INSTANTANEOUS DIRECTION OF ARRIVAL FEATURES FOR DNN-BASED SPEECH RECOGNITION IN NOISY AND REVERBERANT ENVIRONMENTS

CHRISTIAN HUEMMER, PHILIPP STADTER

1. Introduction

The accuracy of automatic speech recognizers with DNN-based acoustic models typically deteriorates significantly if reverberation or interfering noise is present in the recording of the target utterance. A variety of techniques has therefore been developed with the aim of alleviating this decrease in performance, some of which, like e.g. beamforming, exploit spatial relations between the different microphones capturing the speaker's voice. Most commonly, these methods realize a signal enhancement preprocessing stage meaning that they suppress the unwanted reverberation and noise components in the recorded signal before the input features for the DNN-based acoustic model are extracted.

It is however a trend in the field of automatic speech recognition (ASR) to replace explicit signal processing steps by implicit learning. A recent example is [7], where the authors proposed to directly incorporate spatial information about the diffuseness of the soundfield into the DNN-based acoustic model. To this end, they presented a method to extract so-called meldiffuseness features from the microphone signals which are then used as additional inputs to the acoustic model DNN. In their experiments, this approach lead to a lower word error rate (WER) than using the diffuseness information for signal enhancement preprocessing.

Following the same idea, we developed a novel type of features based on direction of arrival information. We use them as additional inputs for the acoustic model and call them *instantaneous direction of arrival features* or *IDOA features*. They are extracted in the STFT-domain by comparing the phase differences between the different microphone signals in every time-frequency bin with the phase differences that would result from plane waves coming from the direction of the speaker. Thus, they indicate wether or not the energy within a given Mel-band and within a given time-frame is received predominantly from the direction of the speaker. The extraction of the IDOA features is strongly inspired by [8] and the notion of the instantaneous direction of arrival space. In the form presented in this report, our method relies on a two-channel input signal (while it can easily be extended to any number of channels $M > 1$) and knowledge or estimation of the direction of arrival of the desired speech.

Evaluation of the presented features on the two-channel task of the REVERB challenge [3] shows that the IDOA features are not promising in their current design and are outperformed by the logmelspec acceleration coefficients as additional DNN inputs.

The rest of this report is structured as follows: The relevant aspects of [8] are summarized in Section 2. There, the notion of the IDOA space will be introduced which is fundamental for our new features. In Section 3 the extraction of the IDOA features will be explained in detail. Section 4 contains descriptions of the speech recognition system and the structure and training of the acoustic model DNN. The results of our experimental evaluation are presented in Section 5 and a conclusion is given in Section 6.

2. Instantaneous Direction of Arrival Space

In [8] a method for direction of arrival based time-frequency masking is proposed which is applied to the output of a microphone array's beamformer in order to improve spatial filtering. This is similar to the preprocessing approach in [7] where the spatial diffuseness information is used for time-frequency masking as well. Before we show how we followed the idea of [7] and moved the direction of arrival information into the acoustic model, we go through the relevant theory of [8] and explain how the utilized IDOA information is obtained in the first place.

We denote by $\mathbf{p}_1, \ldots, \mathbf{p}_M$ the positions of the M microphones, where $\mathbf{p}_m = (x_m, y_m, z_m)^\top$ for all $1 \leq m \leq M$. The discretized signal of the m-th microphone is given by $x_m[k]$ with time index k. Its STFT domain representation is $X_m[f, n]$ where f and n index the frequency bin and time frame, respectively. Each microphone m has a known directivity pattern $U_m(f, \mathbf{c})$ where f is the frequency index again and $\mathbf{c} = (\varphi, \theta, \rho)^\top$ is the position of the sound source in a radial coordinate system centered at $\mathbf{p}_m$. For speaker position $\mathbf{c}$, the captured signal of each microphone m is:

$$X_m[f, n] = D_m(f, \mathbf{c})S[f, n] + N_m[f, n]. \tag{1}$$

In that equation, $S[f, n]$ represents the target speech in time-frequency bin (f, n). $D_m(f, \mathbf{c})$ is a frequency response modeling sound propagation to and recording by the m-th microphone. It is given by

$$D_m(f, \mathbf{c}) = \frac{e^{-2\pi \frac{f}{N_{\mathrm{FFT}}-1} F_s \frac{\|\mathbf{c}-\mathbf{p}_m\|}{v}}}{\|\mathbf{c} - \mathbf{p}_m\|} A_m[f] U_m(f, \mathbf{c}), \tag{2}$$

witt FFT-length N_{FFT} and sampling frequency F_s. The first part describes free-space sound wave propagation from an isotropic source with speed of sound v and $A_m[f]$ is the frequency response of the m-th microphone's preamplifier and ADC system. $N_m[f, n]$ represents noise and reverberation picked up by microphone m in time-frequency bin (f, n).

The instantaneous direction of arrival is obtained at every frequency bin f and time frame n based upon the phase differences between microphone signals $X_2[f, n], \ldots, X_M[f, n]$ and the signal $X_1[f, n]$ of the arbitrarily chosen reference microphone 1. This means that the observed instantaneous direction of arrival can be represented as a vector in $M - 1$ dimensional space given by

$$\boldsymbol{\Delta}[f, n] = (\delta_1[f, n], \ldots, \delta_{M-1}[f, n])^\top, \tag{3}$$

where

$$\delta_l[f,n] = \arg(X_1[f,n]) - \arg(X_{l+1}[f,n]) \tag{4}$$

for all $1 \leq l \leq M-1$. The $M-1$ dimensional space spanning all possible IDOAs is called the instantaneous direction of arrival space.

In the absence of noise, the ideal IDOA vector denoted by $\mathbf{\Psi}[f,n]$ would be obtained from Equation (3). This vector corresponds to the direction of arrival of the speaker and will consequently lie inside a $M-1$-dimensional hypervolume within the IDOA space representing all possible speaker positions $\mathbf{c} = (\varphi, \theta, \rho)^\top$. Under the far field assumption for sound propagation this will turn into a $M-1$-dimensional hypersurface since the distance ρ approaches infinity. Linear microphone arrays can distinguish the incident angle in only one dimension. Hence, the possible speaker positions are represented by a hyperline representing all possible values for θ.

Because reverberation and interfering noise will generally not arrive at the microphone array from the same direction as the target speech, the observed phase differences $\mathbf{\Delta}[f,n]$ will deviate from the ideal IDOA vector $\mathbf{\Psi}[f,n]$. The Euclidian norm of the difference between $\mathbf{\Delta}[f,n]$ and $\mathbf{\Psi}[f,n]$ in IDOA space can be seen as an indicator of wether or not target speech is predominant in a time-frequency bin. A large distance means that most energy at that bin arrived at the microphone array from a different direction than the speaker direction. Thus, that bin will mostly contain reverberation or noise. Conversely, a small distance will be observed for bins where speech from the desired direction overweighs unwanted contributions from other directions. In [8], it is proposed to convert the distance in the IDOA space back into a difference in incident angle θ according to

$$\Gamma[f,n] = \frac{\|\mathbf{\Delta}[f,n] - \mathbf{\Psi}[f,n]\|}{\|\frac{\partial \mathbf{\Psi}[f,n]}{\partial \theta}\|}. \tag{5}$$

The authors of [8] proceed to estimate the variances $\lambda[f]$ of $\Gamma[f,n]$ and estimate the probability that a given time-frequency bin (f,n) originates from the target direction by

$$p[f,n] = \frac{1}{\sqrt{2\pi}\lambda[f]} e^{-\frac{\Gamma^2[f,n]}{2\lambda^2[f]}}. \tag{6}$$

Furthermore, they employ a first order Markov process with the two states "speech activity" and "speech pause" and suitable transition probabilities to model temporal characteristics of speech. Finally, an estimate for the probability of speech activity is obtained for every time-frequency bin. See [8] for the details.

We do not follow these additional steps, but derive our IDOA features directly from $\Gamma[f,n]$ as we will show next.

3. Instantaneous Direction of Arrival Features

In our experiments, we extracted the well known logmelspec features as 24-dimensional vectors for each time frame n of the spectrogram $X_1[f,n]$ by first applying the $\log(|\cdot|^2)$ operator to all entries in $X_1[f,n]$ and then using 24 triangular Mel-scaled weighting filters. We normalize the feature vectors per utterance to have zero mean and unit variance. An

example of the logmelspec features for a reverberated and noisy utterance of the two-channel REVERB challenge corpus [3] can be seen in Figure 1. Figure 1a shows the spectrogram of the clean speech signal. In Figure 1b, the logmelspec features extracted from a noisy and reverberated version of this utterance are shown.

The extraction of our IDOA features from $\Gamma[f,n]$ is almost completely analogous to the extraction of the logmelspec features form $X_1[f,n]$. The same 24 triangular Mel-scaled weighting filters are applied to each time frame of $\log(|\Gamma[f,n]|)$ to obtain 24-dimensional feature vectors. As before, mean and variance normalization is applied to these features. The only minor difference is, that we reduce the variation of the values in $\Gamma[f,n]$ before applying the Mel filters. To this end, we determine an upper and lower threshold value, such that 90% of the entries in $\Gamma[f,n]$ are smaller than the upper threshold and 65% are larger than the lower threshold. We then replace all values exceeding the upper threshold by the upper threshold value and replace all values falling below the lower threshold by the lower threshold value. We show an example of the IDOA features extracted from a reverberated and noisy speech signal in Figure 1c. Under careful comparison with the clean signal spectrogram in Figure 1a, it can be observed that the darkest structures in the IDOA features (indicating lowest distance in IDOA space) coincide with areas of high speech energy. However, overall these relations appear to be rather vague and the IDOA features exhibit an almost random-like character. Presumably, this will limit the extend to which the acoustic model DNN can exploit the information provided by the IDOA features.

In the following, we describe exactly how we interpreted Equations (1), (2), (3), (4) and (5) when extracting the IDOA features for our experiments. First, note that since the number of microphones is $M = 2$ the IDOA space is 1-dimensional. Thus, $\mathbf{\Delta}[f,n]$ and $\mathbf{\Psi}[f,n]$ are 1-dimensional, too. We neglect the influence of the microphones' preamplifier and ADC systems as well as their directivity patterns. Furthermore, we do not consider sound propagation loss. Consequently, we set $A_m[f] = 1$ and $U_m(f,\mathbf{c}) = 1$ and $\|\mathbf{c}-\mathbf{p}_m\| = 1$ in the denominator of Equation (2), yielding

$$D_m(f,\mathbf{c}) = e^{-2\pi \frac{f}{N_{\mathrm{FFT}}-1} F_s \frac{\|\mathbf{c}-\mathbf{p}_m\|}{v}}. \tag{7}$$

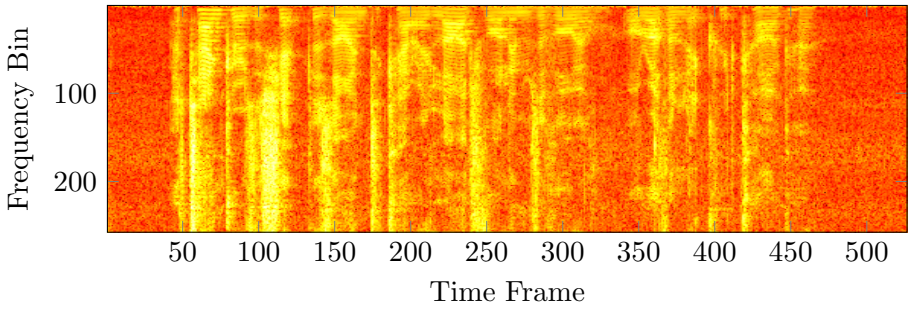
The observed IDOA vectors $\mathbf{\Delta}[f,n]$ are obtained according to Equations (3) and (4):

$$\mathbf{\Delta}[f,n] = \delta_1[f,n] = \arg(X_1[f,n]) - \arg(X_2[f,n]) = \arg(X_1[f,n]X_2^*[f,n]). \tag{8}$$

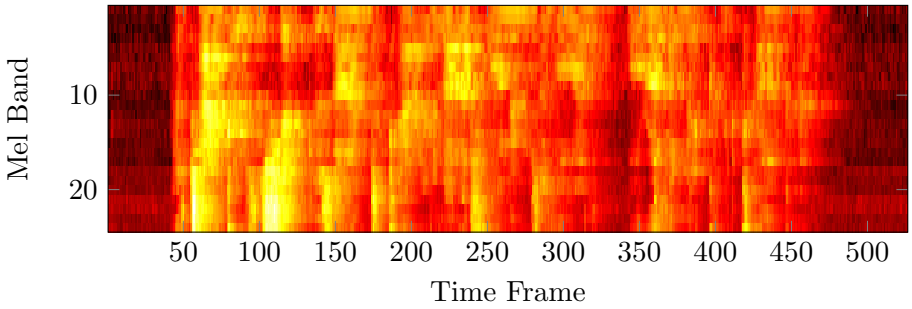
Let θ denote the angle between the line connecting the two microphones and the direction of the plane sound wave coming from the speaker. The distance between the microphones is d. With Equation (7), this gives $D_1(f,\mathbf{c}) = e^{-2\pi \frac{f}{N_{\mathrm{FFT}}-1} F_s \frac{\|\mathbf{c}-\mathbf{p_1}\|}{v}}$ and $D_2(f,\mathbf{c}) = e^{-2\pi \frac{f}{N_{\mathrm{FFT}}-1} F_s \frac{\|\mathbf{c}-\mathbf{p_1}\|+\cos(\theta)d}{v}}$. In the absence of noise, this turns Equation (1) into $X_1[f,n] = e^{-2\pi \frac{f}{N_{\mathrm{FFT}}-1} F_s \frac{\|\mathbf{c}-\mathbf{p_1}\|}{v}} S[f,n]$ and $X_2[f,n] = e^{-2\pi \frac{f}{N_{\mathrm{FFT}}-1} F_s \frac{\|\mathbf{c}-\mathbf{p_1}\|+\cos(\theta)d}{v}} S[f,n]$. The ideal IDOA vectors $\mathbf{\Psi}[f,n]$, which result from Equation (8) when no noise is present, are therefore obtained as:

$$\mathbf{\Psi}[f,n] = \arg(X_1[f,n]X_2^*[f,n]) \tag{9}$$

(A) Spectrogram of the clean speech signal.



(B) Logmelspec features extracted from the reverberated speech signal.



(C) IDOA features extracted from the reverberated speech signal.

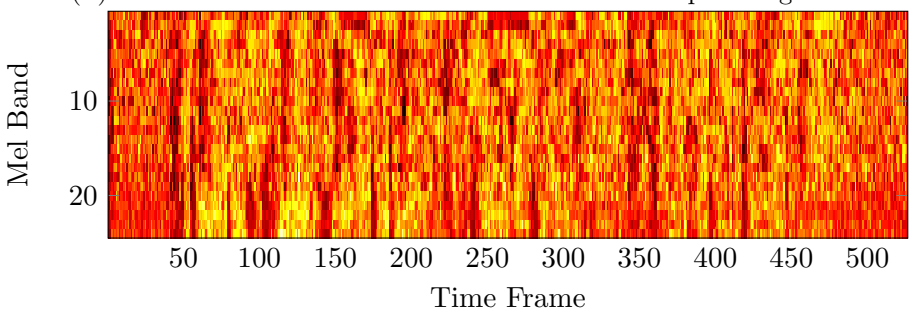


FIGURE 1. Spectrogram of the clean speech signal as well as logmelspec features and IDOA features extracted from the reverberated speech signal for an utterance of the two-channel REVERB challenge corpus [3]. The reverberated and noisy utterance belongs to the "SimData" part of the REVERB corpus and was created using the impulse response of room 3 and source-microphone spacing "far" (see also Section 4).

$$= \arg(e^{-2\pi \frac{f}{N_{\mathrm{FFT}}-1} F_s \frac{\|\mathbf{c}-\mathbf{p_1}\|}{v}} e^{+2\pi \frac{f}{N_{\mathrm{FFT}}-1} F_s \frac{\|\mathbf{c}-\mathbf{p_1}\|+\cos(\theta)d}{v}}) \tag{10}$$

$$= 2\pi \frac{f}{N_{\mathrm{FFT}}-1} F_s \cos(\theta) \frac{d}{v}. \tag{11}$$

Inserting (8) and (11) into Equation (5), we finally arrive at

$$\Gamma[f,n] = \frac{|\arg(X_1[f,n]X_2^*[f,n]) - 2\pi\frac{f}{N_{\mathrm{FFT}}-1}F_s\cos(\theta)\frac{d}{v}|}{|2\pi\frac{f}{N_{\mathrm{FFT}}-1}F_s\sin(\theta)\frac{d}{v}|}. \tag{12}$$

For estimation of the speaker's direction of arrival, we employed a SRP-PHAT implementation [10, 9].

4. DNN-Based Speech Recognition

We use the same ASR system with DNN-based acoustic model as in [2]. We also used the same training method for the DNN and the same training and evaluation data sets. Therefore, some of the descriptions in [2] are directly cited in the following.

The 24 logmelspec features are extended by their 24 delta coefficients (Δ) and depending on the specification in Table 1 by their 24 acceleration coefficients ($\Delta\Delta$) or by the 24 IDOA features. Additionally, context extension is realized using ± 5 frame splicing.

"These extended logmelspec feature vectors of length 792 are used as input of the DNN which is characterized by the following topology:

- 6 hidden layers, each with 2048 nodes and sigmoid activation functions.
- Output layer with softmax nonlinearity and 3463 elements (number of context-dependent HMM states).

" [2]

"We employ the Kaldi toolkit [4] as ASR back-end system using the WSJ0 trigram 5k language mode provided by the REVERB challenge [3]. As first step, we train a GMM-HMM baseline system on the clean WSJCAM0 Cambridge Read News REVERB corpus [6] with feature extraction following the Type-I creation in [5] (which is state-of-the art in the Kaldi recipe): The extraction of 13 mel-frequency cepstral coefficients (MFCCs) is followed by linear discriminant analysis (with splicing optimized to ± 4 input frames), maximum likelihood linear transform and feature-space maximum likelihood linear regression. The state-frame alignment of the trained GMM-HMM baseline system is employed for training the DNN on extended logmelspec feature vectors [...]: A generative pretraining using the contrastive divergence algorithm (on restricted Boltzmann machines) is followed by discriminative fine-tuning using the mini-batch stochastic gradient descent approach (based on the cross-entropy criterion) [1]." [2] The DNN is trained using the the multi-condition training set (of 7861 utterances) provided by the REVERB challenge [3].

"The evaluation of the reverberation-robust DNN-HMM hybrid system [...] is realized using the two-channel task of the REVERB challenge [3]: The evaluation test set consists of about 5000 environmentally-distorted utterances and is split into two categories: First, the utterances of the clean WSJCAM0 Cambridge Read News REVERB corpus are artificially corrupted ("SimData") using measured impulse responses ($T_{60} \approx 0.25$ s, 0.5 s and 0.7 s), recorded noise sequences (added to the microphones signals with a signal-to-noise ratio of 20 dB) and source-microphone spacings of 0.5 m ("Near") and 2 m ("Far"). Second, multichannel recordings ("RealData") in a reverberant ($T_{60} \approx 0.7$ s) and noisy environment are considered with source-microphone spacings of 1 m ("Near") and 2.5 m ("Far"). In

Features	SimData						RealData	
	Room 1		Room 2		Room 3		Room1	
	near	far	near	far	near	far	near	far
logmelspec+Δ+ΔΔ	5.93	6.23	6.74	10.75	7.28	12.28	21.02	21.27
logmelspec+Δ+IDOA	5.86	6.88	6.91	10.51	7.74	13.27	24.47	26.74

TABLE 1. ASR Word Error Rate for the REVERB challenge evaluation test set.

both cases, two microphones with a spacing of 8 cm are selected out of an 8-channel circular microphone array to realize the two-channel task which is evaluated in the following." [2]

5. Evaluation Results

The evaluation results of our new features are presented in Table 1. Unfortunately, the IDOA features cannot outperform the commonly used and easily extracted acceleration coefficients of the logmelspec features. Particularly for the "RealData", the recognition accuracy with IDOA features is significantly inferior to the accuracy with acceleration coefficients. Some considerations as to why this might be the case have already been made in Section 3.

6. Conclusion

In this report, we have presented a novel type of features for DNN-based ASR. They aim at exploiting direction of arrival information of a microphone array in the acoustic model through learning. However, the results of our evaluation show that with their current design our IDOA features are outperformed by the standard logmelspec acceleration coefficients.

References

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, Nov 2012.

[2] C. Hmmer, R. Maas, A. Schwarz, R. F. Astudillo, and W. Kellermann. Uncertainty decoding for dnn-hmm hybrid systems based on numerical sampling. In *Interspeech 2015*, pages 3556–3560, September 2015.

[3] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas. The reverb challenge: Acommon evaluation framework for dereverberation and recognition of reverberant speech. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*, pages 1–4, Oct 2013.

[4] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011. IEEE Catalog No.: CFP11SRW-USB.

[5] S. P. Rath, D. Povey, K. Vesel, and J. Cernock. Improved feature processing for deep neural networks. In F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, L. Lamel, F. Pellegrino, and P. Perrier, editors, *INTERSPEECH*, pages 109–113. ISCA, 2013.

[6] T. Robinson, J. Fransen, D. Pye, J. Foote, S. Renals, P. Woodland, and S. Young. Wsjcam0 cambridge read news for reverb ldc2013e109. Web Download, 2013.

[7] A. Schwarz, C. Huemmer, R. Maas, and W. Kellermann. Spatial diffuseness features for dnn-based speech recognition in noisy and reverberant environments. *CoRR*, abs/1410.2479, 2014.

[8] I. Tashev and A. Acero. Microphone array post-processor using instantaneous direction of arrival. In *Proceedings of International Workshop on Acoustic, Echo and Noise Control IWAENC 2006*, Paris, France, September 2006.

[9] H. L. Van Trees. *Detection, estimation, and modulation theory. Part IV. , Optimum array processing.* Wiley-Interscience, New York, 2002.

[10] C. Zhang, D. Florencio, and Z. Zhang. Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings. *IEEE Transactions on Multimedia*, April 2008.