# Single-ended Prediction of Listening Effort for English Speech

Rainer Huber, Hannah Baumgartner, Vinay Nooromkuttil Krishnan, Stefan Goetze, Jan Rennies
*Fraunhofer IDMT, Hearing-, Speech- and Audio Technology, 26129 Oldenburg*
*E-Mail: rainer.huber@idmt.fraunhofer.de*

## Introduction

The monitoring of speech intelligibility and listening effort of TV audio material is necessary for quality assurance of TV productions. Ensuring (almost) 100% speech intelligibility is not sufficient, since even with such high speech intelligibility, the listening effort required for comprehension can be unacceptably high in the long run. Listening effort is therefore a more sensitive and relevant parameter for quality assurance of broadcast material than speech intelligibility. Its evaluation by individual persons is very subjective, i.e., it can be individually very different so that formal listening tests with many test listeners are required to obtain reliable and reproducible results. In practice, however, such tests are usually too costly and time-consuming. Technical methods for objective evaluation or prediction of the perceived listening effort would therefore be a valuable tool for quality assurance of TV productions, but also for other applications such as the evaluation of signal enhancement algorithms, e.g. in hands-free telecommunication devices.

Basically, reference-based ("double-ended") and reference-free ("single-ended") methods are suitable for this purpose. Reference-based methods compare the test signal to be evaluated (e.g. speech with background noise) with the undisturbed speech signal as reference. This approach is used, for example, in instrumental speech quality assessment by the ITU-T standard method POLQA [1]. A similar approach is based on the comparison of useful and interfering signal (spectra), as applied, e.g., in the Speech Intelligibility Index (SII) [2]. A disadvantage of these approaches is that the clean, undisturbed speech signal must be available or the target and interfering signal must be available separately, which is not always the case (e.g., in TV mixes received by the end user). Even if, as in the mixing process in TV production, the speech and background sound tracks are separate, it is possible that the speech track may already contain disturbances such as soft background noise, reverberation, poor articulation and/or distortion that affect the listening effort. Such disturbances of the reference signal would not be detected by reference-based methods, because the reference signal, as it is, defines the optimal quality or minimum listening effort. Therefore, we proposed a reference-free method for the prediction of listening effort of German TV broadcast audio material in [3]. The method itself was presented for the first time in [4]. The variant presented in [3] was trained and evaluated on German speech only. This paper presents a study for which the proposed method was evaluated with English audio material taken from English and US-American movies and compared to experimental data collected with native listeners. As will be described in the following, a slight modification of the method was necessary in order to become capable of accurately predicting the perceived listening effort of English speech.

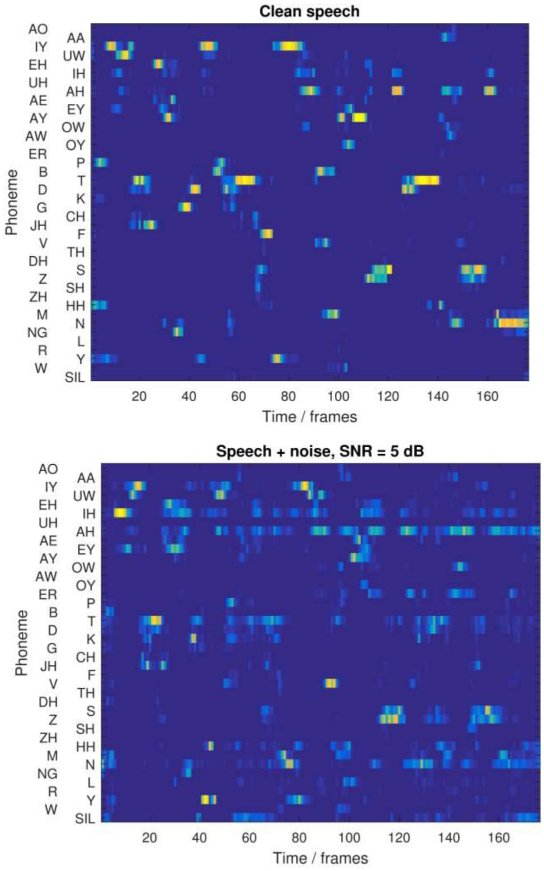


**Figure 1:** Posteriorgrams of clean speech (top) and the same speech utterance with additional noise (SNR = 5 dB, bottom).

## Method

The basic approach of the method is to use a part of an automatic speech recognition (ASR) system, i.e. the processing up to the deep neural network (DNN) which decides for the phoneme classes. Disturbances of speech such as distortions or background noise lead to an increased recognition uncertainty of the ASR system, similar to human speech perception. This uncertainty can be detected and quantified in the recognition system as follows: The deep neural network produces phoneme posterior probabilities ("posteriorgrams"). A posteriogram represents the temporal course of the probability for the activity of individual phonemes (see Fig. 1). Disturbances of speech lead to the posteriorgrams being "smeared" (see Fig. 1, lower panel). The degree of smearing is quantified by a mathematical measure. This measure is used as a predictor of listening effort. The generation of the posteriorgrams and the measure for quantifying the degree of smearing are described below.

### Posteriorgram generation

The same ASR system as in [5] was used to generate the posteriorgrams, which will therefore only be briefly described here (for details see [5]):

Short-term energies of a 40-channel Mel filter bank are used as acoustic features. Splicing is applied, i.e. in each filter bank channel, -15...+15 blocks with time shift of 10 ms are used (= 310 ms temporal context window) are combined and passed to a deep time-delay neural network (TDNN) as a feature vector. This deep TDNN has seven hidden layers with 700 rectified linear units each. The output layer consists of 6448 neurons, i.e. one per triphone (a triphone is a sequence of three phonemes). The network was trained approx. 1000 hours of clean speech. This database was extended to about 8000 hours by augmentation, i.e. mixing the clean speech segments with different types of noise at different SNR.

### Posteriorgram measure

From the output of the deep TDNN, i.e. the posteriorgram, the "mean temporal distance" or "*M*-Measure" was calculated according to Hermansky et al. [6]. The *M*-Measure calculates the average mathematical distance between two vectors of phoneme posteriors $p_{t\text{-}\Delta t}$ and $p_t$ (i.e. two columns of the posteriorgram) with a temporal distance $\Delta t$:

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

$T$ is the time length of the analyzed posteriorgram (which is equal to the length of the analyzed audio file, i.e. about 10 s in this study). $D$ is the symmetrical Kullback-Leibler divergence between two vectors $x$ and $y$ with the components $x(i)$ and $y(i)$:

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

In this study, $N$ was equal to the dimensionality of the TDNN output layer (6448) and $M$ was calculated for $\Delta t$ = 350 to 800 ms (in 50 ms steps) and then averaged to give the final listening effort predictor $\overline{M}$. This predictor was then mapped onto the same response scale as used in the subjective listening tests to allow for a quantitative comparison.

## 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 with (amongst others) TV audio signals [7]. Mel-Frequency Cepstral Coefficients (MFCCs) are used as feature vectors at the input of the SAD's neural network. (For details on automatic speech activity detection see [7].)

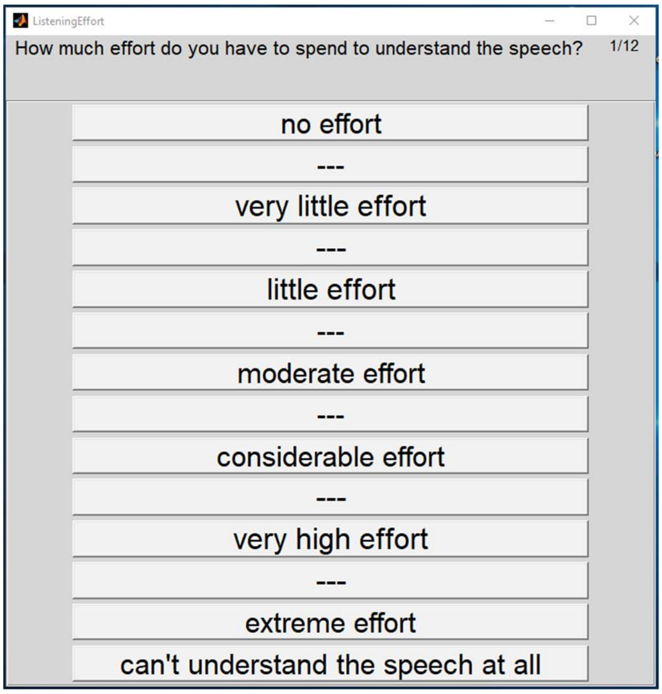
## Listening effort data

### Stimuli

39 audio excerpts of about 10 s each were taken from English and American movies; 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, [8]) mixed with speech-simulating noise were added, so that the measured listening efforts could be compared to the results from an earlier study [9], which contained the same six stimuli.

### Subjects and rating procedure

Fifteen normal hearing subjects aged 22-44 years (median = 27 years), six male, nine female, participated in the study. They rated the perceived listening effort of the 146 stimuli using a 14-step rating scale with eight named and six unnamed categories (Fig. 2) on a touch screen. The subjects were asked: "How much effort do you have to spend to understand the speech?" The stimuli were presented via headphones (Sennheiser HD 650) in a sound attenuating booth. Video was not provided. The selected rating categories were mapped to numerical values from 1 (corresponding to the rating category "effortless") to 14 **(**"can't understand the speech at all").



How much effort do you have to spend to understand the speech?

**Figure 2**: Listening effort rating scale (adapted from [10])

## Results

Subjective listening effort ratings, averaged across subjects (LE-MOS - „Listening Effort Mean Opinion Scores") are plotted vs. the corresponding values of the listening effort predictor $\overline{M}$ in a scatter plot shown in Fig. 3. A linear relation between LE-MOS and $\overline{M}$ values could be observed. The linear (Pearson) correlation coefficient is $r = 0.87$.

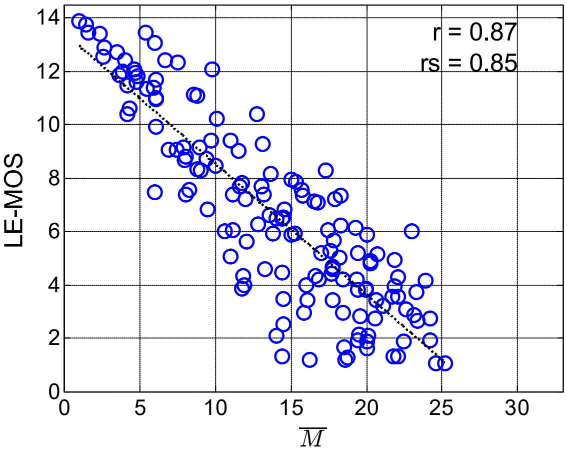


**Figure 3**: Relation between averaged subjective listening effort ratings LE-MOS for English audio material (ordinate) and corresponding values of the listening effort predictor $\overline{M}$ (abscissa). The black dotted line represents the best fit after linear regression. r and rs indicate Pearson's correlation coefficient and Spearman's rank correlation coefficient, respectively.

### Mapping $\overline{M}$ to LE-MOS for German and English speech

In order to actually predict LE-MOS values by the proposed method, the objective measure $\overline{M}$ had to be mapped to the LE-MOS scale. A mapping function *f*: $\overline{M}$→ LE-MOS was derived empirically by means of linear regression. For German speech, the listening effort data presented in [3] were used. Fig. 4 shows the relation between LE-MOS and $\overline{M}$ for this German dataset consisting of more than 400 data points. In comparison to the results for the English dataset shown in Fig. 3, the $\overline{M}$ values spanned a larger range towards higher values for German speech, up to about $\overline{M}$=35, whereas for English speech, the maximum value of $\overline{M}$ was about 25. Consequently, the relation between LE-MOS and $\overline{M}$ values was steeper for English speech than for German speech.

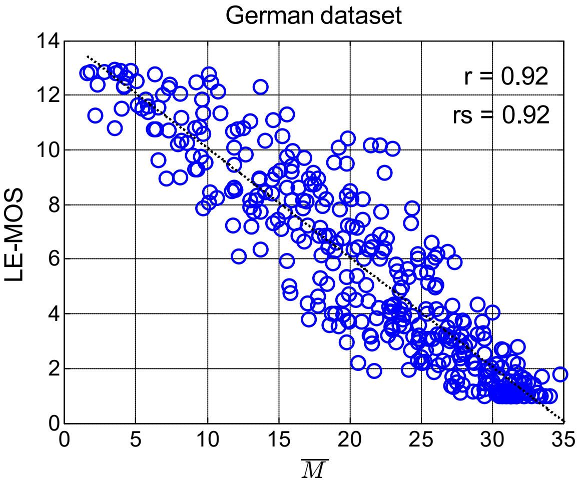


**Figure 4:** As Fig. 3, but for the German dataset presented in [3]

Applying the new, steeper mapping function to the $\overline{M}$ values of the English dataset, the relation between transformed $\overline{M}$ values and LE-MOS shown in Fig. 5 is obtained.

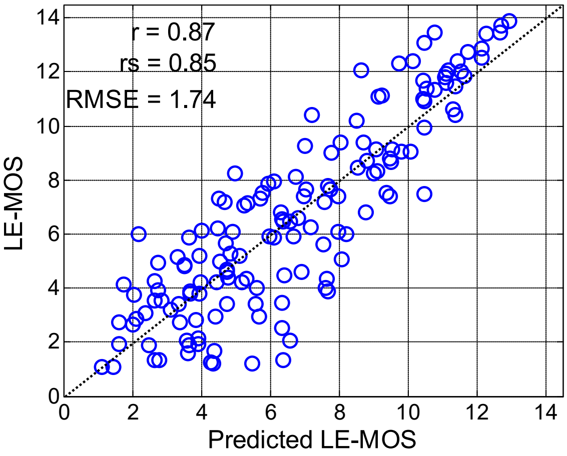


**Figure 5:** LE-MOS prediction results for the English dataset. The LE-MOS predictions are obtained by applying a new linear transformation to $\overline{M}$. RMSE: Root Mean Squared Error between actual and predicted LE-MOS. Black, dotted line: Perfect prediction, i.e. predicted LE-MOS = LE-MOS.

## Discussion

The results 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 $\overline{M}$ 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 $\overline{M}$ values is smaller for English speech than it is for German speech. The maximum $\overline{M}$ 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. Figs. 3 and 4). 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 $\overline{M}$ 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 predicted LE-MOS values for low actual LE-MOS values (see Fig. 5). For some audio clips with actual LE-MOS values near 1, the corresponding predicted values 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 and RMSE is comparable to German audio material, indicating that the proposed approach can be applied to English broadcast applications, e.g., automatic listening effort monitoring of movies or TV material, provided that the language mode

(German/English) is adapted accordingly. At present, this has to be set manually, but in future versions this might be done by means of an automatic language recognizer.

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.

## Conclusions

The single-ended method for assessing the listening effort of German TV broadcast audio material presented in [3] also works for English movie audio, if a different (steeper) mapping function $f$: $\overline{M} \rightarrow$ LE-MOS is used. The computation of $\overline{M}$ is not changed. For the application of the method to other languages, further, individual mapping functions have to be derived from listening tests. For languages with quite different sets of phonemes (like Chinese), however, this approach might not work anymore. Instead, the underlying ASR part of the method might have to be retrained with the target language.

## Acknowledgment

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

## References

[1] ITU-T Rec. P.863. Perceptual Objective Listening Quality Assessment. (2018) Geneva, Switzerland

[2] ANSI S3.5–1997. American National Standard Methods for the Calculation of the Speech Intelligibility Index. New York: ANSI, 1997

[3] Huber, R., Baumgartner, H., Rollwage, C., Goetze, S., Rennies, J.: Erfassung der Höranstrengung fertiger TV-Mischungen. In Fortschritte der Akustik. DAGA 2019, S. 919-922; DEGA, Berlin

[4] Huber, R., Spille, C., Meyer, B.T.: Single-Ended Prediction of Listening Effort Based on Automatic Speech Recognition. In Proceedings Interspeech 2017, 1168-1172

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

[6] Hermansky, H., Variani, E., Peddinti, V.: Mean temporal distance: Predicting ASR error from temporal properties of speech signal. In Proceedings ICASSP 2013, 38th IEEE Int. Conf. Acoust. Speech Signal Process. doi: 10.1109/ICASSP.2013.6639105

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

[8] Kreisman, B. M., Carroll, R., Zokoll, M. A., Warzybok, A., Allen, P., Folkeard, P., Wagener, K. C., & Kollmeier, B.: 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.

[9] Rennies, J., Best, V., Roverud, E., & Kidd Jr., G.: Energetic and informational components of speech-on-speech masking in binaural speech intelligibility and perceived listening effort. Trends in Hearing (2019) 23, 1-p21.

[10] Krueger, M., Schulte, M., Brand, T., & Holube, I.: Development of an adaptive scaling method for subjective listening effort. The Journal of the Acoustical Society of America, 2017, 141(6), 4680-4693.