitrate. The ASR performance has also been shown to be affected by degradations of speech quality estimated by the Perceptual Evaluation of Speech Quality (PESQ) model [3, 4] and by the E-model for uncorrelated white noise and for signal-correlated noise [5]. The PESQ and the E-model are, respectively, an intrusive signal-based model and a parametric model which can estimate subjective quality ratings in terms of Mean Opinion Scores (MOS).

The automatic evaluation of speech intelligibility has been shown to be successful employing ASR in [6] (considering different NB codecs and bit errors) and in [7, 8, 9] (considering noise conditions). While these works focus on intelligibility modeling, no publication is known to the authors that addresses the prediction of ASR results from subjective and objective intelligibility measures, as presented in this paper.

The speech bandwidth considered in the reviewed studies that employed transmitted speech was limited to conventional narrowband (NB, 300–3,400 Hz), which transmits only a small portion of the frequencies of human speech. With the advent of wideband channels (WB, 50–7,000 Hz), e.g. encountered in IP-based voice communications (VoIP), it has been shown that the added frequency components account for better word intelligibility and higher voice naturalness [10]. Besides, super-wideband transmission (SWB, 50–14,000 Hz), offering an even more extended bandwidth, is currently gaining adoption in the marketplace, usually combined with high-definition video streams.

Signal quality has been found to be improved by 30% when switching from NB to WB [11]. Further benefits of the enhanced bandwidth have been shown for human and automatic speaker recognition [12], speech intelligibility [13, 1], and ASR performance [14]. The superiority of SWB over NB and WB has been revealed in [15] for subjective quality judgments and in [12] for automatic speaker verification. The possible advantage of SWB over WB for ASR still remains to be assessed, which is a secondary goal of the present paper. Possible advantages may partly determine investments on SWB channels for telecommunication operators.

Only the study in [14] is known to the authors that examined the ASR performance comparing NB and WB channels. It was reported that high MOS values (estimated by PESQ) did not always correspond to high recognition rates. Their results revealed almost no variation in the ASR performance across WB codecs. Differently, in this work we apply the Perceptual Objective Listening Quality Assessment (POLQA) [16] and the POLQA-based intelligibility models to compute the instrumental measures to be related to ASR. The POLQA model is the successor of PESQ and is the only ITU-T standard for estimating SWB speech quality. It can operate in NB and in SWB mode, the latter covering a bandwidth wider than that considered in WB-PESQ [17]. The analysis in [18] asserted that the correlation between POLQA and the subjective MOS was higher than for PESQ for WB data.

The POLQA-intelligibility model (V1490intellV2) is a further development of PESQ intelligibility [19] using the latest developments in the objective assessment of speech quality [20, 21]. The main differences between POLQA and POLQA-intelligibility are:

- The zero impact of noise degradations in the silent intervals which have an impact on quality but not on intelligibility
- The difference of the impact of loud degradations in loud speech parts which lead to extremely low-quality speech that still remains intelligible
- Extended modeling of extreme severe degradations which all lead to the same MOS sore of 1.0 while the intelligibility may differ significantly
- Extended modeling of the impact of local severe loss of signal, e.g. time clipping, which may lead to high quality, but not intelligible speech

To the best of our knowledge, no other intelligibility model (e.g. SII, STOI) can satisfactorily predict subjective scores under the effects of communication channel degradations such as those considered in this work.

2. Data preparation

For the study of the ASR performance under different channel distortions, it was required that the speech data were recorded in clean conditions and have a sampling frequency of at least 32 kHz for SWB transmission. Unfortunately, most publicly available databases, with speech sampled at 8 kHz or at 16 kHz, were not suitable for our analysis. The speech data employed in this work was a small portion of the AusTalk database [22]. Only the speech from 37 speakers of Australian English (21 females and 16 males), clean and with a sampling frequency of 44.1 kHz, was available to the authors. Each speaker uttered the same set of 322 words in three separate recording sessions which took place in different days. This speech, totaling around eight hours (after voice activity detection), was selected for the ASR experiments in this work.

Differently, for the intelligibility test, eight VCV logatomes were chosen, varying the middle consonant: “ama”, “aba”, “afa”, “ana”, “apa”, “asa”, “awa”, and “ascha”, pronounced in German. These logatomes were selected based on the high phoneme confusions previously found in [13]. The logatomes were extracted from words in purposely created sentences recorded by four native German speakers (2m, 2f, age range 25–36 years). The recordings were made in clean conditions with 48 kHz sampling frequency. The test stimuli are thus excerpts of natural speech, which can presumably be less carefully articulated than words or logatomes spoken in isolation as in the OLLO logatome speech database [23, 13], but on the other hand reflect a realistic pronunciation.

All original speech files (AusTalk and logatomes) were transmitted through simulated communication channels, creating different degraded versions of the same speech. Conditions with no bandwidth filter or codec applied, i.e. direct speech sampled as 8, 16, or 32 kHz were also considered. The ASR experiment and the intelligibility test examined a common set of 19 conditions, listed in Table 1, which are considered in this paper for building the prediction models for ASR.

The channel simulation process was conducted as follows. First, the speech was level-equalized to -26 dBov, a characteristic level of telephone channels, using the voltmeter algorithm of the ITU-T Rec. P.56. To simulate NB channels, the speech was downsampled to 8 kHz and then band-passed according to the ITU-T Rec. G.712 (300–3,400 Hz). For WB channels, the signals were downsampled to 16 kHz and band-filtered complying with ITU-T Rec. P.341 (50–7,000 Hz). For the SWB channels, the speech was downsampled to 32 kHz and then processed with the 14KBP filter (50–14,000 Hz), of the ITU-T Rec. G.191. Anti-aliasing low-pass FIR filters were always employed before downsampling. After band-filtering the signals, codecs were applied employing standard ITU and ETSI tools or the open-source Speex codec [24]. Finally, the speech was again level-equalized to -26 dBov.

3. Automatic speech recognition

The CMU Sphinx toolkit (Sphinx 4, [25]) was adopted for the ASR experiments. Separate context-dependent Hidden Markov Models (HMM) were trained and tested with speech data of each channel condition, that is, the distortion of the training files matched that of the testing files. The words of the first and second recording session from all speakers were pooled to build the training models, whereas the words uttered in the third recording session were retained for testing.

Our ASR system makes use of continuous HMMs with 39 base-phones without stress, three filler phones, and 9679 triphones. Five states per HMM are used, with each state modelled by eight Gaussians. The language model contains 324 uni-grams, 646 bi-grams and 966 tri-grams. 19 Mel-Frequency Cepstral Coefficients (MFCC) including the energy coefficient and the corresponding delta and delta-delta were extracted, constituting feature vectors of 57 components. 40 mel filters were employed for feature extraction, with the low and upper limits set as 300–3,400 Hz, 50–7,000 Hz, and 50–14,000 Hz, for NB, WB, and SWB distortions, respectively. Some of these parameters, i.e. the number of MFCCs and of filters in the filterbank, were defined by preliminary experimentation.

The recognition results were analyzed in terms of Word Error Rate (WER), calculated as the percentage of incorrectly recognized words in the test set. The WERs are shown for the different distortions of the training and test speech in Table 1 along with results of the intelligibility test, objective intelligibility estimations and POLQA MOS, obtained as described in next sections.

We are aware that the limited amount of training data may be the primary cause of the low ASR accuracy reached. On the other hand, the low, not saturated performance enabled us to detect effects between transmission conditions. Contrastingly, almost no variability was found within the ASR accuracy from WB codecs in [14]. The authors achieved around 98% accuracy in WB employing the TIMIT database for their experiments.

Table 1: *WERs, subjective intelligibility, predictions by the POLQA-intelligibility model, and POLQA MOS for NB (first block), WB (second block), and SWB (third block) degradations. The bitrates are indicated in kbit/s.*

Distortion	WER (%)	subj-intell	obj-intell	MOS
8kHz, nocodec	19.6	92.5	96.3	3.68
G.711@64	20.1	93.2	92.9	3.43
G.723.1@6.3	26.4	88.6	91.2	2.64
GSM-EFR@12.2	21.8	90.6	90.9	2.76
AMR-NB@4.75	27.3	85.9	85.1	2.03
AMR-NB@12.2	22.8	90.5	91.5	2.91
Speex-NB@2.15	41.8	70.5	82.0	1.86
Speex-NB@11	23.2	87.7	91.6	2.81
Speex-NB@24.6	20.3	91.8	92.2	3.05
16kHz, nocodec	15.3	96.8	99.8	4.43
G.722@64	16.6	95.5	97.7	4.14
AMR-WB@12.65	17.9	94.1	97.0	3.43
AMR-WB@23.05	16.7	94.8	97.5	3.78
Speex-WB@3.95	36.3	78.8	87.6	1.55
Speex-WB@23.8	23.4	94.0	95.2	3.79
Speex-WB@42.2	16.6	95.3	95.8	3.93
32kHz, nocodec	16.6	96.9	99.1	4.66
G.722.1C@24	18.9	92.7	97.2	3.46
G.722.1C@48	18.3	93.9	97.2	3.78

4. Intelligibility listening test

The complete set of stimuli presented in the intelligibility test consisted of 4 speakers x 8 VCV logatomes x 27 channel conditions = 864 segments. It has to be noted that some of these channel conditions were not examined in the ASR experiment. The test was performed by 30 listeners (15m, 15f), mean age 25 years (range 18–38 years) and with German as mother tongue. They were instructed to choose among the eight alternatives after listening to each stimulus. There was no possibility to listen to one stimulus more than once. Short breaks were included every 15 minutes approximately to avoid listeners' loss of focus. Before the test started, a brief familiarization phase was conducted in which the participants listened to samples of each logatome as many times as they wished. The complete test session had a duration of about one hour. It was performed in a quiet room using a laptop and Shure SRH240 headphones (diotic listening, frequency range 20–20,000 Hz). Listeners were not allowed to control the speech loudness level.

The test results were computed as the percentage of correct answers over all speakers and logatomes for each condition and are presented in Table 1, 3rd column. Further aspects about this test and its results can be read in [1].

5. Instrumental speech intelligibility and quality measurements

The intelligibility predictions were obtained by applying the POLQA-intelligibility model (V1490intellV2) to the logatome speech signals of our study, concatenated for each degradation separately. These objective scores attempt to predict the average accuracy of the intelligibility of test listeners and are shown in Table 1, 4th column. A second-order curve could be fitted between the objective and subjective intelligibility scores with $R^2 = 0.870$, $RMSE = 2.10$ (root mean square error).

Also using the degraded logatome speech, the POLQA standard V2.4.1 was employed to estimate objective quality in terms of MOS, shown in Table 1, 5th column. The speech recognition performance tends to improve with higher quality (WER decreases), contrasting with the findings in [14]. It can be observed that, overall, the ASR performance and the estimated MOS improve when transiting from NB to WB. Nevertheless, no clear advantage of SWB over WB transmission is manifested. This might be due to undesired noisy artefacts incorporated in this extended frequency band, from which no speech-specific features can be extracted. Interestingly, the coded-decoded SWB speech does not offer better quality or higher recognition rates than G.722, AMR-WB, and Speex WB operating at their higher bitrates. More SWB conditions should be tested in the future to investigate which codecs can provide both enhanced (or, at least, comparable) quality and ASR performance with respect to WB. The ASR performances and MOS in SWB are improved, however, compared to the NB conditions.

The POLQA model always operated in SWB mode, which permits the direct comparison between the signal qualities in the three bandwidths. The MOS were estimated on a joint scale in the range [1–5] for all distortions. A second-order curve was fitted between POLQA measures and subjective intelligibility scores with $R^2 = 0.858$, $RMSE = 2.20$. In [14], differently, NB-PESQ and WB-PESQ were applied for the NB and WB degradations, respectively. This impeded the comparison of all MOS values on a common scale.

The intelligibility test and the correlations of subjective results with objective intelligibility and with objective quality scores are described in more detail in [1].

6. Predicting WER from speech intelligibility and quality

The plausibility of predicting ASR WER from transmitted signal intelligibility (subjective and objective) and objective quality are investigated by fitting polynomial models.

6.1. Prediction from subjective intelligibility

A strong linear correspondence was found between subjective intelligibility and WER considering all distortions listed in Table 1. A linear fit yielded $R^2 = 0.951$, $RMSE = 1.49$ and is presented in Figure 1. The data points corresponding to the labeled distortions present higher residuals than the rest.

The linear fit found also indicates that it can be possible to predict subjective intelligibility scores from the ASR performance - a more realistic case given the costs of conducting listening tests. This was also addressed in [6] yet only considering NB codecs and bit errors conditions.

6.2. Prediction from POLQA-intelligibility

A quadratic curve was fitted to the pairs POLQA-intelligibility–WER (4th and 2nd columns of Table 1, respectively) with $R^2 = 0.834$, $RMSE = 2.75$, and shown in Figure 2.

Data points with high residuals correspond to the distortions AMR-NB@4.75 (high negative residual) and Speex-WB@3.95, Speex-WB@23.8, and Speex-NB@2.15 (high positive residual). An expert listening analysis showed that data points that are significantly above the optimal 2nd order regression sounded noise like distorted while data points significantly below the regression mostly did not show noise like speech degradations. Apparently the effect of noise like speech signal degradations is not modeled correctly by POLQA-intelligibility regarding their impact on the WER.

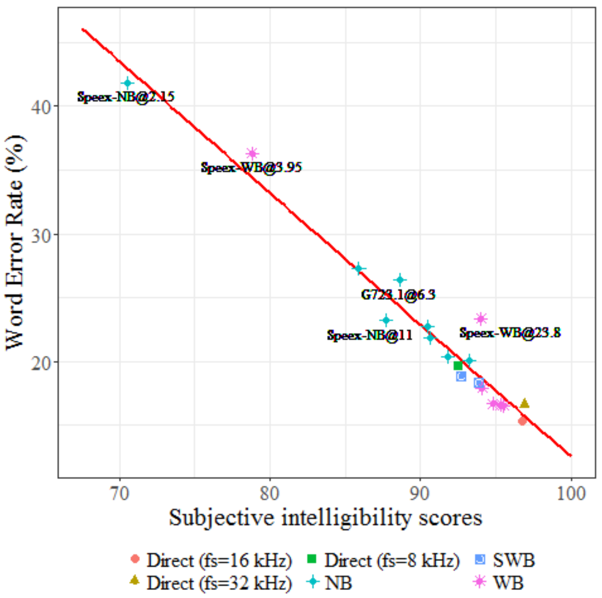

Figure 1: *Linear fit to predict WER using subjective intelligibility scores.* $l(x) = 115.7 - x$*,* $R^2 = 0.951$*,* $RMSE = 1.49$*.*

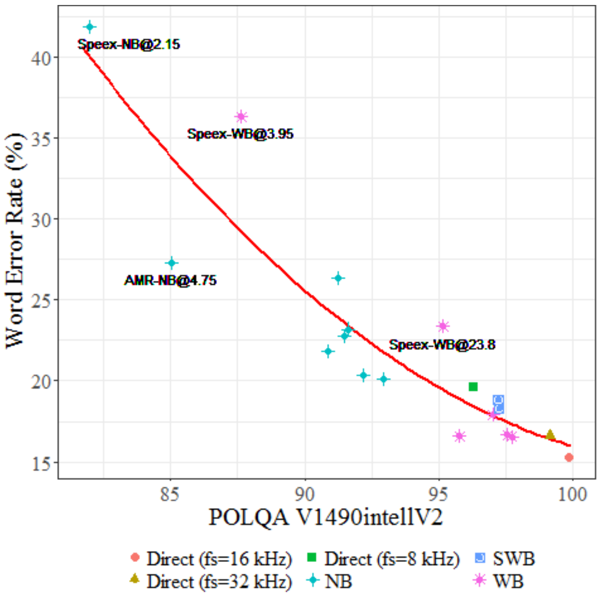

Figure 2: *Second-order polynomial fit to predict WER using POLQA-intelligibility values.* $q_P intell(x) = 22.1 - 26.3x + 5.1x^2$*,* $R^2 = 0.834$*,* $RMSE = 2.75$*.*

6.3. Prediction from POLQA

The ASR WER can also be predicted by objective quality estimations, that is, MOS derived by the POLQA model. The best fit with a second-order function yields $R^2 = 0.843$, $RMSE = 2.67$ (Figure 3). As also imposed for the POLQA-intelligibility–WER fit, the curve is monotonically decreasing.

An expert listening analysis showed that data points that are significantly above the optimal 2nd order regression (Speex-NB@2.15 and Speex-WB@23.8) sounded noise like distorted while the data point below the regression (AMR-NB@4.75) does not show this noise like speech degradations. The effect of noise like speech degradations seems to be different for WER compared to speech quality, for which POLQA is optimized.

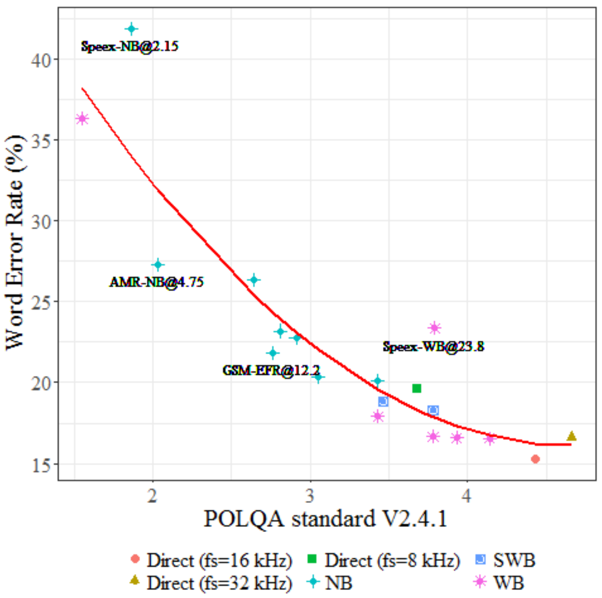

Figure 3: *Second-order polynomial fit to predict WER using POLQA MOS values.* $q_P MOS(x) = 65.7 - 21.3x + 2.3x^2$*,* $R^2 = 0.843$*,* $RMSE = 2.67$*.*

The fit with POLQA MOS is marginally better than that obtained when employing POLQA-intelligibility as predictor, while the fit that describes the relation between subjective intelligibility and POLQA-intelligibility is marginally better than that obtained with POLQA. Currently we have no explanation for this counter intuitive result.

The presented fits in Figures 2 and 3 can be advantageous when a channel codec has to be selected for communications and its merits need to be evaluated. If the objective scores are at hand, the costs of conducting ASR experiments can be saved in terms of time and resources.

7. Conclusions

This work has examined the relationships between ASR performance, speech quality, and intelligibility, considering distortions in three different bandwidths. It has been shown that ASR accuracy, intelligibility, and quality benefit from the transition from NB to WB, yet no meaningful improvement of SWB over WB could be observed.

Polynomial models have been fitted to data points corresponding to NB, WB, and SWB channels, which enable the prediction of the ASR accuracy from intelligibility subjective or objective scores ($R^2 = 0.951$ and $R^2 = 0.834$, respectively) or from MOS given by POLQA ($R^2 = 0.843$). Given the efforts and costs required to run ASR experiments, network planners can be assisted by these polynomial models involving instrumental measures to select among different configurations in the communication channel design process.

The low ASR performance reached in this study is not representative of today's state-of-the-art ASR systems trained with copious amounts of speech data. A future study involving an improved ASR system and SWB codecs is subject to the availability of clean speech databases of sufficient bandwidth. Also, a greater number of speakers and languages would be necessary to explore the generalization of our results and possibly to create more precise estimation models.

8. References

[1] L. Fernández Gallardo, “Speech Intelligibility Measurements over VoIP Channels,” in *Annual German Congress on Acoustics (DAGA)*, 2017.

[2] H. G. Hirsch, “The Influence of Speech Coding on Recognition Performance in Telecommunication Networks,” in *International Conference on Spoken Language Processing (ICSLP)*, 2002, pp. 1877–1880.

[3] H. Sun, L. Shue, and J. Chen, “Investigations into the Relationship Between Measurable Speech Quality and Speech Recognition Rate for Telephony Speech,” in *ICASSP*, vol. 1, 2004, pp. 865–868.

[4] T. Triyason and P. Kanthamanon, *Effect of Codec Bit Rate and Packet Loss on Thai Speech Recognition over IP*. Springer, 2013, pp. 232–241.

[5] S. Möller and H. Bourlard, “Analytic Assessment of Telephone Transmission Impact on ASR Performance Using a Simulation Model,” *Speech Communication*, vol. 38, no. 3–4, pp. 441–459, 2002.

[6] Y. Teng and R. F. Kubichek, “Speech Intelligibility Evaluation of Low Bit Rate Speech Codecs,” in *12th Digital Signal Processing Workshop - 4th Signal Processing Education Workshop*, 2006, pp. 251–256.

[7] J. Barker and M. Cooke, “Modelling Speaker Intelligibility in Noise,” *Speech Communication*, vol. 49, no. 5, pp. 402–417, 2007.

[8] 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.

[9] M. R. Schädler, A. Warzybok, S. Hochmuth, and B. Kollmeier, “Matrix Sentence Intelligibility Prediction using an Automatic Speech Recognition System,” *International Journal of Audiology*, vol. 54, no. Suppl. 2, pp. 100–107, 2015.

[10] J. Rodman, “The Effect of Bandwidth on Speech Intelligibility,” 2003, polycom, White Paper.

[11] S. Möller, A. Raake, N. Kitawaki, A. Takahashi, and M. Wältermann, “Impairment Factor Framework for Wideband Speech Codecs,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1969–1976, 2006.

[12] L. Fernández Gallardo, *Human and Automatic Speaker Recognition over Telecommunication Channels*, ser. T-Labs Series in Telecommunication Services. Singapore: Springer-Verlag, 2016.

[13] L. Fernández Gallardo and S. Möller, “Phoneme Intelligibility in Narrowband and in Wideband Channels,” in *Annual German Congress on Acoustics (DAGA)*, 2015, pp. 121–124.

[14] A. V. Ramana, L. Parayitam, and M. S. Pala, “Investigation of Automatic Speech Recognition Performance and Mean Opinion Scores for Different Standard Speech and Audio Codecs,” *IETE Journal of Research*, vol. 58, no. 2, pp. 121–129, 2012.

[15] M. Wältermann, I. Tucker, A. Raake, and S. Möller, “Extension of the E-Model Towards Super-Wideband Speech Transmission,” in *ICASSP*, 2010, pp. 4654–4657.

[16] ITU-T Recommendation P.863, *Perceptual Objective Listening Quality Assessment*, International Telecommunication Union, CH-Geneva, 2011.

[17] ITU-T Recommendation P.862.2, *Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs*, International Telecommunication Union, CH-Geneva, 2007.

[18] A. Hines, J. Skoglund, A. Kokaram, and N. Harte, “Robustness of Speech Quality Metrics to Background Noise and Network Degradations: Comparing ViSQOL, PESQ and POLQA,” in *ICASSP*, 2013, pp. 3697–3701.

[19] J. G. Beerends, R. A. van Buuren, J. van Vugt, and J. A. Verhave, “Objective Speech Intelligibility Measurement on the Basis of Natural Speech in Combination with Perceptual Modeling,” *Journal of the Audio Engineering Society*, vol. 57, no. 5, pp. 299–308, 2009.

[20] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullman, J. Pomy, and M. Keyhl, “Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part II Perceptual Model,” *Journal of the Audio Engineering Society*, vol. 61, no. 6, pp. 385–402, 2013.

[21] ——, “Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part I Temporal Alignment,” *Journal of the Audio Engineering Society*, vol. 61, no. 6, pp. 366–384, 2013.

[22] D. Burnham, D. Estival, S. Fazio, F. Cox, R. Dale, J. Viethen, S. Cassidy, J. Epps, R. Togneri, Y. Kinoshita, R. Göcke, J. Arciuli, M. Onslow, T. Lewis, A. Butcher, J. Hajek, and M. Wagner, “Building an Audio-Visual Corpus of Australian English: Large Corpus Collection with an Economical Portable and Replicable Black Box,” in *Interspeech*, 2011, pp. 841–844.

[23] B. T. Meyer, T. Jürgens, T. Wesker, T. Brand, and B. Kollmeier, “Human Phoneme Recognition Depending on Speech-Intrinsic Variability,” *The Journal of the Acoustical Society of America*, vol. 128, no. 5, pp. 3126–3141, 2010.

[24] J.-M. Valin, *The Speex Codec Manual Version 1.2 Beta 3*, Xiph.org Foundation, 2007.

[25] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, *Sphinx-4: A Flexible Open Source Framework for Speech Recognition*, Sun Microsystems, Inc., 2004.