Learning Acoustic Features from the Raw Waveform for Automatic Speech Recognition

Tobias Menne, Zoltan Tüske, Ralf Schlüter, Hermann Ney
Human Language Technology and Pattern Recognition, Computer Science Department,
RWTH Aachen University, Aachen, Germany,
Email: {menne, tuske, schlueter, ney}@cs.rwth-aachen.de

Abstract

Automatic speech recognition (ASR) usually is performed by using hand crafted preprocessing, which extracts relevant information from the speech waveform, while reducing the redundancy of the resulting feature vectors. Prominent examples are the Mel frequency cepstral coefficients (MFCCs) or the Gammatone (GT) filter bank, originally designed for the use in Gaussian mixture hidden Markov models. However, the successful introduction of neural network (NN) acoustic models has raised the following question: can preprocessing become part of the acoustic modeling and training, taking unprocessed waveforms as direct input? Recent work shows that indeed a fully connected feed-forward NN, is able to learn the feature extraction as part of the acoustic model to a large extent. Introducing convolutional layers in the first stages of the NN further closed the performance gap to hand crafted preprocessing. Improvements, even for multi-channel speech input, are reported on top of manually designed preprocessing, using large amounts of training data for a proprietary task. In this work, waveform based ASR modeling and training is investigated and analyzed for a publicly available medium sized data set, namely the CHiME-4 data set, which supplies real multichannel noisy data for training and evaluation.

Introduction

For ASR systems based on Gaussian mixture model (GMM)-hidden Markov models (HMMs), manually defined feature extraction was a key part in developing high performing noise-robust acoustic models [1, 2, 3]. With the recent advance of deep neural network (DNN) based acoustic models [4], there have been several attempts to include the feature extraction into the acoustic model. E.g. in [5, 6] Mel-like filters are optimized during training. Recent studies also show, that the complete feature extraction can be learned automatically [7, 8] from the raw time signal. Although thus far large amounts of data are necessary to reach the performance of systems based on cepstral features [9]. Using network architectures, which imitate steps from the standard feature extraction, can somewhat mitigate the amount of data needed [10, 8, 11]. In [12, 13, 14] work on proprietary data has been presented on not only learning the feature extraction from the raw time signal, but also learning the beamforming operation for multi-channel, noise robust ASR. This work has been done on 2000 h of noisy training data. Here we present a single channel system using the raw time signal as input for a multi-channel, noisy ASR dataset, namely the CHiME-4 dataset [15]. This system can be used as a baseline system for future work on multi-channel ASR on the raw time signal using a non proprietary dataset.

The remaining paper is structured as follows. The first section sets the work presented here into the context of prior work. The acoustic model architecture used here to do ASR on the raw time signal is described in the following section, before the experimental setup is described. This is followed by a presentation and discussion of the results of the experiments before a brief conclusion and outlook are given.

Relation to prior work

This work analyses the architecture presented in the companion paper [16], when applied to a small, noisy dataset. Here the CHiME-4 dataset [15] is used. The architecture which is presented in [16] and used here is an extension of the architecture in [8], where the usual max pooling layer is substituted by a second level time convolution, which enables the network to exploit various sampling rates. This work describes a baseline system for future work on ASR on multi-channel raw time signals on a non proprietary dataset, as e.g. presented in [12].

System overview

The following is a review of the system presented in [16].

Figure 1 shows an overview of the system. In a first step a time convolutional layer is used:

$$y_{k,t'} = s_t * h'_{k,t} \overset{\text{FIR}}{=} \sum_{\tau=0}^{N_{\text{TF}}-1} s_{t+\tau-N_{\text{TF}}+1} \cdot h_{k,\tau} \qquad (1)$$

Where s_t is the input signal, $y_{k,t'}$ is the output after applying filter $h'_{k,t}$ of the filterbank. Sub-sampling in time is assumed, e.g. by a factor of 10, such that $t = 10 \cdot t'$. This step is motivated by the time-frequency decomposition (TF) of the signal as e.g. done by the short-time Fourier transform (STFT) with a fixed filterbank [1]. The filterbank has a finite impulse response N_{TF}. The next step extracts the amplitude spectrum from the down-sampled TF filter outputs by envelope detection:

$$x_{i,k,t''} \overset{\text{FIR}}{=} f_2\left(\sum_{\tau=0}^{N_{\text{ENV}}-1} f_1\left(y_{k,t'+\tau-N_{\text{ENV}}+1}\right) \cdot l_{i,\tau}\right) \qquad (2)$$

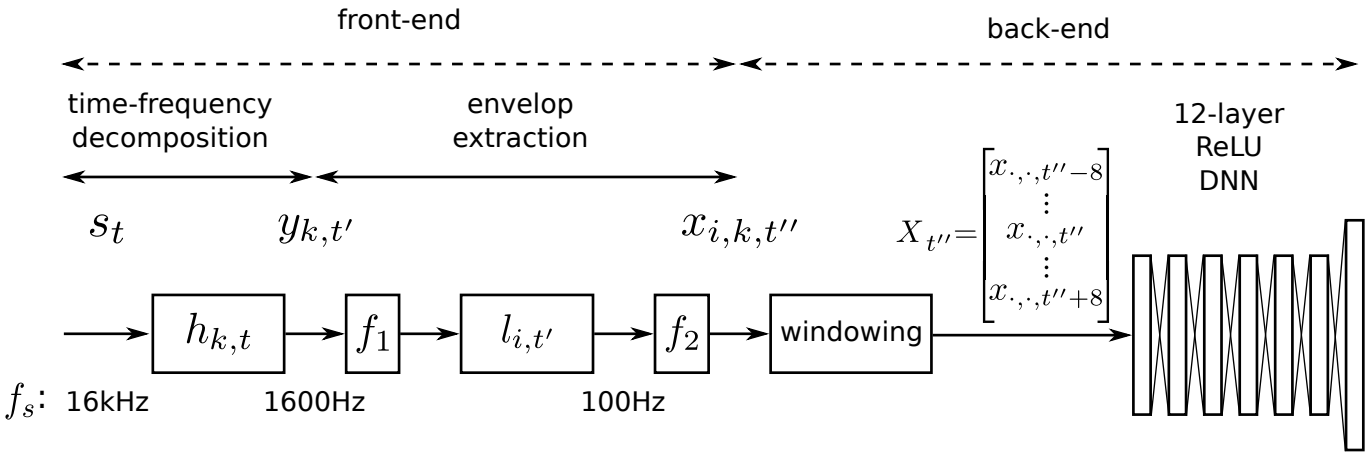

Figure 1: Multi-resolution and convolutional processing of speech signal with neural networks.

Where $l_{i,t'}$ are assumed to be low pass FIR filters with a response of N_{ENV} samples f_1 describes the rectification and is a rectified linear unit (ReLU) [17] or absolute value function in the network architecture. Envelope detection is usually integrated as non-parameterized function in acoustic models, which are working on the raw time signal, e.g. max pooling [18], average [9], p-norm [11]. f_2 describes a logarithmic or root compression, which follows the envelope detection, accounting for the hearing working on a non-linear scale [1, 2, 3]. Multiple filters $l_{i,t'}$ are learned, which are shared not only in time but also between the TF filters. If N_{ENV} and $\max(i)$ are chosen large enough, this allows for multi-resolutional sampling of $f_1(y_{k,t'})$, despite each $x_{i,k,t''}$ having the same sampling rate. After concatenating the extracted feature vectors of neighbouring frames, this network structure is followed by feed-forward ReLU layers. The system is compared to a baseline which uses 16 dimensional mean and variance normalized MFCC features.

Experimental setup

The network architecture described in the previous section is evaluated on the data of the CHiME-4 speech recognition task [15]. The CHiME-4 dataset consists of real and simulated 16 kHz, multi-channel audio data. A microphone array with six channels was arranged around a tabled device, where five of the microphones were facing towards and one microphone was facing away from the speaker. The corpus is based on the 5k WSJ0-Corpus. The real recordings and simulations have been done in four different real-world noise environments. Those environments are on a bus, in a cafe, in a pedestrian area and at a street junction. The training set contains approximately 18 h of data per channel recorded from 87 different speakers.

For the experiments presented here the feature extraction structure described in the system overview section is followed by 12 feed-forward layers, with 2000 units each. The data used for the training of the acoustic model is the audio data of the five microphones facing towards the speaker. Every channel is considered as a separate recording for the training. Furthermore optional fine-tuning of the acoustic model is done by processing the training data with the baseline beamformer (BFIT) [19], which is provided with the CHiME-4 data. The fine-tuning is then done by training the acoustic model on the preprocessed data and using the model trained on the unprocessed data as initialization. The cross-entropy criterion was used for training the acoustic model using stochastic gradient descent with momentum, $l2$ regularization and discriminative pretraining [20]. The sub-sampling was done with $160 \cdot t'' = 10 \cdot t' = t$, for s_t sampled at 16kHz. The TF decomposition was performed with 150 filters with a length of 512 samples. and the $l_{i,t'}$ filters have a length of 40 samples and $\max(i) = 5$. 17 frames were concatenated after the feature extraction structure to increase the context for the computation of a single posterior vector. Results are provided for the real development and real evaluation set of the 6-channel track of the CHiME-4 speech recognition task. The recognition is done on the 5^{th} channel of the development and evaluation data and on the data preprocessed by the BFIT algorithm. Decoding is done with the 5-gram language model provided with the CHiME-4 data. The RASR toolkit [21] was used to carry out the experiments.

Experimental results and discussion

Table 1 shows the results of the experiments on the CHiME-4 dataset for the various subsets of different noise environments. It contains two mismatch scenarios, where either the acoustic model is fine-tuned with the preprocessed training data or the test data is preprocessed, but not both. The results show, that among the mismatch scenarios it is more beneficial to preprocess the test data and use it as input for the system, which is not fine-tuned, than using unprocessed test data as input to the fintuned system. Even more, it can be seen that if the unprocessed test data is used, the word error rate (WER) using the fine-tuned system is generally worse than with the system, which is not fine-tuned. On the other hand, the system, which is not fine-tuned, generally works better on the processed test data. This is despite the mismatch between training and test data. But the results also show, that the system applied to the raw time signal, which is

Table 1: WER (%) of the described system on the CHiME-4 data. The column *Fine-tuned* indicates whether the acoustic model has been fine-tuned with the preprocessed training data. The *Beamformed* column indicates whether the decoding data has been preprocessed by the beamforming algorithm, otherwise the decoding data is the unprocessed channel 5.

Features	Fine-tuned	Beamformed	Dev					Eval				
			Bus	Caf	Ped	Str	**Avg.**	Bus	Caf	Ped	Str	**Avg.**
MFCC	-	-	19.7	14.2	9.7	13.7	**14.3**	30.4	25.2	20.2	16.6	**23.1**
		+	12.2	10.9	8.6	10.6	**10.6**	19.0	16.6	14.7	13.8	**16.0**
	+	-	19.4	13.6	9.1	14.4	**14.1**	32.2	25.7	21.1	17.9	**24.2**
		+	10.4	9.1	6.6	9.2	**8.8**	17.0	14.2	12.4	12.2	**13.9**
raw TS	-	-	30.7	19.9	17.5	21.3	**22.4**	63.6	38.6	34.0	25.0	**40.3**
		+	30.1	15.6	16.1	17.9	**19.9**	49.9	27.4	24.3	23.6	**31.3**
	+	-	27.8	27.3	19.0	24.9	**24.7**	51.9	48.9	39.8	26.1	**41.7**
		+	14.8	11.8	8.8	11.7	**11.8**	29.5	21.5	18.6	16.3	**21.5**

not fine-tuned, can not benefit from the preprocessing of the test data to the same extend as the system using MFCC features.

Both systems generally perform better when the acoustic model is fine-tuned to the preprocessing, when preprocessed test data is used. The results show, that the system applied to the raw time signal benefits more strongly from this adaptation than the MFCC system.

This suggests, that the features learned by the system, at least when trained on the small amount of data used here, are not as robust to a mismatch scenario as the MFCC features.

Generally the system using MFCC features outperforms the system applied to the raw time signal by a relatively large margin. But the results suggest, that the raw time signal system is able to learn usable features even on the small and noisy CHiME-4 dataset. Furthermore the experiments show, that when using a system applied directly to the raw time signal, applying preprocessing both during the training and during decoding has a more significant impact on the performance than for MFCC systems. Therefore it would be interesting to compre the system presented here to a system attempting to also learn the preprocessing of the multi-channel signals directly from the raw time signal.

Conclusion and outlook

In this work we applied the acoustic model architecture proposed in [16] for ASR on the raw time signal to the 18 h, small, noisy CHiME-4 dataset. The results indicate, that a system using MFCC features still outperforms the raw time signal system by a significant margin, when using such a small amount of data. Compared to the MFCC features, the features learned by the network are less robust to different kinds of noises and more sensitive to a mismatch between training and evaluation data preprocessing. Nevertheless the results are a promising baseline to use this non proprietary dataset for multi-channel ASR on the raw time signal, which has thus far only been published on proprietary tasks, e.g. [12].

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program grant agreement No. 694537. This work has also been supported by European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 644283. The work reflects only the authors' views and the European Research Council Executive Agency is not responsible for any use that may be made of the information it contains. The GPU cluster used for the experiments was partially funded by Deutsche Forschungsgemeinschaft (DFG) Grant INST 222/1168-1.

Literatur

[1] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[2] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[3] R. Schlüter, I. Bezrukov, H. Wagner, and H. Ney, "Gammatone features and feature combination for large vocabulary speech recognition," in *ICASSP*, 2007, pp. 649–652.

[4] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, J. Navdeep, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[5] A. Biem, E. McDermott, and S. Katagiri, "A discriminative filter bank model for speech recognition," in *Eurospeech*, 1995, pp. 545–548.

[6] T. N. Sainath, B. Kingsbury, A.-r. Mohamed, and B. Ramabhadran, "Learning filter banks within a deep neural network framework," in *ASRU*, 2013, pp. 297–302.

[7] Z. Tüske, P. Golik, R. Schlüter, and H. Ney, "Acoustic modeling with deep neural networks using raw time signal for LVCSR," in *Interspeech*, 2014, pp. 890–894.

[8] P. Golik, Z. Tüske, R. Schlüter, and H. Ney, "Convolutional neural networks for acoustic modeling of raw time signal in LVCSR," in *Interspeech*, 2015, pp. 26–30.

[9] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *Interspeech*, 2015, pp. 1–5.

[10] D. Palaz, R. Collobert, and M. Magimai.-Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," in *Interspeech*, 2013, pp. 1766–1770.

[11] P. Ghahremani, V. Manohar, D. Povey, and S. Khudanpur, "Acoustic modelling from the signal domain using CNNs," in *Interspeech*, 2016, pp. 3434–3438.

[12] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, M. Bacchiani, and A. Senior, "Speaker location and microphone spacing invariant acoustic modeling from raw multichannel waveforms," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, AZ, Dec 2015, pp. 30–36.

[13] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, and M. Bacchiani, "Factored spatial and spectral multichannel raw waveform cldnns," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar 2016, pp. 5075–5079.

[14] B. Li, T. N. Sainath, R. J. Weiss, K. W. Wilson, and M. Bacchiani, "Neural network adaptive beamforming for robust multichannel speech recognition," in *Proc. Interspeech*, San Francisco, CA, Sep 2016, pp. 1976–1980.

[15] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, Nov 2017.

[16] Z. Tüske, R. Schlüter, and H. Ney, "Acoustic modeling of speech waveform based on multi-resolution neural network signal processing," in *accepted for publication in ICASSP*, Calgary, Canada, Apr. 2018.

[17] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *the 27th Int. Conf. on Machine Learning*, 2010, pp. 807–814.

[18] Y. Hoshen, R. J. Weiss, and K. W. Wilson, "Speech acoustic modeling from raw multichannel waveforms," in *ICASSP*, 2015, pp. 4624–4628.

[19] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[20] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *ASRU*, Hawaii, USA, 2011, pp. 24–29.

[21] D. Rybach, S. Hahn, P. Lehnen, D. Nolden, M. Sundermeyer, Z. Tüske, S. Wiesler, R. Schlüter, and H. Ney, "RASR - the RWTH Aachen University open source speech recognition toolkit," in *ASRU*, 2011.