PREDICTING ERROR RATES FOR UNKNOWN DATA IN AUTOMATIC SPEECH RECOGNITION

Bernd T. Meyer[1]*, Sri Harish Mallidi*[1]*, Hendrik Kayser*[2]*, Hynek Hermansky*[1]

[1]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA
[2]Medizinische Physik and Cluster of Excellence Hearing4all,
Carl von Ossietzky Universität, Oldenburg, Germany

ABSTRACT

In this paper we investigate methods to predict word error rates in automatic speech recognition in the presence of unknown noise types, which have not been seen during training. The performance measures operate on phoneme posteriorgrams that are obtained from neural nets. We compare average frame-wise entropy as a baseline measure to the mean temporal distance (M-Measure) and to the number of phonetic events. The latter is obtained by learning typical phoneme activations from clean training data, which are later applied as phoneme-specific matched filters to posteriorgrams (MaP). When exceeding a threshold after filtering, we register this as phonetic event. For test sets using 10 unknown noise types and a wide range of signal-to-noise ratios, we find M-Measure and MaP to produce predictions twice as accurate as the baseline measure. When excluding noise types that contain speech segments, a prediction error of 3.1% is achieved, compared to 15.0% for the baseline measure.

Index Terms— performance measure, error prediction, automatic speech recognition

1. INTRODUCTION

Error rates produced by automatic speech recognition (ASR) systems depend on many factors such as noise type and level, task complexity and speaker-specific characteristics. Methods that are able to estimate or predict the word error rate (WER) are often referred to as performance measures [8]. A typical application for such measures is the selection or weighting of streams in a multi-stream ASR system [1, 15], which should be especially useful in unknown conditions [5]. This paper explores three performance measures and their application to predict the WER in unseen scenarios, i.e., in noise types that have not been used during multi-condition training. All measures are calculated from phoneme posteriorgrams, which are obtained from softmax activations of a deep neural network (DNN). The baseline measure is frame-wise entropy, which has been motivated by the fact that noisy frames of the posteriorgrams often exhibit many class activations (and hence high entropy), while clean conditions often result in single class activations and low entropy [10, 12]. Second, the mean temporal distance (or M-Measure) has been proposed for performance monitoring for a phoneme recognition task and was successfully applied later in a multistream ASR setup in our earlier work [4, 8]. Third, we propose a matched filter approach using average phoneme activation patterns (MaP) learned from clean training data that - in contrast to the other two measures - takes into account the average duration of phonemes. We test the relation of the number of phonetic events that exceed a certain threshold with WER. Similar processing was suggested in [7], but in the context of keyword spotting rather than ASR. In [9], we suggested a related measure for WER prediction that also made use of matched filtering. However, that study has a different scope since it aims on signals obtained from behind-the-ear hearing aids. This study for the first time explores the predictive power of M-Measure and MaP while explicitly targeting unknown noise types.
In the following, the performance measures are introduced and the ASR experiments are outlined. In the results section, we investigate the effect of thresholding for matched filtering, before we compare the predictive power of each measure for a wide range of SNRs and unknown noise types with very different characteristics. Finally, we analyze the absolute WER prediction error and also report how prediction error relates to the observation time window or the number of utterances.

2. PERFORMANCE MEASURES

In this section, the performance measures used to predict word error rates are introduced. They are calculated from phoneme posteriorgrams obtained from deep neural nets, and later compared to average frame-wise entropy, which is used as a baseline measure.

2.1. Mean temporal distance: M-Measure

The mean temporal distance or M-Measure was proposed in [4] for performance monitoring in ASR and was shown to

This work was funded by Google via a Google faculty award to Hynek Hermansky and by the Cluster of Excellence 1077/1‘ “Hearing4all”.

be a good predictor of error rates in a phoneme classification task for two different noise types. It was also shown to outperform entropy in multistream ASR [8]. The measure takes into account the average difference of two vectors of phoneme posteriors $\mathbf{p}_{t-\Delta t}$ and $\mathbf{p}_t$ with a temporal distance Δt, and is given by

$$\mathcal{M}(\Delta t) = \frac{1}{T-\Delta t} \sum_{t=\Delta t}^{T} \mathcal{D}(\mathbf{p}_{t-\Delta t}, \mathbf{p}_t), \qquad (1)$$

where T is the duration of the analyzed posteriorgram. The Kullback-Leibler divergence is chosen as distance measure $\mathcal{D}$ between phoneme posterior vectors $\mathbf{p}_{t-\Delta t}$ and $\mathbf{p}_t$. The M-Measure is motivated by the fact that distant clean posterior frames will often be different since they are likely to represent different phonetic classes. On the other hand, uniform posteriorgrams with similar class activations over time often emerge from high noise levels. This results in lower differences between distant phoneme vectors. We consider a range of Δt from 50 ms to 800 ms (in steps of 50 ms), which results in 16 data points for each utterance [8]. These are averaged to obtain one scalar value per utterance.

2.2. Matched filtering and phonetic events

To distinguish between good and bad posteriorgrams (that should result in low or high WER), we explore matched filtering of phoneme trajectories followed by thresholding. We refer to supra-threshold activations as phonetic events. Our intuition is that the number of phonetic events per second should be informative for the quality of posteriorgrams. In the following, the two steps for calculating this measure are described:

1. Matched filters: Phoneme-specific filters are obtained from clean training data as suggested in [3]: A random subset of utterances was pushed through a DNN trained on Aurora 4 multi-condition data; 330 utterances were selected for this, which is the same number as for a standard testing set. Posteriors are calculated from softmax activations obtained from the final layer. They are converted to monophone activations by grouping the corresponding context-dependent triphone activations. A low threshold of 0.1 is applied which effectively separates phonetic islands of activation. The islands are centered in a 41-frame segment, averaged and normalized, which results in the filters shown in Figure 1 with phonetic classes denoted in ARPABET. Note how these filters capture average phoneme duration, with most vowels exhibiting relatively wide activations, while /P, T, K/ are comparatively short. This approach also has the potential to model asymmetric activations, which were however not observed on our data.

The maximum numerical value of the filter *output* depends on width and shape of the filter. Since we want to use a single threshold in the next step (in contrast to filter-specific thresholds), the filters are re-normalized using the normalization

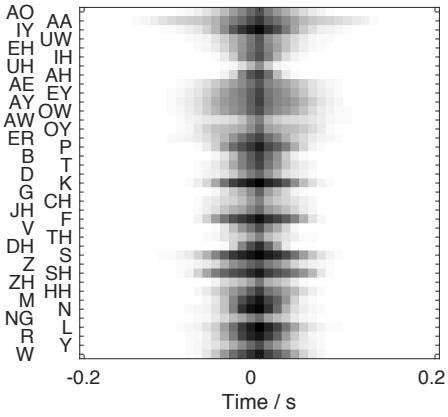

Fig. 1. *Phoneme-specific filter activations derived from clean training data.*

constant c_p for phoneme p. It is obtained by collecting the maximum values of filtered, clean posteriorgrams from training data. The 95% quantile of the maximum values is then chosen as c_p, which ranges from a value of 2.1 (for phoneme UH) to 8.4 (phoneme S). This procedure ensures the majority of post-filter values to be in the range from 0 to 1, and avoids that outliers dominate the rescaling.

2. Thresholding: A low threshold should result in low selectivity by producing a large number of phonetic events and false alarms. Vice versa, a very high threshold results in rare events that might be too sparsely distributed to cover the whole range from low to high-noise conditions. We therefore perform experiments to determine reasonable values for this threshold. An example of pre- and post-filter activations is shown in Fig. 2 for a noisy and clean activation of the same speech segment. Note how the dynamic range between noisy and clean signals is increased in the lower panel, which makes them easier to separate by simple thresholding.

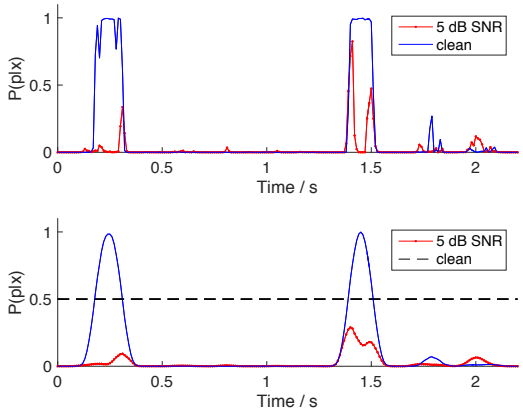

Fig. 2. *Example of activations for phoneme K obtained from high- and low-noise speech before and after filtering (top and bottom panel)*

3. ASR SYSTEM AND TEST DATA

3.1. Training data and system architecture

The ASR system was trained using the standard kaldi DNN recipe for Aurora 4 multi-condition data[14]. The DNN used six hidden layers, 2048 units per layer, and an additional softmax output layer. It was pre-trained as an RBM using contrastive divergence (CD-1) and supervised fine-tuning with the triphone targets via cross entropy. Every phone was modeled with three Hidden-Markov-Model (HMM) states except for the silence phone which was modeled with five states. 40-dimensional Mel-filterbank (FBANK) features were extracted from the 16 kHz audio data and fed to the DNN using an additional temporal context of 5+5 frames, resulting in 440-dimensional input to the neural net. Phoneme posteriorgrams were derived from the activations of the softmax output layer. Monophone posteriorgrams were obtained by grouping all triphones belonging to the same phone and subsequent summation of the corresponding activations. This was done since we found monophone processing to give similar results than its triphone equivalent at a lower computational cost. In the standard Aurora 4 task employed here, the following noise types are contained in the multi-condition training set: airport, babble, car, street, and restaurant. In total, 3569 utterances are used from training, all derived from the WSJ0 corpus using a vocabulary size of 5,000 words.

3.2. Unknown noise types in testing data

To explore the effect of unknown noises, ten maskers were selected that cover different types of long-term spectra and temporal modulations (Table 1). The standard clean Aurora 4 test set (eval92) was used as a basis to create noisy test sets with SNRs ranging from -5 to 25 in 5 dB steps. This results in 10 (noise types) $\times$ 7 (SNRs) $\times$ 330 (clean test files) utterances (23,100 in total), which are grouped in 70 test sets.

Index	Name	Source	Symbol
1	Vacuum Cleaner	BBC	o
2	Factory 1	Noisex	x
3	Factory 2	Noisex	+
4	Propeller Plane	BBC	*
5	Gym	BBC	⊡
6	ICRA1	DRE01	♢
7	Mall	BBC	v
8	Playground	BBC	^
9	Rain	BBC	<
10	Shower	BBC	>

Table 1. *Overview of unseen noise types and the corresponding origin (DRE01 [2], the BBC Sound Effects Library, or the Noisex database [11]).*

4. RESULTS

4.1. Threshold selection for matched filtering

To determine a reasonable threshold for MaP, the relation between WER and the number of phonetic events per second was analyzed for thresholds from 0.15 to 0.95. For a reliable estimation of WER without prior knowledge, the relation of WER and performance measure should be monotonic and exhibit a small variance across noise types. Examples for different thresholds are shown in Fig. 3. Each data point represents one noise type and SNR as denoted in Table 1.
For $T = 0.2$, the MaP value is not very informative for determining the WER: For instance, 15 events per second are observed for test conditions with more than 90% WER (*shower*) or less than 10% (*rain*). Similar variance was found for $T = 0.95$ (not shown for space restrictions). For thresholds around 0.5, the WER-MaP curve approaches a sigmoid function and exhibits less variance. To quantify this relation, a sigmoid function was fitted to the data and used to linearize the WER-MaP relation, from which the linear correlation is calculated. We found T not to be a very sensitive parameter: For values from 0.45 to 0.65, correlation values above 0.95 were obtained. The highest value ($r = 0.98, p < 0.0001$) was found for $T = 0.55$, which is therefore used in the following experiments.

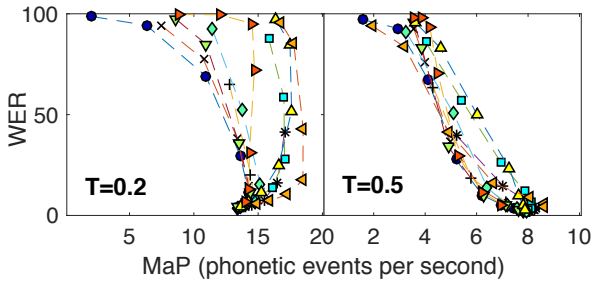

Fig. 3. *Relation of WER and phonetic events per second for different thresholds.*

4.2. Comparison of performance measures

The relation of word error rates obtained with the test sets described in the previous section to the performance measures under consideration are shown in Fig. 4. As described in Subsection 4.1, each color/symbol denotes a different noise type (cf. Table 1) that was not seen during training. A sigmoid fit is used to linearize each data set. From this, we obtain correlation values that indicate how well each performance measure is suited for WER prediction. The lowest value is obtained for entropy ($r = 0.89$), while higher coefficients are obtained with M-Measure ($r = 0.97$) and MaP ($r = 0.98$). In each case, p is below 0.0001.

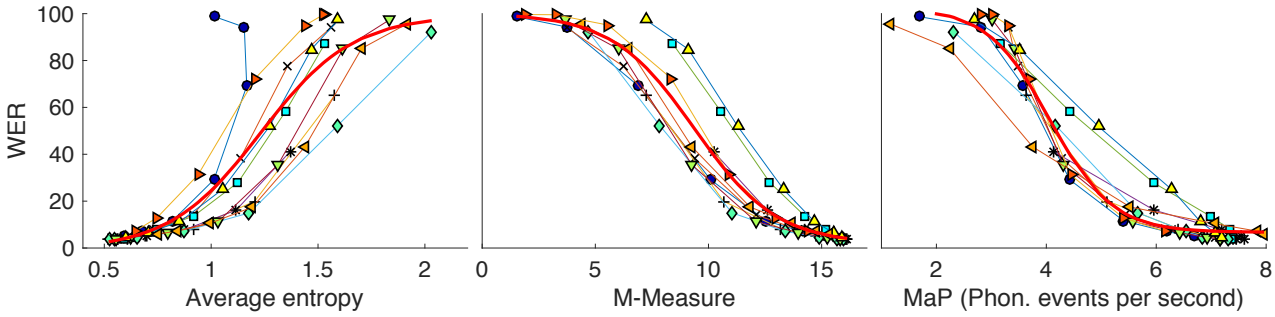

Fig. 4. *Relation of WER and performance measure. Each data point represents one noise type (denoted by the marker, cf. Table 1), and one SNR. A sigmoid function is fitted to each performance measure (thick solid line).*

4.3. Prediction error of performance measures

To evaluate how well the WER can be predicted, we analyze the absolute prediction error (PE), which is obtained by averaging the differences between WER data points and the corresponding fit. To ensure that unseen noise types are truly unseen, the fit is estimated for nine noise types, to which the WER for the remaining noise type is compared. This procedure is repeated for all ten noises. By taking into account the *absolute* WER difference, under- and overestimates of WER are treated equally. The trend observed for correlation is also reflected in the PE shown in Table 2 (*all noises*): MaP produces a slightly lower PE than M-Measure. Both measures outperform entropy, for which the relative PE is at least higher by a factor of 2.1. Since the number of phonetic events by matched filtering is tailored to speech, it should be less affected when the masking noise also contains speech elements, as masking speech will also produce supra-threshold phonetic events, which gradually replace the events from the target speech as the noise level increases. We therefore performed a separate analysis for non-speech noise types, omitting *mall*, *gym*, and *playground*. MaP was found to benefit from excluding maskers with speech elements (Table 2, columns *non-speech*), but an even stronger effect was observed for the M-Measure, with an average prediction error of only 3.1%. On the other hand, PE with entropy is further increased.

Finally, it was investigated how PE depends on the size of the observation window, i.e., the number of utterances used for obtaining WER and PE for each measure. Fig. 5 shows the PE for all three measures for reduced test sets, ranging from

	Noise types			
	a) all noises		b) non-speech	
	PE	Std	PE	Std
Entropy	11.3	12.1	15.0	15.5
M-Measure	6.0	6.5	**3.1**	3.6
MaP	**5.9**	5.4	4.8	4.2

Table 2. *WER prediction error (PE) and its standard deviation for the performance measures considering (a) all noise types or (b) only noises that contain no speech elements .*

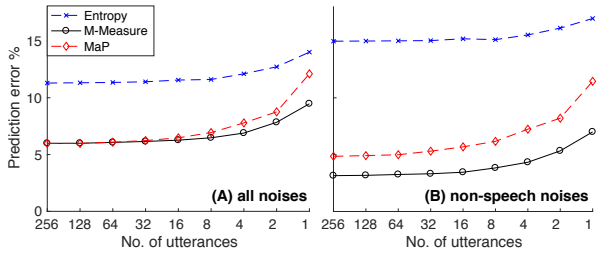

Fig. 5. *Prediction error over the number of utterances observed to estimate the WER for all conditions.*

256 to just one utterance from Aurora 4. All measures are affected by reduced sets, yet the results are relatively stable, losing less than 1% of predictive power when using 16 instead of 256 sentences. For noise set (B), M-Measure consistently outperforms MaP and entropy. However, when considering all noises (A) and long observation windows, MaP should be preferred over M-Measure since it provides almost identical PE, yet the computational complexity is 80 times lower than for M-Measure. The lower computational cost is achieved through linear filtering, in contrast to the calculation of KL-Divergence of many vectors for the M-Measure.

5. SUMMARY

This paper investigated three measures for predicting WER in ASR experiments. All measures were evaluated in noise types not used during training. Two measures from our earlier work (M-Measure [4, 8] and matched filtering of posteriorgrams (MaP) [7, 9]) were applied for the first time in this context, and were found to outperform the baseline with a prediction error (PE) of 6.0% and 5.9% compared to a 11.3% baseline. When considering all noise types and long observation windows, MaP produces a slightly better prediction error (PE) than M-Measure despite its low computational cost. The design choices of MaP motivated an analysis of results for noise types that do not contain speech segments. In this scenario, the PEs for M-Measure and MaP are reduced to only 3.1% and 4.8%, while the PE for entropy is further increased. Shortening the observation window to predict WER only gradually reduces the predictive power of all measures.

6. REFERENCES

[1] Bourlard, H., Dupont, S., Hermansky, H., Morgan, N. (1996). "Towards subband-based speech recognition," in Proc. EUSIPCO, pp. 1579-1582.

[2] Dreschler, W. A., H, V., Ludvigson, C., and Westermann, S. (2001). "ICRA Noises: Artificial Noise Signals with Speech-like Spectral and Temporal Properties for Hearing Instrument Assessment," Audiology, 40, pp. 148-157.

[3] M. Lehtonen, P. Fousek, and H. Hermansky, Hierarchical approach for spotting keywords, IDIAP Research Report, no.05-41, 2005.

[4] Hermansky, H., Variani, E., and Peddinti, V. (2013). "Mean temporal distance: Predicting ASR error from temporal properties of speech signal," ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process.

[5] Hermansky, H. (2013). "Multistream recognition of speech: Dealing with unknown unknowns," Proc. IEEE, 101, 1076-1088. doi:10.1109/JPROC.2012.2236871

[6] Hirsch, H. G., and Pearce, D. (2000). "The AURORA Experimental Framework For The Performance Evaluation of Speech Recognition Systems Under Noisy Conditions," Proc. Autom. Speech Recognit. Challenges for the new Millenium, pp. 29-32.

[7] Kintzley, K., Jansen, A., and Hermansky, H. (2011). "Event selection from phone posteriorgrams using matched filters," Proc Interspeech, pp. 1905-1908.

[8] Mallidi, S. H., Ogawa, T., and Hermansky, H. (2016). "Uncertainty estimation of DNN classifiers," IEEE Work. Autom. Speech Recognit. Understanding (ASRU), pp. 283-288.

[9] Meyer, B.T., Mallidi, H., Castro Marti nez, A.M., Paya-Vaya, G., Kayser, H., Hermansky, H. (2016). "Performance monitoring for automatic speech recognition in noisy multi-channel environments," submitted to the IEEE Workshop on Speech and Language Technologies.

[10] Misra, H., Bourlard, H., and Tyagi, V. (2003). "Entropy-Based Multi-Stream Combination," Proc. ICASSP, pp. 741-744.

[11] Varga, A., and Steeneken, H. J. M. (1993). "Assessment for automatic speech recognition: II NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech Commun., 12, 247-251.

[12] Okawa, S., Bocchieri, E., and Potamianos, A. (1998). "Multi-band speech recognition in noisy environments," Proc. 1998 IEEE Int. Conf. Acoust. Speech Signal Process (ICASSP), pp. 641-644.

[13] Parihar, N., Picone, J., Pearce, D., and Hirsch, H. (2003). "Performance analysis of the Aurora large vocabulary baseline system," Proc. of Eurospeech, pp. 10-13.

[14] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K. (2011). "The Kaldi Speech Recognition Toolkit," in Proc. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, pp. 1-4.

[15] Ravuri, S., and Morgan, N. (2010). "Using Spectro-Temporal Features to Improve AFE Feature Extraction for ASR," In Proc. INTERSPEECH.

[16] Spille, C., Kayser, H., Hermansky, H., Meyer, B.T. (2016). "Assessing speech quality in speech-aware hearing aids based on phoneme posteriorgrams," in Proc. Interspeech