ASR-based Measures for Microscopic Speech Intelligibility Prediction

Mahdie Karbasi[*], *Dorothea Kolossa*[*+]

[*]Cognitive Signal Processing Group, Ruhr-Universität Bochum, 44801 Bochum, Germany
[+]Kavli Institute for Theoretical Physics, UC Santa Barbara, USA
{mahdie.karbasi, dorothea.kolossa}@rub.de

Abstract

Automatic and accurate prediction of human speech perception performance is of great benefit for developing speech processing algorithms. Automatic speech recognizers (ASR) can be designed with the goal of mimicking human performance in speech recognition, hence, they can also be employed for predicting the intelligibility of speech. This paper presents two new objective measures for predicting speech intelligibility at the word level. The normalized likelihood difference (NLD) and the time alignment difference (TAD) are the proposed measures, extracted utilizing the hidden Markov models (HMMs) trained for an ASR system. Experimental results show that the proposed measures can accurately predict the normal-hearing listeners' performance in a keyword recognition task.

Index Terms: speech intelligibility prediction, automatic speech recognition, microscopic approach, objective measure

1. Introduction

The number of applications of devices working with speech signals is growing every day. For instance, many researchers are developing speech processing algorithms for hearing aids, which are widely needed in our aging societies. For these developments, it has always been a requirement to assess the intelligibility or quality of the signal at hand before or after processing. Partially automating this task rather than purely relying on listening tests is beneficial considering the time and cost required in human intelligibility assessment.

In the last decades, many objective measures have been published, which aim to predict the speech intelligibility from a macroscopic point of view. Well-known objective measures like the speech intelligibility index (SII) [1], speech transmission index (STI) [2], short time objective intelligibility (STOI) [3], and mutual-information-based models [4] compare the degraded speech with a reference in long segments, e.g. over an entire sentence, and predict only the average number of speech units, like words, heard correctly. The speech-based envelope power spectrum model (sEPSM) [5] is another example of macroscopic measures, which uses an auditory model to analyze the speech signal and computes the signal-to-noise ratios in modulation frequency bands as a measure of intelligibility. This model was later extended to mr-sEPSM [6] and sEPSM-corr [7] in order to account for non-linear degradations as well.

Macroscopic measures typically require longer input signals in order to obtain a sufficient accuracy in intelligibility prediction. In contrast to such methods, microscopic approaches process smaller segments of speech and attempt to predict the individual listener's response to a speech signal on a word-by-word or phoneme-by-phoneme basis. As an example, the microscopic method proposed in [8] uses an auditory model to extract features from speech signals and the dynamic time warping algorithm to compare the features extracted from a degraded signal to its clean counterpart for predicting the intelligibility of single words.

In another microscopic framework [9], Kollmeier et al. have considered the outputs of an ASR system as predictors of speech perception in both normal-hearing and hearing-impaired listeners. In this method, in contrast to the previously mentioned intelligibility prediction methods, it is not required to have access to the clean signal as a reference for predicting the speech intelligibility. Also, this method can benefit from the language knowledge implemented as a grammar in ASR systems. In [10], it has been shown that in listening tests, humans are taking advantage of their prior knowledge about the characteristics of speech units such as phonemes. Therefore, the authors have suggested to take the phonetic information into account in the design of instrumental quality or intelligibility measures. Otherwise, comparing the processed speech only to a signal-based reference can lead to unreasonably low quality estimates in scenarios like artificial speech bandwidth extension. A non-intrusive prediction of intelligibility has been introduced in [11, 12] that uses either the oracle transcriptions or the ASR-recognized transcriptions of the speech signal and synthesizes the clean features, required inside an intrusive intelligibility prediction method.

Microscopic methods promise to be more precise in estimating intelligibility and in diagnosing problems due to specific phoneme confusions. We have previously proposed an approach [13] that uses the logarithm of likelihood ratio of the true and the ASR-recognized word as an objective metric for the intelligibility prediction. In this paper, we introduce two other new heuristic measures for predicting the speech intelligibility from a microscopic point of view. The proposed measures are extracted utilizing an HMM-based ASR system, which will be explained and inspected in detail in the following sections.

2. ASR-based Microscopic Intelligibility Measures

Within the process of recognizing a speech signal, an HMM-based ASR system can compute some intermediate features that are indicative of its confidence. The likelihood of N-best state sequences or N-best word choices are primary examples [14]. Consequently, it can be hypothesized that such features contain information about the intelligibility of speech units as well. Moreover, it can be stated that the less intelligible a speech signal is, the more errors are expected in the ASR output. Hence, the time alignment information, estimated during the recognition process, can be used as another source of information on the intelligibility of speech units. In order to exploit such HMM-based features in the context of speech intelligibility prediction, the normalized likelihood difference (NLD) and the time alignment difference (TAD) are introduced in this paper.

Prior to extracting the NLD and TAD, an HMM-based

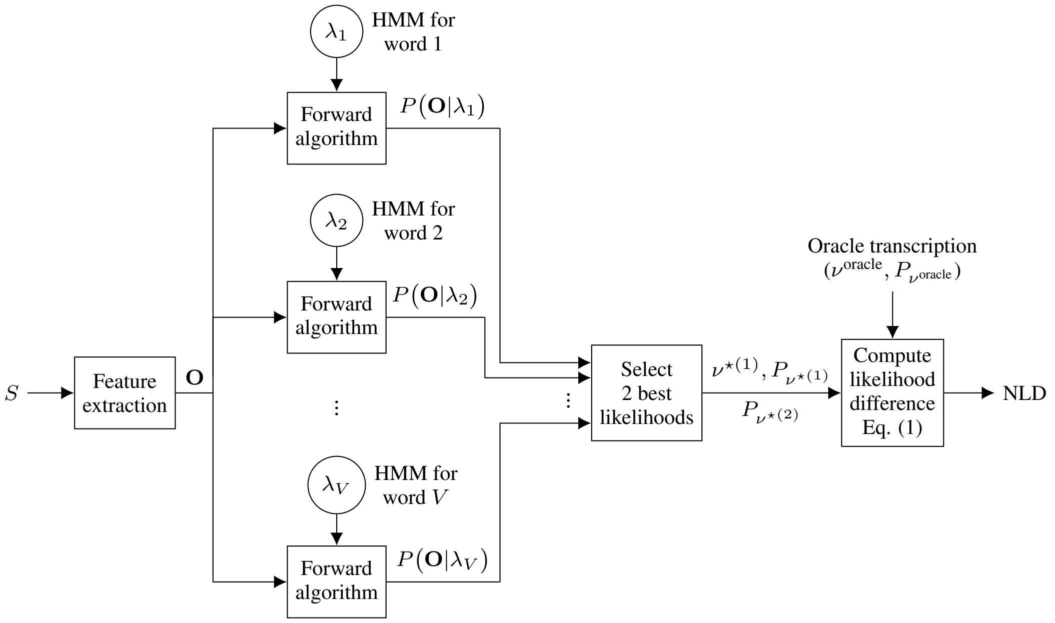
λ_1
HMM for word 1
Forward algorithm
$P(\mathbf{O}|\lambda_1)$
λ_2
HMM for word 2
Forward algorithm
$P(\mathbf{O}|\lambda_2)$
⋮
λ_V
HMM for word V
Forward algorithm
$P(\mathbf{O}|\lambda_V)$
S
Feature extraction
$\mathbf{O}$
Select 2 best likelihoods
$\nu^{\star(1)}, P_{\nu^{\star(1)}}$
$P_{\nu^{\star(2)}}$
Oracle transcription
$(\nu^{\text{oracle}}, P_{\nu^{\text{oracle}}})$
Compute likelihood difference Eq. (1)
NLD

Figure 1: *Block diagram of the first proposed speech intelligibility measure, the NLD.*

speech recognition system must be trained. Here, for each word ν in the vocabulary, one HMM λ_ν is built. The parameters of each HMM can be estimated by optimizing the likelihood of the training set observation vectors for the associated word or by discriminative training. Based on these trained models, Figure 1 and Figure 2 illustrate the schematic diagram of extracting the proposed measures NLD and TAD, respectively. The detailed description of these measures is provided below. It is notable that the introduced measures, here, are extracted at the word level, however, it is possible to extend the current framework for predicting the perception of phonemes as well.

2.1. Normalized Likelihood Difference (NLD)

In order to extract the NLD per word, the speech signal is segmented into the constituent words and each segment is fed into the system as an input. The first step in extracting the NLD is to apply a feature extraction method to the input signal S and estimate the observation sequence $\mathbf{O} = \{\mathbf{o}_1\mathbf{o}_2 \dots \mathbf{o}_T\}$. Then, the model likelihoods given the observation sequence are computed for all possible words, $P\big(\lambda_\nu|\mathbf{O}\big), 1 \leq \nu \leq V$. All model likelihoods are sorted to find the first $\nu^{\star(1)}$ and second $\nu^{\star(2)}$ most likely word:

$$\begin{aligned}\nu^{\star(1,2)} &= \underset{1\leq\nu\leq V}{\arg\max}^{(1,2)}\big[P\big(\lambda_\nu|\mathbf{O}\big)\big] \\ &= \underset{1\leq\nu\leq V}{\arg\max}^{(1,2)}\big[\frac{P\big(\mathbf{O}|\lambda_\nu\big)P\big(\lambda_\nu\big)}{P\big(\mathbf{O}\big)}\big] \qquad (1)\end{aligned}$$

Here, $P\big(\lambda_\nu|\mathbf{O}\big)$ is the likelihood of the word model λ_ν given the observation sequence $\mathbf{O}$ and V is the number of all possible words.

Since the probability of the observation sequence, $P\big(\mathbf{O}\big)$, is independent of word models and the prior probability of each model, $P\big(\lambda_\nu\big)$, is equal here for all possible words, Equation (1) can be reformulated to

$$\nu^{\star(1,2)} = \underset{1\leq\nu\leq V}{\arg\max}^{(1,2)}\big[P\big(\mathbf{O}|\lambda_\nu\big)\big]. \qquad (2)$$

As shown in Figure 1 and according to Equation (2), following the feature extraction, the forward algorithm is applied to the observation sequence in order to compute the probability of the input sequence given the model λ_ν for all possible words. Next, all probabilities are sorted and the words with the first and second highest probability are selected. The word with the highest probability $\nu^{\star(1)}$ is compared to the oracle transcription of the signal ν^{oracle}. If they are equal, the NLD is defined to be the difference between the likelihoods of the first $\lambda_{\nu^{\star(1)}}$ and the second $\lambda_{\nu^{\star(2)}}$ best word models given the observation sequence and normalized with the best likelihood. Otherwise, the order of the difference between the two model likelihoods is interchanged:

$$\text{NLD} = \begin{cases} \dfrac{P\big(\lambda_{\nu^{\star(1)}}|\mathbf{O}\big) - P\big(\lambda_{\nu^{\star(2)}}|\mathbf{O}\big)}{P\big(\lambda_{\nu^{\star(1)}}|\mathbf{O}\big)}, & \text{if } \nu^{\star(1)} = \nu^{\text{oracle}} \\ \dfrac{P\big(\lambda_{\nu^{\text{oracle}}}|\mathbf{O}\big) - P\big(\lambda_{\nu^{\star(1)}}|\mathbf{O}\big)}{P\big(\lambda_{\nu^{\star(1)}}|\mathbf{O}\big)}, & \text{if } \nu^{\star(1)} \neq \nu^{\text{oracle}} \end{cases} \qquad (3)$$

Similar to Equation 2, here, $P\big(\lambda_\nu|\mathbf{O}\big)$ can be replaced by $P\big(\mathbf{O}|\lambda_\nu\big)$ and the NLD is computed in practice using the probability of the observation sequence given the word models.

2.2. Time Alignment Difference (TAD)

As a second metric, we consider the time alignment difference (TAD) between the recognized and oracle transcriptions of the input signal. The computational steps of this measure are shown in Figure 2. Like the NLD, the TAD is estimated at the word level, but the input to this method is the entire sentence. At first, a feature extraction algorithm is applied to the input signal. Then, a decoder is employed to perform a continuous speech recognition on the observation sequence $\mathbf{O}$. Lastly, the recognized time alignment for a specified keyword is compared to its oracle time alignment and their relative difference is computed as the TAD measure

$$\text{TAD} = \frac{|Tb_{Rec} - Tb_{Oracle}| + |Te_{Rec} - Te_{Oracle}|}{L}, \qquad (4)$$

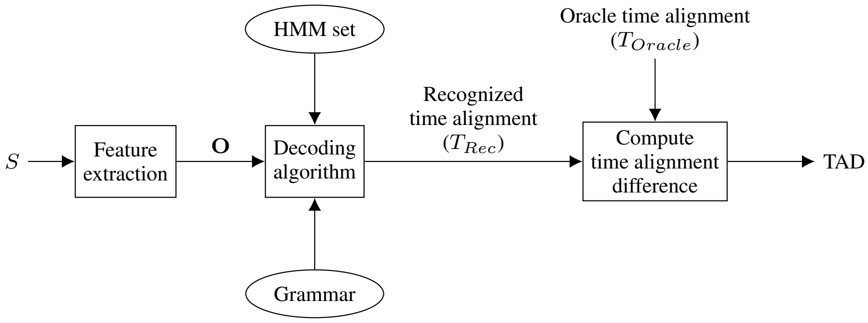

Figure 2: *Block diagram of the second proposed speech intelligibility measure, the TAD.*

where Tb_{Rec} and Tb_{Oracle} are the recognized and the oracle beginning frame index of a single word, respectively. Similarly Te_{Rec} and Te_{Oracle} represent the recognized and the oracle ending frame index of the same word and lastly, L is the word length in frames.

3. Experiments

3.1. Speech Intelligibility Database

The original Grid corpus [15] and its noisy version [16] have been used in the following experiments. The original corpus contains 34000 clean speech signals in total, recordings of 34 English speakers made at the University of Sheffield. Each Grid utterance is a 6-word sentence with a fixed grammar: <Verb (4)- Color (4)- Preposition (4)- Letter (25)- Digit (10)- Adverb (4)>, where the numbers in parentheses represent the number of available choices for each word type.

In addition to the clean Grid database, there is also a noisy version, which has been created by adding speech-shaped noise to the clean signals at 12 different signal-to-noise ratios (SNRs) from -14 dB up to 6 dB in steps of 2 dB, plus 40 dB (labeled as clean). For the noisy database, the results of a listening test conducted on 20 listeners with normal hearing are available. Each participant has listened to 2000 utterances and has been asked to recognize three keywords, the color, letter, and digit [16].

3.2. Experimental Setups

The first step in all experiments is the extraction of features from the speech signals. As features, the first 13 Mel frequency cepstral coefficients (MFCCs) plus their first (Δ) and second order derivatives ($\Delta\Delta$) were used. Hamming-windowed frames with a length of 25 ms and a frame shift of 10 ms were chosen for the MFCC extraction algorithm. The sampling frequency was set to 25000 Hz in all experiments.

For ASR, each word was modeled using a linear left-to-right HMM, resulting in 51 whole-word HMMs plus one silence model. The number of states were chosen as three times the number of phonemes of the modelled word. A 2-mixture diagonal covariance GMM represents the state output distribution of all HMMs.

In order to be able to use the entire data collected in the listening test for evaluating the intelligibility measures, all experiments were carried out with 5-fold cross validation. During each fold, the speech database was divided into the disjoint training (60%), development (20%), and test (20%) sets. To raise the accuracy, noise-dependent models were trained separately at each SNR and development sets were used to assess the accuracy of HMMs during the training.

To evaluate the proposed intelligibility measures, single Gaussian models (GM) were utilized to predict the intelligibility of the Grid keywords. For each test, two GMs were trained; one to represent the distribution of the intelligibility measure for correctly recognized words and another one representing the distribution of the same intelligibility measure but for words misrecognized by human listeners. Hence, in this framework, intelligibility measures were used as input features for GMs. After training, GMs were employed to predict whether an input speech signal can be recognized correctly at the word level. Here, the development set data were used for training the GMs and the test set data were used to evaluate them.

3.3. Evaluation

In the current work, the proposed intelligibility measures are assessed for predicting the normal-hearing listeners' performance. The results, presented below, are averaged over all 20 listeners available in the Grid database. Table 1 contains the accuracy of the proposed intelligibility measures NLD and TAD in predicting human keyword recognition results, averaged over 12 SNRs. Also, the accuracy of the ASR system in the same task is given in this table, which is computed by a direct comparison of ASR and human recognition outcomes. In addition to the mentioned methods, the well-known macroscopic intelligibility measure STOI [3] was used as a baseline to predict the intelligibility of Grid keywords. Please note that the STOI method needs longer units of speech for its computations and normally, it can not be used for computing the intelligibility of a single word. Therefore, we have augmented the length of every speech signal corresponding to a word by repeating the same signal several times to allow for a computation of the STOI per word. This repetition was implemented in the one-third octave band domain. After framing and extracting the one-third octave band representation of the signal in each frame, all frames were repeated several times for both degraded and reference speech signals without inserting any silence gaps. Therefore, no artifacts have been introduced to the signal, which might have disadvantaged the STOI method.

Considering the results in Table 1, it is evident that both proposed measures, NLD and TAD, have higher accuracies, on average, in comparison to the STOI and to the direct use of the ASR output. Moreover, the average accuracy of the TAD exceeds that of the NLD measure. A statistical significance analysis using Fisher's exact test [17] has shown that the TAD is statistically different from all competitors at a significance level of 0.01. Furthermore, the comparison of the NLD and ASR results with that of the STOI using the same test has shown that

Table 1: *Average accuracy of all considered intelligibility measures in predicting the keyword recognition performance of 20 normal-hearing listeners.*

ASR	STOI	NLD	TAD
78.70	77.24	78.76	**80.92**

both methods are statistically different from the STOI, at a significance level of 0.01 as well.

An SNR-based comparison of the above intelligibility prediction methods is provided in Figure 3. One can observe that the TAD has the highest accuracy in most SNRs down to -8 dB. The STOI has a comparable performance to that of the TAD at 2 dB and higher SNRs but its accuracy drops steeply in the middle of the plot. The NLD and ASR have a similar pattern and are less accurate at higher SNRs than the STOI and TAD measures. The NLD is performing better than the ASR in most SNRs except for the very high (greater than 4 dB) and very low (smaller than -12 dB) ones.

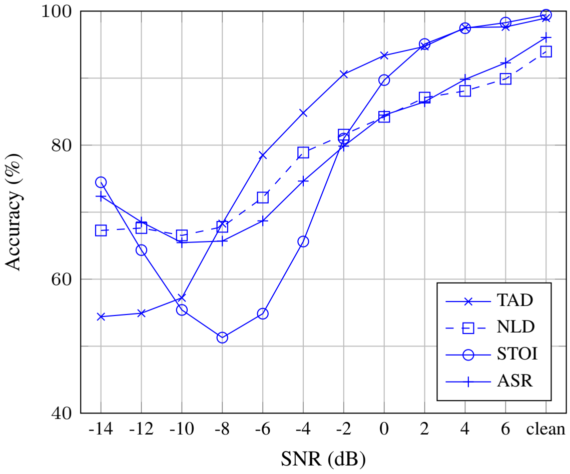

Figure 3: *Accuracy of all considered intelligibility measures per SNR in predicting the the keyword recognition performance of 20 normal-hearing listeners.*

4. Conclusions and Future Work

In this work, we have introduced two new intelligibility measures, NLD and TAD, both derived from a simple ASR system. These measures are proposed for predicting the intelligibility from a microscopic point of view. The NLD is computed based on the likelihood difference of the 2 best word choices and the TAD depends on the time alignment information. It was shown that, on average, both measures outperform the STOI method as well as the direct ASR system output. The TAD achieves a higher accuracy than the NLD. In some SNRs, both measures have lower accuracies in comparison to the baseline methods which needs more analysis. Extracting and appending complementary information to the proposed measures, and employing a discriminatively trained, DNN-based ASR can be considered as possible solutions for elevating the accuracy of our measures. The capacity of these measures for predicting the individual performance of hearing-impaired listeners will be examined in future works. Also, an extension of the NLD based on the likelihoods of n-best word hypotheses should be investigated in future work, with the goal of predicting likely word confusions.

5. Acknowledgements

This research has received funding from the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement n°[317521]. It was also supported in part by the National Science Foundation under Grant No. NSF PHY-1125915. The authors would like to thank Jon Barker for providing a noisy version of the Grid database with comprehensive listening test results.

6. References

[1] *Methods for the Calculation of the Speech Intelligibility Index*, S3.5-1997, ANSI, New York, NY, USA, 1997.

[2] H. J. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, Jan. 1980.

[3] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sept. 2011.

[4] J. Taghia and R. Martin, "Objective intelligibility measures based on mutual information for speech subjected to speech enhancement processing," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 1, pp. 6–16, Jan. 2014.

[5] S. Jørgensen and T. Dau, "Predicting speech intelligibility based on the envelope power signal-to-noise ratio after modulation-frequency selective processing," *J. Acoust. Soc. Am.*, vol. 129, no. 4, pp. 2384–2384, Apr. 2011.

[6] S. Jørgensen, S. D. Ewert, and T. Dau, "A multi-resolution envelope-power based model for speech intelligibility," *J. Acoust. Soc. Am.*, vol. 134, no. 1, pp. 436–446, 2013.

[7] H. Relaño-Iborra, T. May, J. Zaar, C. Scheidiger, and T. Dau, "Predicting speech intelligibility based on a correlation metric in the envelope power spectrum domain," *J. Acoust. Soc. Am.*, vol. 140, no. 4, pp. 2670–2679, 2016.

[8] T. Jürgens, S. Fredelake, R. M. Meyer, B. Kollmeier, and T. Brand, "Challenging the speech intelligibility index: macroscopic vs. microscopic prediction of sentence recognition in normal and hearing-impaired listeners," in *Proc. Interspeech 2010*, Sep. 2010, pp. 2478–2481.

[9] B. Kollmeier, M. R. Schädler, A. Warzybok, B. T. Meyer, and T. Brand, "Sentence recognition prediction for hearing-impaired listeners in stationary and fluctuation noise with fade empowering the attenuation and distortion concept by Plomp with a quantitative processing model," *Trends in Hearing*, vol. 20, pp. 1–17, 2016.

[10] T. Fingscheidt and P. Bauer, "A phonetic reference paradigm for instrumental speech quality assessment of artificial speech bandwidth extension," in *Proc. 4th International Workshop on Perceptual Quality of Systems*, 2013.

[11] M. Karbasi, A. H. AbdelAziz, and D. Kolossa, "Twin-HMM-based non-intrusive speech intelligibility prediction," in *Proc. ICASSP 2016*, Mar. 2016, pp. 624–628.

[12] M. Karbasi, A. H. Abdelaziz, H. Meutzner, and D. Kolossa, "Blind non-intrusive speech intelligibility prediction using twin-HMMs," in *Proc. Interspeech 2016*, Sep. 2016, pp. 625–629.

[13] M. Karbasi and D. Kolossa, "A microscopic approach to speech intelligibility prediction using auditory models," in *Proc. DAGA 2015*, Mar 2015, pp. 16–19.

[14] H. Jiang, "Confidence measures for speech recognition: A survey," *Speech communication*, vol. 45, no. 4, pp. 455–470, 2005.

[15] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 120, no. 5, pp. 2421–2424, 2006.

[16] J. Barker and M. Cooke, "Modelling speaker intelligibility in noise," *Speech Communication*, vol. 49, no. 5, pp. 402–417, 2007.

[17] A. Agresti, "A survey of exact inference for contingency tables," *Statistical science*, pp. 131–153, 1992.