ASR-Based, Single-Ended Modeling of Listening Effort ? A Tool for TV Sound Engineers

Rainer Huber, Hannah Baumgartner, Stefan Goetze, Jan Rennies

ASR-Based, Single-Ended Modeling of Listening Effort – A Tool for TV Sound Engineers

Rainer Huber[1] **Hannah Baumgartner**[1] **Stefan Goetze**[2] **Jan Rennies-Hochmuth**[1]
[1] Fraunhofer Institute for Digital Media Technology IDMT, Oldenburg, Germany
[2] Department of Computer Science, The University of Sheffield, Sheffield, UK
Rainer.Huber@idmt.fraunhofer.de

ABSTRACT

This paper reviews our research approaches towards a listening effort model and its applications as a tool to automatically measure and display the perceived listening effort required to understand speech in a variety of different background sounds. It is single-ended, i.e. it does not require a clean speech reference, and is based on an automatic speech recognition (ASR) system. Speech distortions and interfering background sounds increase the uncertainty of the ASR system, which can be quantified and mapped to a perceptually interpretable scale using a psychoacoustic modeling approach. This performance measure correlates well with mean subjective listening effort ratings for a variety of distortions and acoustic backgrounds typical for TV broadcast material ($r > 0.9$). In principle, the tool is applicable to be integrated as a software plugin for digital audio workstations (DAWs) to support the work of sound engineers, or in other applications such as speech quality monitoring of communication channels or real-time control of signal-enhancement algorithms.

1. INTRODUCTION

The perceived speech intelligibility of many TV productions is a subject of complaints by TV viewers. In the production process of broadcast material, sound engineers mixing speech with background sounds (e.g. music or "atmosphere") typically listen to the sound material many times under ideal acoustical conditions with high-quality equipment, which is different from the listening situation of TV viewers at home. Hence, the sound engineer might underestimate the listener's effort required to understand the speech. This contribution reviews our research approaches towards better understanding and eventually counteracting these problems, including a series of psychoacoustic studies.

We focus on listening effort rather than speech intelligibility, because in contrast to speech intelligibility (typically measured as word recognition rate), listening effort can still be affected by changes in noise levels at realistic SNRs, where speech intelligibility is already close to 100% (e.g., [1-3]; see also Figure 1). Consequently, Rennies et al. [2] conclude that "intelligibility is an insensitive measure to evaluate many everyday listening conditions".

The measurement of listening effort by means of formal subjective listening tests is time consuming and costly and cannot be used for, e.g., online-monitoring. Hence, instrumental methods to predict the perceived listening effort would be valuable tools for the automatic evaluation of, e.g, TV broadcast material or speech enhancement schemes.

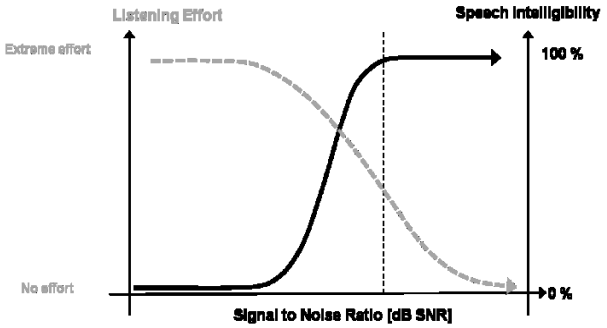

Figure 1. Schematic illustration of the relation between listening effort (grey) and speech intelligibility (black) as functions of the SNR (taken from [3]).

Signal-based instrumental methods for the prediction of, e.g., speech quality, speech intelligibility or listening effort can be classified into single-ended (or "reference-free", "non-intrusive") and double-ended (or "reference-based", "intrusive") methods. Double-ended methods (such as, e.g. [4, 5] for speech/audio quality assessment) typically achieve more accurate predictions than single-ended methods (such as, e.g. [6, 7]), but have the disadvantage that they need a clean or nearly clean reference signal, which is often not available. Consequently, we focus on single-ended methods for the prediction of perceived listening effort in the following.

Huber et al. [8] introduced a single-ended approach for listening effort prediction from acoustic parameters (LEAP) based on an ASR system. The ASR system employs a deep neural network (DNN) to compute phoneme posterior probabilities (or "posteriorgrams") of the input speech. Distortions or additive noise increase the uncertainty of the ASR system, which is reflected by smeared posteriorgrams (see Figure 2). The degree of posteriorgram degradation is quantified by a performance measure, i.e., the *mean temporal distance* or *M*-Measure proposed in [9]. It has been found that the *M*-Measure correlates well with measured listening effort data of several data sets [8, 10]. In related work, this modeling approach was also explored in the context of speech quality prediction [11]. One limitation of the original method presented in [8] was that the ASR system was trained on the specific background noises of the test data

sets. This issue was addressed with a modified LEAP model [10]. In that study, we investigated a deep time-delay neural networks (TDNNs) trained with a much bigger speech corpus with speaker-independent training and mismatched noises (see Section 2.4 for details).

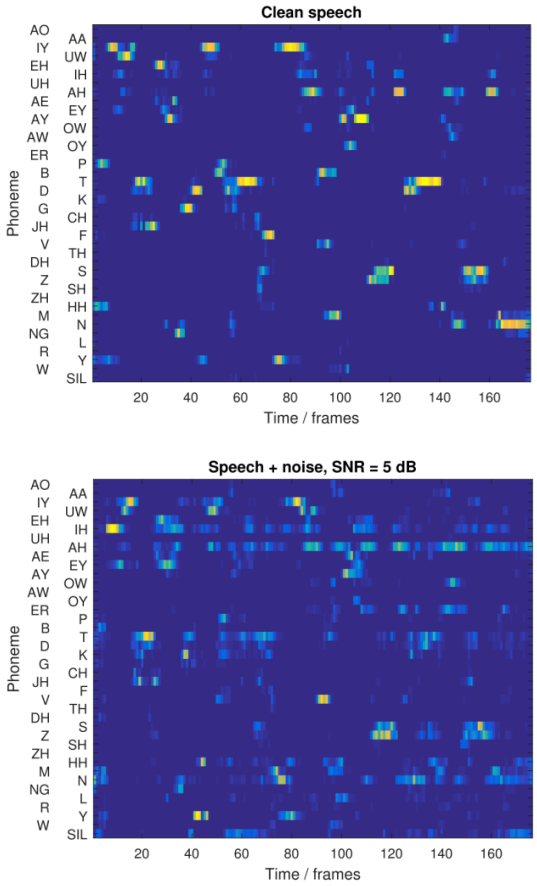

Figure 2: Posteriorgrams of clean speech (top) and the same speech utterance with additional noise at an SNR of 5 dB (bottom).

2. METHODS

2.1 Listening Effort Dataset I

Separate sound tracks (language/background) from different German TV productions were available. From this material, 234 audio clips with an average length of about 10s were cut out and mixed with strongly varying (partly unrealistic) mixing ratios to cover a range of the expected listening effort as large as possible.

20 normal-hearing German subjects aged 20-30 years (median = 25 years), 10 of them male, 10 female, participated in the study. They rated the perceived listening effort of the 234 audio clips on the scale shown in Figure 3 on a touch screen. The rating task for the subjects was: ”Wie anstrengend ist es für Sie, die Sprache zu verstehen?“ ("How much effort do you have to spend to understand the speech?”). The acoustic presentation was performed via headphones (Sennheiser HD 650) in a listening booth, no video was presented. The selected listening effort categories were mapped to a numerical value from 1 ("no effort") to 13 ("extreme effort", see Figure 3).

Figure 3. 13-step listening effort rating scale after [12].

2.2 Listening Effort Dataset II

A second dataset was created from disjunct further German TV shows, again comprising separate sound tracks for language and background. From this material, 210 audio clips with an average length of about 10s were cut out and mixed with strongly varying mixing ratios.

20 normal-hearing, German subjects aged 20-29 years (median = 23 years), 9 of them male, 11 female, participated in the study. The rating procedure was the same as for Dataset I (see above).

2.3 Listening Effort Dataset III

39 audio excerpts of about 10 s each were taken from English and American movies were used to create dataset III; 19 containing clean speech, 20 containing background sounds without speech. From these 39 excerpts, 140 audio clips were mixed with various SNRs in order to cover a broad range of expected listening efforts. Moreover, six sentences from an American English speech intelligibility test (matrix test, [13]) mixed with speech-shaped noise were added, so that the measured listening efforts could be compared to the results from an earlier study [14], which contained the same six stimuli.

Fifteen normal-hearing, native English speaking subjects aged 22-44 years (median = 27 years), 6 male, 9 female, participated in the study. They rated the perceived listening effort of the 146 stimuli using a slightly adapted version of the listening effort rating scale described above on a touch screen. The adaptation consisted in the addition of a 14th rating step with the verbal description: “can't understand the speech at all”.

Apart from that, the rating procedure was the same as described above for Datasets I and II.

2.4 Posteriorgram generation

An acoustic model for automatic speech recognition (ASR) was trained prior to the extraction of context-dependent triphone posteriorgrams. For ASR training, approx. 1.000 hours of unprocessed German speech data of an in-house training data set were used, augmented to about 8.000 hours in a multi-condition training setup. A deep time-delay neural network (TDNN) [15, 16], which is also known as a one-dimensional temporal convolutional neural network, was trained with the lattice-free maximum mutual information (LF-MMI) criterion [17]. To save computational time, the LF-MMI trained neural network modeled output posterior probabilities at one third of the frame rate of conventional acoustic ASR models, which usually run at a frame rate of 100 Hz. The TDNN topology was similar to a setup described in [18] that had a total context size of +/- 15 input feature frames (resulting in a temporal context of 310 ms), which were analyzed in 7 hidden layers of dimension 700 with rectified linear units. The dimensionality of the final output layer amounted to 6448, which was the result of decision tree clustering of context-dependent Hidden Markov Model output distributions. As acoustic features input to the TDNN, we used 40-dimensional log-Mel filterbank energies. Note that during training, the TDNN used here had two output layers, one that followed the LF-MMI objective function and one that followed a cross-entropy (CE) objective function. The latter one is usually used to regularize training only, while the former one is used for ASR purposes. In this work we used the CE output layer for generating posteriorgrams instead, due to better results.

2.5 Performance measure

From the posteriorgrams, the *M*-Measure as proposed by Hermansky et al. [9] was computed. The *M*-Measure computes the average difference between two vectors of phoneme posteriors $p_{t-\Delta t}$ and p_t (i.e., two columns of the posteriorgram) with a temporal distance Δt:

$$M(\Delta t) = \frac{1}{T-\Delta t}\sum_{t=\Delta t}^{T} D(p_{t-\Delta t}, p_t), \qquad (1)$$

with T being the temporal length of the analyzed posteriorgram (which is equal to the length of the analyzed sound file, i.e. around 10 s in the present study), and D being the symmetric Kullback-Leibler divergence between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ with components $x(i)$ and $y(i)$:

$$D(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{N} x(i)\log\left(\frac{x(i)}{y(i)}\right) + \sum_{i=1}^{N} y(i)\log\left(\frac{y(i)}{x(i)}\right) \qquad (2)$$

In the present study, N equals the dimensionality of the TDNN output layer (6448) and M was computed for Δt =350 to 800 ms (in 50 ms steps) and averaged to yield the final listening effort predictor $\overline{M}$.

2.6 Speech activity detection

The posteriorgram calculation was only performed for sections of the audio signal in which speech activity was detected. To this end, an automatic speech activity detection (SAD) was employed, which was also based on a deep neural network that had also been trained on TV audio signals (amongst other signals) [19]. Mel-frequency cepstral coefficients (MFCCs) were used as feature vectors at the input of the SAD's neural network. (For details on automatic speech activity detection see [19].)

3. RESULTS

3.1 Datasets I+II (German material)

The relation between mean subjective listening effort ratings (shown as *listening effort mean opinion scores*, LE-MOS, on the ordinate) and the model's listening effort predictor $\overline{M}$ is shown by a scatter plot for the combined datasets I (blue circles) and II (red stars) in Figure 4. The scatter plot shows that LE-MOS and $\overline{M}$ are linearly related. A linear regression analysis yields the following mapping function:

$$\text{LE-MOS} = -0.4 \cdot \bar{M} + 14 \qquad (3)$$

Pearson's correlation coefficient is $r = 0.92$ and Spearman's rank correlation coefficient amounts to $rs = 0.92$.

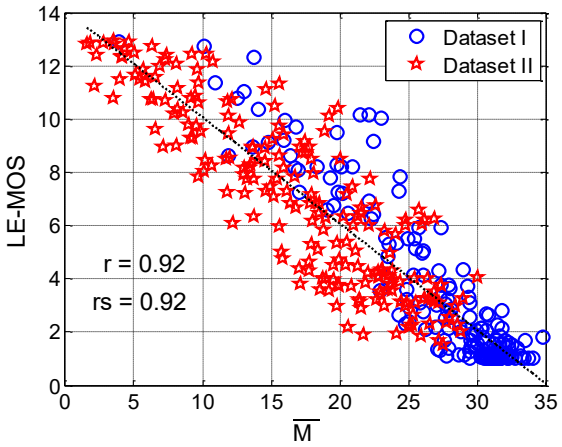

Figure 4: Relation between averaged subjective listening effort ratings LE-MOS (ordinate) and corresponding values of the listening effort predictor $\bar{M}$ (abscissa) of dataset I (blue circles) and dataset II (red stars). The black dotted line represents the best fit after linear regression. r and rs indicate Pearson's correlation coefficient and Spearman's rank correlation coefficient, respectively.

3.2 Dataset III (English/American material)

Figure 5 shows the model prediction results for the English/American movie material. Again, a linear relation between subjectively measured and predicted listening effort data is obtained. The correlation is somewhat lower than for the joint German datasets

($r = 0.87$, $rs = 0.85$), and regression analysis yields a slightly steeper mapping function:

$$\text{LE-MOS} = -0.49 \cdot \bar{M} + 13.4 \qquad (4)$$

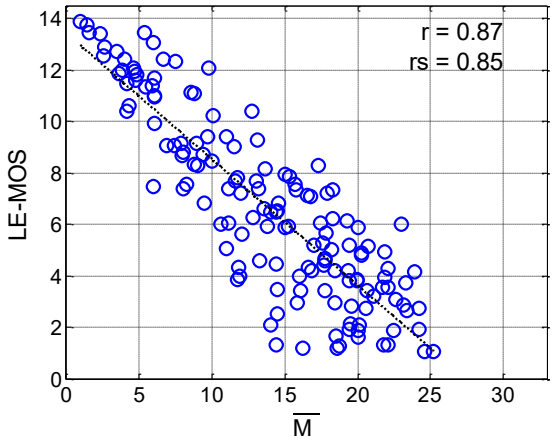

Figure 5: As Figure 4, but for Dataset III.

4. DISCUSSION

The application of the reference-free instrumental method for the prediction of listening effort as introduced in [8] and evaluated for speech enhancement schemes in [10] has also been shown to be very promising for the use in broadcast applications, especially since the results of listening effort prediction were achieved without any training or optimization of the method on the data used here. Further improvements of the results are conceivable if the speech recognizer would also be trained with typical background "noise" signals as they occur in broadcast material.

The results obtained for Dataset III show that the single-ended method for listening effort prediction also works reasonably well for English audio material - although the used ASR system was trained with German speech data - if a different linear function for mapping the metric $\bar{M}$ to the subjective listening effort scale is used. The necessity for using a steeper mapping function is a consequence of the fact that the range of $\bar{M}$ values is smaller for English speech than it is for German speech. The maximum $\bar{M}$ values for speech requiring minimum listening effort (i.e., clean speech) are markedly smaller in case of English speech compared to German speech (25 vs. 35, cf. Figures 4 and 5). This might be explained by less clear posteriorgrams for clean English speech than for clean German speech, meaning less distinct, less high phoneme probabilities, which is a consequence of the mismatch between German ASR training data and English test data. The sets of German and English phonemes have considerable overlap and similarity, but they are not identical. Apart from a steeper relation between $\bar{M}$ and LE-MOS, another effect of the higher uncertainty of the ASR system in case of clean English speech compared to German speech can be observed in the larger variance of $\bar{M}$ values for low LE-MOS values (see Figure 5). For some audio clips with actual LE-MOS values near 1, the corresponding predicted LE-MOS values (after mapping $\bar{M}$ to LE-MOS following Eq. (4)) are above 6. Such erroneous high predicted LE-MOS values might be caused by the mismatch between German and English phonemes. Despite these effects, the overall prediction accuracy as indicated by correlation values is comparable to German audio material, indicating that the proposed approach can also be applied to English broadcast, e.g., automatic listening effort monitoring of movies or TV material, provided that the language mode (German/English) is adapted accordingly.

Although the present results indicate that the proposed approach can be extended to English without adapting the underlying ASR engine to the target language, care should be taken when considering a further generalization to other languages. It is possible for the observed differences to become larger if languages with more dissimilar phoneme sets like, e.g., Chinese language are considered. In such cases, the underlying ASR system of the method might have to be re-trained with the target language.

The model is implemented as a real-time library, which can be easily integrated into, e.g., DAWs or communication systems to monitor listening effort and, potentially, to control automatic signal adjustment strategies. First evaluations with professionals not reported here indicate the model's potential and applicability, thus providing a successful example of transition from basic psychoacoustic studies employing simple and highly controlled stimuli to complex conditions with considerable potential for applications.

5. CONCLUSIONS

The single-ended, ASR-based method for the prediction of listening effort reviewed here has been shown to correlate well with mean subjective listening effort ratings obtained in the evaluation of speech enhancement schemes and TV and movie audio. The method runs online and in real time and is thus applicable to be integrated as a software plugin for DAWs to support the work of sound engineers, or in other applications such as speech quality monitoring of communication channels or real-time control of signal-enhancement algorithms.

6. ACKNOWLEDGMENT

This study was supported by the German Ministry of Education Research (BMBF), project SITA; FKZ: 01/S17017

7. REFERENCES

[1] M. Krueger, M. Schulte, M. Zokoll, K. Wagener, M. Meis, T. Brand, and I. Holube, "Relation between listening effort and speech intelligibility in noise," *Am. J. Audiol.,* vol. 26, pp. 378-392, Oct. 2017.

[2] J. Rennies, H. Schepker, I. Holube, and B. Kollmeier, “Listening effort and speech intelligibility in listening situations affected by noise and reverberation,” *J. Acoust. Soc. Am.*, vol. 136, no. 5, pp. 2642-2653, Nov. 2014.

[3] Klink, K., Schulte, M., Meis, M.: Measuring listening effort in the field of Audiology – a literature review of methods, part 1. *Z Audiol* 51 (2) 60–67, 2012.

[4] ITU-T. Methods for Objective and Subjective Assessment of Speech Quality (POLQA): Perceptual Objective Listening Quality Assessment. Recommendation P.863, International Telecommunication Union, Geneva, Switzerland. 2014

[5] R. Huber and B. Kollmeier, “PEMO-Q - a new method for objective audio quality assessment using a model of auditory perception,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 6, pp. 1902-1911, Nov. 2006.

[6] ITU-T. Single-ended Method for Objective Speech Quality Assessment in Narrow-band Telephony Applications. Recommendation P.563, International Telecommunication Union, Geneva, Switzerland. , 2004.

[7] D. S. Kim and A. Tarraf, “ANIQUE+: a new American national standard for nonintrusive estimation of narrowband speech quality,” *Bell Labs Tech. J.* vol. 12, no. 1, pp. 221-236, May 2007.

[8] R. Huber, M. Krüger, and B. T. Meyer, “Single-ended prediction of listening effort using deep neural networks”, *Hear. Res.*, vol. 359. pp. 40-49, Mar. 2018

[9] H. Hermansky, E. Variani, and V. Peddinti, “Mean temporal distance: predicting ASR error from temporal properties of speech signal,” in *Proc. IEEE Conf. Acoust. Speech, Signal Process. (ICASSP)*, Vancouver, Canada, May. 2013, pp. 7423-7426.

[10] R. Huber, A. Pusch, N. Moritz, J. Rennies, H. Schepker, and B. T. Meyer, B.T.: “Objective Assessment of a Speech Enhancement Scheme with an Automatic Speech Recognition-Based System.” *Proceedings ITG Conference on Speech Communication* (2018), 86-90

[11] R. Huber, J. Ooster, and B. T. Meyer, “Single-ended Speech Quality Prediction Based on Automatic Speech Recognition”, *J. Aud. Eng. Soc.*, vol. 66, no. 10. pp. 759-769, 2018.

[12] M. Schulte, M. Meis, K. Wagener: “Listening Effort and Speech Intelligibility,” *Proceedings 8th EFAS Congress / 10th Congress of the German Society of Audiology*, 2007.

[13] B. M. Kreisman, R. Carroll, M. Zokoll, A. Warzybok, P. Allen, P. Folkeard, K. C. Wagener, and B. Kollmeier: “Design, Optimization, and Evaluation of an American English Matrix Sentence Test in Noise,” *Talk presented at AudiologyNow! – 25th Annual Meeting of the American Association of Audiology*. Anaheim, CA (USA), April 03-06, 2013.

[14] J. Rennies, V. Best, E. Roverud, and G. Kidd Jr.: “Energetic and informational components of speech-on-speech masking in binaural speech intelligibility and perceived listening effort,” *Trends in Hearing* 23, pp. 1-21., 2019

[15] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang: "Phoneme recognition using time-delay neural networks," *IEEE Transaction on Acoustics, Speech, and Language Processing*, vol. 37, no. 3, pp. 328-339, 1989.

[16] V. Peddinti, D. Povey, and S. Khudanpur: "A time delay neural network architecture for efficient modeling of a long temporal contexts," *Proc. Interspeech*, Dresden, pp. 2440-2444, 2015.

[17] D. Povey, V. Peddinti, D. Galvez, P. Ghahrmani, V. Manohar, X. Na, Y. Wang, S. Khudanpur: "Purely sequence-trained neural networks for ASR based on lattice-free MMI," *Proc. Interspeech*, San Francisco, 2016.

[18] V. Peddinti, Y. Wang, D. Povey, and S. Khudanpur, "Low latency acoustic modeling using temporal convolution and LSTMs," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 373-377, 2018.

[19] N. Moritz, J. Drefs, H. Baumgartner, J. Rennies: „Sprachaktivitätserkennung basierend auf Deep Neural Networks für Anwendung in Film und Fernsehen,“ *Fortschritte der Akustik*. DAGA 2016, S.960-963; DEGA, Berlin, 2016