PERFORMANCE MONITORING FOR AUTOMATIC SPEECH RECOGNITION IN NOISY MULTI-CHANNEL ENVIRONMENTS

Bernd T. Meyer[1], *Sri Harish Mallidi*[1], *Angel Mario Castro Martínez*[2,4], *Guillermo Payá-Vayá*[3,4], *Hendrik Kayser*[2,4], *Hynek Hermansky*[1]

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

ABSTRACT

In many applications of machine listening it is useful to know how well an automatic speech recognition system will do before the actual recognition is performed. In this study we investigate different performance measures with the aim of predicting word error rates (WERs) in spatial acoustic scenes in which the type of noise, the signal-to-noise ratio, parameters for spatial filtering, and the amount of reverberation are varied. All measures under consideration are based on phoneme posteriorgrams obtained from a deep neural net. While frame-wise entropy exhibits only medium predictive power for factors other than additive noise, we found the medium temporal distance between posterior vectors (M-Measure) as well as matched phoneme filters (MaP) to exhibit excellent correlations with WER across all conditions. Since our results were obtained with simulated behind-the-ear hearing aid signals, we discuss possible applications for speech-aware hearing devices.

Index Terms— automatic speech recognition, performance measures, spatial filtering, hearing aids

1. INTRODUCTION

Human listeners usually know how well they are doing in terms of speech recognition in a given acoustic scene that involves spoken language. This kind of knowledge would provide a useful measure in automatic speech recognition (ASR) as well, i.e., given an utterance in noisy or reverberant conditions, it would be advantageous to have an estimate for the error rate. The module which performs the estimation is typically referred to as performance monitoring [14]. One possible application lies in multistream ASR, in which each stream provides a different view on the signal. When effectively selected or combined, this approach resulted in robust ASR [1, 20].

In this paper, we explore measures based on phoneme posteriorgrams obtained from neural nets for performance monitoring. Average frame-wise entropy of posteriors has been proposed earlier for this task, based on the observation that noise often results in multiple phone activations per frame, thereby increasing its entropy [16, 13]. It has also been applied for estimation of speech quality in speech-aware hearing aids using posteriorgrams [22]. Second, the mean temporal distance (or M-Measure) has been proposed in [7] for performance monitoring for a phoneme recognition task and was successfully applied later in a multistream ASR setup [14]. An approach that takes into account different average durations of phoneme classes is matched filtering of posteriorgrams (which we refer to as MaP). In this approach, filters are learned from clean posteriors or labels and convolved with posterior activations with the aim of obtaining high values for phonetic events. Our intuition is that low-noise speech should produce more of these events, which could be useful information for performance monitoring. This is supported by results obtained in [11], in which matched filters improved a keyword spotting system. We explore a modified version for performance monitoring in ASR and compare the results to entropy and M-Measure.

Optimally, a performance measure should predict the error rate of a system not just for one specific setting, but generalize over different situations that occur in typical applications in speech processing. We therefore test the three measures in different acoustic scenes: Spatial scenes are simulated in which a speaker is positioned on the left side of a virtual listener, where signals are obtained using behind-the-ear hearing aids (Fig. 1). The hearing aid signals provide multiple channels, which allows to perform beamforming for enhancing different angles. When the beamformer aims at the speaker, word error rate (WER) is expected to be lowest, and a good performance measure should also differentiate the correct angle from other directions. Further, we investigate the effect

of localized and diffuse noise types in anechoic and reverberant conditions. Although the main focus of the study is on ASR, we briefly discuss the feasibility of our methods to be used in speech-aware hearing aids given their hardware limitations. Finally, it is analyzed which phoneme representation (i.e., context-dependent triphones or monophones) should be preferred for applying performance measures.

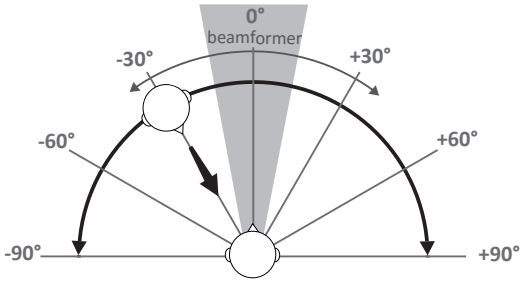

Fig. 1. *Illustration of the spatial scenes investigated in this study: We simulate signals captured by behind-the-ear hearing aids with a lateral speaker. Diffuse or spatial noise (at $-40°$) is added to create two different acoustic scenarios.*

2. PERFORMANCE MEASURES

In the following, the performance measures that are applied to predict the word error rate are described. Both measures are later compared to average frame-wise entropy of phoneme posteriorgrams, which is used as a baseline.

2.1. Mean temporal distance: M-Measure

The mean temporal distance or M-Measure was proposed in [7] for performance monitoring in ASR and was shown to 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 [14]. 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. As in [7], the Kullback-Leibler divergence was chosen as distance measure $\mathcal{D}$ between phoneme posterior vectors $\mathbf{p}_{t-\Delta t}$ and $\mathbf{p}_t$. We consider a range of Δt from 10 to 50 ms (in steps of 10 ms) and 100 to 800 ms (in steps of 50 ms), which results in 20 data points for each utterance as shown in the example in Fig. 2. For small values of Δt, a small average KL-divergence is obtained which reflects that neighbouring phoneme frames will often be similar. When moving to Δt of 200 ms and higher, M-Measure values typically saturate reflecting the effect of phoneme duration and coarticulation on the shape of the curve. High noise levels often result in similar class activations over time (cf. third panel in Fig. 4), hence the maximum distance is decreased for noisy conditions. Both effects can be observed in the example in Fig. 2. Since we assume that the average value of the saturated curve is informative for ASR WER, we average data points of the M-Measure curve from 5 to 80 frames to obtain a scalar performance measure.

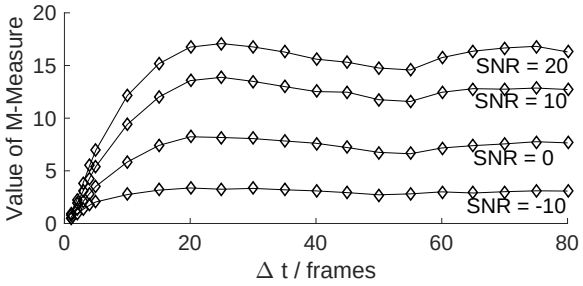

Fig. 2. *Average temporal distance (or M-Measure) for one utterance from TIMIT for stationary speech-shaped noise at various SNRs.*

2.2. Matched Phoneme (MaP) filters

A third measure is obtained from matched phoneme filters which are convolved with the temporal trajectories of posteriorgrams. The idea is that filters matched to the average activation pattern of phonemes should produce a high peak for robust phoneme representations. For degraded posteriorgrams, these peaks should be less pronounced; this feature could be used to assess the quality of the posteriorgram representation. With this approach, we extract the number of peaks per second (PPS) from the processed posteriorgram.

The procedure is as follows: First, phoneme-specific filters are obtained from clean data or labels. In this study, filters were estimated from TIMIT data by averaging labels for each phoneme class. Compared to the other performance measures considered here, this has the advantage that prior knowledge about phoneme duration (e.g., on average 'OW' is longer than 'P') is explicitly taken into account in this measure. The estimated filters are shown in Fig. 3. Second, posteriorgrams are obtained from computing the softmax activations from the last layer of the DNN. For the analysis of monophone posteriorgrams, the activations were grouped accordingly. To obtain a sparse representation from the posteriorgram, a threshold is applied to its values, as has been suggested in [9]. Next, each row of the resulting representation is convolved with the corresponding matched filter. In related study on keyword spotting [11], matched filtering was performed *after* a threshold was applied. This strategy was also tested since it preserves the shape of activations that the filters are supposed to match.

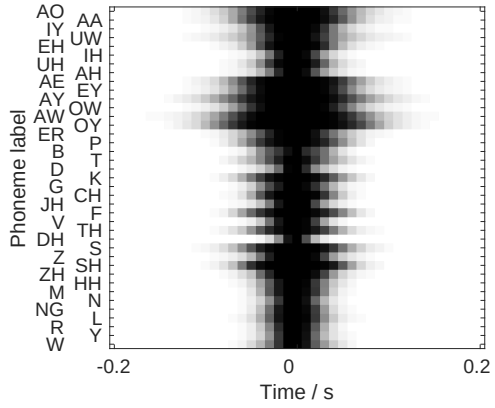

Fig. 3. *Phoneme-specific filter activations derived from TIMIT labels.*

When the threshold is applied first, parts of the activations that we try to match are removed. On the other hand, early thresholding also removes small (and potentially unreliable) estimates of posteriors, which should be helpful to focus on reliable activations, i.e., there is a trade-off between these two effects. In pilot experiments, thresholding ($T = 0.5$) followed by filtering produced good correlations of SNR and the performance measure, which hence motivated its use to predict WER. Localized activation patterns are obtained through the filter process; in each localized trajectory, the position of the local maximum is determined. Finally, the number of phoneme peaks per second is obtained from this representation. An example for unfiltered posteriorgrams (clean and noisy) and the processed output with localized maxima is shown in Figure 4. In this example, the number of peaks is heavily reduced by adding speech-shaped, stationary noise. In the next section it will be investigated if this measure is also useful for different noise types, varying parameters of spatial filtering and in different environments.

3. SIGNALS, SCENES, AND RECOGNIZER

3.1. Generation of spatial scenes

To investigate the results of spatial filtering and additive noise on performance measures and WER, the standard Aurora 4 clean eval92 test set was used as a basis. Spatially localized and diffuse sound sources are simulated using a database of head-related impulse responses (HRIR), which features impulse responses recorded with three microphones from each of two behind-the-ear (BTE) hearing aids attached to left and the right ear. The HRIRs used in this study are a subset of the database described in [10]: Anechoic free-field HRIRs and reverberated HRIRs from the frontal horizontal half-plane were measured at a distance of 3 m and 1 m between microphones and loudspeaker, respectively. All HRIRs (anechoic and reverberated) from the database were measured with 5° resolution for the azimuth angles, which was limited in this study to 10° to obtain a feasible number of ASR test sets. Reverberated HRIRs were measured in a typical office environment with a reverberation time of ∼300 ms.

Spatial signal enhancement is conducted in the frequency domain by multiplying the multi-channel STFT $\mathbf{x}(\omega, t)$ of the input signal from the six BTE input channels with a spatial filter vector $\mathbf{w}(\alpha, \omega)$, α being the steering direction, yielding the single-channel output $y(\alpha, \omega, t)$:

$$y(\alpha, \omega, t) = \mathbf{w}^{\mathsf{H}}(\alpha, \omega)\, \mathbf{x}(\omega, t). \tag{2}$$

We apply MVDR beamforming ('minimum variance distortionless response', [3, 2]) and obtain $\mathbf{w}(\alpha, \omega)$ from the steering vector $\mathbf{d}(\alpha, \omega)$; the noise covariance matrix $\mathbf{R}(\omega)$ is calculated according to

$$\mathbf{w}(\alpha, \omega) = \frac{\mathbf{R}^{-1}(\omega)\, \mathbf{d}(\alpha, \omega)}{\mathbf{d}^{\mathsf{H}}(\alpha, \omega)\, \mathbf{R}^{-1}(\omega)\, \mathbf{d}(\alpha, \omega)}. \tag{3}$$

In the current approach solely head-related characteristics of sound propagation are included in the signal enhancement setup and no further information about room acoustics is exploited. Hence $\mathbf{d}(\alpha, \omega)$ is computed from the anechoic HRIRs according to a given α and $\mathbf{R}(\omega)$ from the whole set of anechoic HRIRs, resembling a spatially diffuse, white noise field as captured by the BTE devices.

Figure 1 shows a sketch of the first acoustic scene under consideration. Spoken utterances from a fixed azimuth angle of -30° were mixed with random parts of a spatially diffuse stationary speech-shaped noise at signal-to-noise ratios (SNR) from -10 to 10 dB in 5 dB steps. In a second scenario, the diffuse noise was replaced with a localized vacuum cleaner noise positioned at +40° azimuth using the same range of SNRs; the noise signal was taken from the BBC Sound Effects Library.

3.2. ASR setup

The ASR system was trained using the standard kaldi DNN recipe for anechoic multi-condition data (Aurora 4)[19]. Clean-condition training was also considered, but produced generally very high error rates for the chosen SNR range from -10 to 10 dB in pilot experiments and was hence not included in the experiments presented here. The DNN used six hidden layers, 2048 units per layer, and an additional softmax output layer. It was pre-trained as a 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

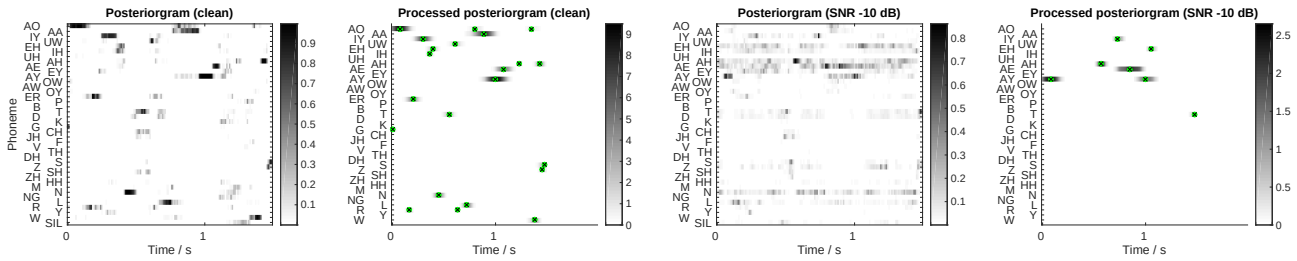

Fig. 4. *Comparison of clean and noisy posteriorgram, as well as the processed posteriorgrams obtained for a segment of one utterance. In the clean case, 22 peaks are obtained with a threshold of $T = 0.5$, while at -10 dB SNR only 7 peaks occur.*

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. Test signals were generated as outlined in the previous subsection. In total, 330 (clean test sentences) $\times$ 2 (rooms) $\times$ 2 (noise types) $\times$ 19 (directions) $\times$ 5 (SNRs) were produced (approx. 125k test utterances).

4. RESULTS

4.1. Comparison of performance measures

Fig. 5 compares three performance measures with word error rate. Two noise scenarios, five SNRs and 19 directions the beamformer was steered to are pooled in each subplot to analyze how well the measures generalize over different conditions. In case of entropy (left panels) there is no consistent relation, neither for the anechoic case nor the office situation. The lower left panel shows data points with a similar entropy (values between 1.5 and 2) that belong to the same SNR (-10 dB, black diamonds) but are scattered across a large WER range, which is caused by different beamforming angles. This means that in this condition the degradation caused by spatial filtering is not captured by entropy. In contrast to this, for the M-Measure and MaP, a clear relation with WER emerges showing that effects of SNR, spatial filtering and noise types introduced to the posteriorgram are reflected by both measures. The data shows ceiling effects for the lowest SNR (-10 dB). For higher SNRs (> 10 dB, not considered here), we expect a flooring of WER, which would result in a sigmoid shape of the data. We therefore fit a sigmoid function to the data obtained with M-Measure and MaP and report the root-mean-square difference of data points and the fitted curve to evaluate the measures. The fit is also used to linearize the data and calculate Pearson's correlation coefficient. In an anechoic environment, a higher correlation with WER is obtained with the M-Measure ($r = 0.962$) than with MaP ($r = 0.842$). The opposite trend is observed in a reverberated office environment (M-Measure: $r = 0.669$, MaP: $r = 0.874$).

In comparison to standard Aurora 4 results, the WERs we obtain are comparatively high with error rates often reaching 100% for high-noise conditions. This can be attributed to the low SNRs chosen here that range from -10 to +10 dB, while the original Aurora 4 test set exhibits random SNRs between 10 and 20 dB [18]. A second important factor is reverberation introduced in the office condition. To decrease WERs, strategies to cope with reverberated signals could be applied [23], but are out of scope for this study.

The data presented in Fig. 5 was obtained with the full Aurora 4 test set (330 utterances, each approx. 5 seconds). This could be useful for scenarios in which acoustic conditions do not change rapidly, but ultimately a performance measure would be much more useful when estimation on shorter time scales is possible. We investigated this by looking at a smaller number of utterances, and report the corresponding correlations in Table 1. Correlations over 0.7 are obtained even if just a single randomly selected utterance is used.

Fig. 6 shows an example for one utterance (speaker at 30°) and how M-Measure and MaP are affected by different SNRs and beamforming angles. Both measures peak at the location of the speech source and show lower values for off-speaker directions as well as higher SNRs.

#Utts.	Anechoic		Office	
	M-Meas.	MaP	M-Meas.	MaP
1	0.911	0.705	0.722	0.724
2	0.935	0.773	0.644	0.777
4	0.955	0.802	0.538	0.717
8	0.956	0.847	0.730	0.837
16	0.959	0.858	0.687	0.885
32	0.962	0.853	0.660	0.793
64	0.964	0.844	0.637	0.875

Table 1. *Correlation values for M-Measure and MaP for different number of Aurora 4 utterances from which WER and performance measure were calculated.*

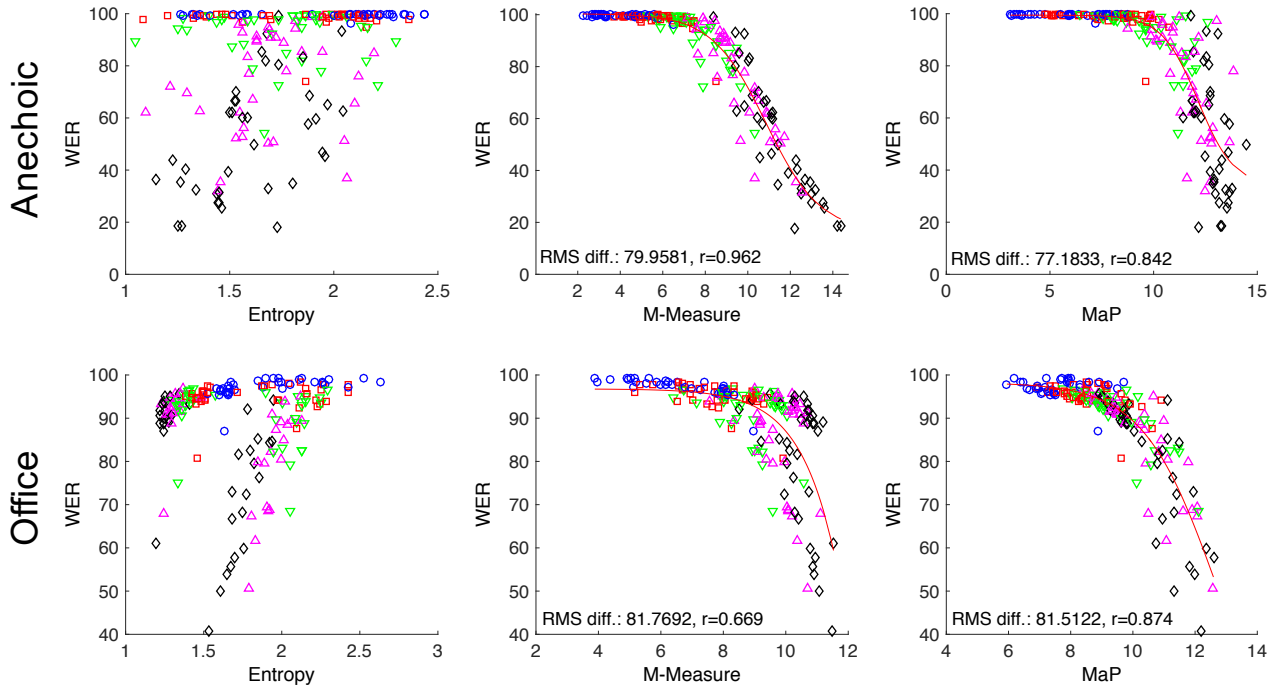

Fig. 5. *Relation of average WER and performance measures for an anechoic environment (top row) and an office (bottom). Color encodes the test SNR, ranging from -10 dB (blue circles) to +10 dB (black diamonds). The data shown here covers two different noise types and 19 beamforming angles per condition. For M-Measure and MaP, a sigmoid function was fitted to the data; in the subplots, the RMS difference between fit and actual data is shown, as well as the correlation with WER.*

4.2. Triphone vs monophone activations

An important question when investigating performance monitoring is to which specific representation the measures should be applied. Previous work using the M-Measure has been carried out on monophone posteriorgrams for phoneme recognition [7] or context-dependent triphone activations, which have been become standard in DNN-based ASR systems [14]. In this study, the relation of this factor to predictive power of the measure is analyzed by application of the measures to both monophone as well as triphone activations. Note that the matched filter approach only used monophone activations, since the filters were obtained from monophone TIMIT label data. Table 2 compares the correlation values obtained for both measures after linearizing the data through a sigmoid fit, as described in the previous section. While for anechoic data almost identical correlations are obtained, using triphone representation increases the quality of the M-Measure, i.e., in this reverberant condition, it is beneficial not to merge the triphone classes to monophones.

	Room	Posteriorgram	r
M-Measure	Anechoic	Tri	0.9620
M-Measure	Anechoic	Mono	0.9621
MaP	Anechoic	Mono	0.8420
M-Measure	Office	Tri	0.8191
M-Measure	Office	Mono	0.6694
MaP	Office	Mono	0.8736

Table 2. *Correlation values for the M-Measure applied to triphone and monophone activations as well as for the MaP method, which was only test on monophones. All correlation values were highly significant ($p < 0.001$).*

4.3. Technical feasibility for hearing devices

The main focus of this study is on performance monitoring for ASR in multi-channel scenarios. However, the signals under consideration were obtained from impulse responses recorded in behind-the-ear hearing aids, which motivated an assessment of the applicability of our findings in assistive hearing technologies. We assume that for the parameters varied in this work, the relation between speech intelligibility in normal-hearing and hearing-impaired listeners on the one hand, and ASR WER on the other is straight-forward: The best performance can be expected for a lower SNR and a beamformer directed to a spatial source. The measures analyzed could therefore be valuable in optimizing parameters of hearing devices, e.g., by monitoring several beamforming directions with the

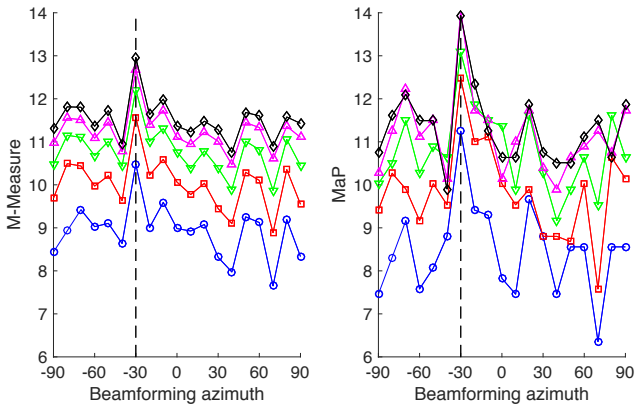

Fig. 6. *Relation of performance measures with SNR and beamforming azimuth α for the office condition with diffuse noise, using one utterance from Aurora 4 to obtain each data point. As in Fig. 5, color encodes the test SNR, ranging from -10 dB (blue circles) to +10 dB (black diamonds).*

aim of selecting the optimal azimuth indicated by high values for the corresponding performance measure. It is however unclear if this is technically feasible, since it requires the generation of posteriorgrams on small-footprint hardware.

On the Kavuaka processor, an application-specific integrated circuit (ASIC) developed at the Cluster of Excellence Hearing4all for hearing aids and digital signal processing, performing a feed forward pass on a deep learning architecture as the one used in this study (i.e., six hidden layers with 2048 units per layer, and an additional softmax output layer) will approximately take 300 ms per frame, taking profit on the available subword and instruction parallelism mechanisms implemented on the Kavuaka processor[12]. Comparable performance can be yielded by ASICs that are used in current hearing aids, such as the CoolFlux DSP [21]. The main constraint in these ASICs is the dynamic range of the parameters, for instance the Kavuaka processor can efficiently emulate floating-point arithmetic (i.e., 24-bit signficant and 8-bit exponent floating-point format) [6] which provides high precision within a narrow range. Therefore, methods like batch normalization on the DNN side and CMVN on the features side are imperative for using DNN-based processing on actual hearing aid hardware.

Although the current implementation of our DNN is not capable of running in real time on the hardware mentioned (with a real-time factor (RTF) of 300 ms/10 ms), it should be possible to reach an RTF below 1.4 by reducing the number of hidden neurons to 512, which should not compromise performance but potentially decreases the processing time by an order of magnitude: As reported in [17], relatively small nets can be used (e.g., with only 256 hidden neurons) while still maintaining low phoneme classification errors on TIMIT. Hence, we believe it is worth to further explore this application in performance monitoring.

5. SUMMARY

Both M-Measure and MaP appear to be suitable for performance monitoring and to generalize over different acoustic scenarios. In terms of correlation with WER, we obtained better results for M-Measure in anechoic situations, while MaP should be preferred in the reverberant condition. This was consistently observed for a full modified test set from Aurora 4 as well as for a single utterance. Since the calculation of measures is rather different, we believe there is room for improvement by combining them to exploit their potential complementarity.

We also investigated if performance measures should be applied to posteriorgrams that represent context-dependent triphones or rather monophones, where the latter can be derived by grouping the according triphones. Very similar correlations were obtained for M-Measure in the anechoic case ($r = 0.96$), but in the office condition better results were achieved with triphone activations. This comparison was restricted to the M-Measure, since MaP used matched filters which were derived from TIMIT monophone labels. In future research, it should be tested if a similar benefit is obtained for MaP when using triphone posteriorgrams. This could be achieved by learning phoneme trajectories from clean DNN output rather than from labels.

In this work, we used signals from behind-the-ear hearing aids, which is an unusual setting for ASR but is useful for assistive technologies, e.g., for providing transcripts for the hearing impaired. It also motivated us to estimate if a standard DNN forward pass could be done on current hearing aid hardware. This was not the case, but rather simple and straight-forward modifications to our setup would allow to do the processing in real-time.

Our findings should be useful for multi-microphone ASR for distant speech recognition in general: Since the distance of microphones considered here is very small, the effects of beamforming or channel selection are limited. Still, it was shown that our performance measures capture the effect of spatial filtering. Devices for home automation typically exhibit several microphones with a larger distance (e.g., arranged in a circular array), in which beamforming has a larger effect and hence should be reflected by these measures as well.

6. ACKNOWLEDGEMENTS

This work was funded by Google via a Google faculty award to Hynek Hermansky, by the DFG (Research Unit FOR 1732 “Individualized Hearing Acoustics” and the Cluster of Excellence 1077/1‘ “Hearing4all”). We thank Constantin Spille for valuable discussions and providing the illustration shown in the first figure.

7. REFERENCES

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

[2] Bitzer, J., Simmer, K. U. (2001). "Superdirective Microphone Arrays," in *Microphone Arrays*, Brandstein, M., Ward, D., Eds., Springer, 2001, pp. 1021–1042.

[3] Cox, H., Zeskind, R., Owen, M. (1987). "Robust adaptive beamforming," IEEE Trans. Acoust. Speech Signal Process., vol. 35, no. 10, pp. 1365–1376, 1987.

[4] Davis, S. and Mermelstein, P. (1980). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Transactions on Acoustics, Speech, and Signal Processing, 28 (4), pp. 357-366.

[5] Exter, M., Meyer, B.T. (2016). "DNN-based automatic speech recognition as a model for human phoneme perception," in Proc. Interspeech.

[6] Gerlach, L., Payá-Vayá, G., and Blume, H. (2016). "Efficient Emulation of Floating-Point Arithmetic on Fixed-Point SIMD Processors," in Proc. IEEE Workshop on Signal Processing Systems (SiPS).

[7] 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.

[8] 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.

[9] Jansen, A. and Niyogi, P. (2009) "Point Process Models for Spotting Keywords in Continuous Speech", IEEE Trans. Audio, Speech and Language Proc., 17(8):1457-1470, 2009.

[10] Kayser, H., Ewert, S. D., Anemüller, J., Rohdenburg, T., Hohmann, V., Kollmeier, B. (2009). "Database of Multichannel In-Ear and Behind-the-Ear Head-Related and Binaural Room Impulse Responses," EURASIP Journal on Advances in Signal Processing, 2009.

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

[12] Krämer, S. (2015). "Implementierung und Evaluation eines Spracherkennungsalgorithmus auf einem VLIW-SIMD Signalprozessor", Bachelor thesis, Institute of Microelectronic Systems, Leibniz Univertät Hannover.

[13] 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.

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

[15] Meyer, B.T. (2013). "What's the difference? Comparing humans and machines on the Aurora2 speech database," in Proc. Interspeech 2013, pp. 2634-2638.

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

[17] Nagamine, T., Seltzer, M.L., Mesgarani, N. (2016). "On the Role of Nonlinear Transformations in Deep Neural Network Acoustic Models," personal communication, to appear in Proc. Interspeech (2016).

[18] 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.

[19] 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.

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

[21] Roeven, H., Coninx, J., Ade, M. (2004). "CoolFlux DSP - The embedded ultra low power C-programmable DSP core," Proc. Intl. Signal Proc. Conf. (GSPx).

[22] 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

[23] Xiong, F., Meyer, B. T., Moritz, N., Rehr, R., Anemüller, J., Gerkmann, T., Doclo, S., et al. (2015). "Front-end technologies for robust ASR in reverberant environments-spectral enhancement-based dereverberation and auditory modulation filterbank features," EURASIP J. Adv. Signal Process., 70.