Channel mapping using bidirectional long short-term memory for dereverberation in hands-free voice controlled devices

Zixing Zhang, Joel Pinto, Christian Plahl, Björn Schuller, Daniel Willett

Angaben zur Veröffentlichung / Publication details:

Zhang, Zixing, Joel Pinto, Christian Plahl, Björn Schuller, and Daniel Willett. 2014. "Channel mapping using bidirectional long short-term memory for dereverberation in hands-free voice controlled devices." *IEEE Transactions on Consumer Electronics* 60 (3): 525–33. https://doi.org/10.1109/tce.2014.6937339.

Channel Mapping using Bidirectional Long Short-Term Memory for Dereverberation in Hands-Free Voice Controlled Devices

Zixing Zhang, Joel Pinto, Christian Plahl, Björn Schuller, *Member*, IEEE, Daniel Willett

Abstract — *In this article, the reverberation problem for hands-free voice controlled devices is addressed by employing Bidirectional Long Short-Term Memory (BLSTM) recurrent neural networks. Such networks use memory blocks in the hidden units, enabling them to exploit a self-learnt amount of temporal context. The main objective of this technique is to minimize the mismatch between the distant talk (reverberant/distorted) speech and the close talk (clean) speech. To achieve this, the network is trained by mapping the cepstral feature space from the distant talk channel to its counterpart from the close talk channel frame-wisely in terms of regression. The method has been successfully evaluated on a realistically recorded reverberant French corpus by a large scale of experiments of comparing a variety of network architectures, investigating different network training targets (differential or absolute), and combining with common adaptation techniques. In addition, the robustness of this technique is also accessed by cross-room evaluation on both, a simulated French corpus and a realistic English corpus. Experimental results show that the proposed novel BLSTM dereverberation models trained by the differential targets reduce the word error rate (WER) by 16% relatively on the French corpus (intra room scenario) as well as 8% relatively on the English corpus (inter room scenario)*[1].

Index Terms — Hand-Free Voiced Controlled Devices, Bidirectional Long Short-Term Memory, Indirect Feature Enhancement, Dereverberation.

I. Introduction

Human computer interaction via voice is increasingly being used and accepted in consumer electronics because of the advantages of hands-free operation: simplicity, mobility, customizability, etc. For some personal computing devices such as notebooks and smart phones, the user is close to the microphones due to the inherent nature of these devices and application (e.g., personal assistant). In many other applications such as digital television [1], set-top boxes, home automation [2], car navigation system [3], and human robot interaction, the ultimate user experience is the ability to communicate hands free from a distance typically a few meters. In this case, however, the distant controlled speech can underlie significant distortion due to room reverberation, echo from loud speaker, and additive noise sources, which leads to high word error rate of speech recognition, and consequently results in poor user experience.

Reverberation is an undesired acoustic phenomenon in the context of speech recognition where the speech signal from the user reaches the microphone with different time delays and amplitude attenuations, caused by the reflection of various surfaces in the acoustic enclosure such as a living room. The speech signal acquired by a microphone is a sum of three components: (a) the direct path signal whose power is inversely proportional to the square of the distance from the speaker [4]; (b) the early reflections from the walls, floor, ceiling, etc., and these depend on the position of the speaker; (c) the late reverberation which depends mainly on the size of the room and reflective properties of the room surface. This is considered to be less dependent on the position of the speaker [4], [5].

In the past decades, extensive research has been carried out to handle such harmful effects. Based on what is addressed, they can broadly be sorted into three categories: signal, feature, and model-based approaches. The *signal*-based approaches are to enhance the reverberant signal from temporal or spectral information. Typical methods include blind deconvolution by inverse filtering [6], beamforming (e.g., delay-and-sum method) which is based on multi-microphones [5], etc. The *feature*-based approaches attempt to remove the influence of reverberation directly from the corrupted feature vectors. Well-known techniques involve feature normalization like cepstral mean normalization (CMN), which is effective for mitigating early reverberation [7], exacting expert crafted features like RASTA-PLP [8], and so on. Both signal- and feature based approaches are located in the front-end of ASR system according to ETSI standard ES 202 212. The *model*-based approaches are applied in the

[1] The research leading to these results is sponsored by Nuance Communications, Inc., where Zixing Zhang pursued his internship from August 2013 to December 2013.

Z. Zhang is with the Machine Intelligence & Signal Processing Group, Institute for Human-Machine Communication, Technische Universität München, München, 80333, Germany (e-mail: zixing.zhang@tum.de).

J. Pinto is with the Nuance Communications, Inc., Aachen, 52072, Germany (e-mail: joel.pinto@nuance.com).

C. Plahl is with the Nuance Communications, Inc., Aachen, 52072, Germany (e-mail: christian.plahl@nuance.com).

B. Schuller with the Machine Intelligence & Signal Processing Group, Institute for Human-Machine Communication, Technische Universität München, München, 80333, Germany (e-mail: schuller@tum.de). He is also with the Department of Computing, Imperial College London, London, SW7 2AZ, United Kingdom (e-mail: bjoern.schuller@imperial.ac.uk).

D. Willett is with the Nuance Communications, Inc., Aachen, 52072, Germany (e-mail: daniel.willett@nuance.com).

back-end of an ASR system, and adjusting the parameters of the acoustic model to the statistical properties of reverberant feature vectors or tailoring the decoder to the reverberant feature vectors. One or more adaptation techniques are applied, for example, maximum a posteriori (MAP) [9], maximum likelihood linear transformation (MLLR) [10], and feature-space MLLR (fMLLR or CMLLR) [11], to reduce the mismatch of Hidden Markov Models (HMMs) trained on clean speech and reverberant speech.

In the recent past, a prominent technique is to train *deep neural networks* (DNNs) [12] using a wide variety of reverberated data sources. The key objective is to derive the original speech features to a high level representation. Its potential capability for noise robust automatic speech recognition (ASR) has been demonstrated previously [13], [14]. Another approach which has lately received increasing attention is to use neural networks for feature enhancement, which aims to remove the reverberation characteristic information from the distant talk speech on the means of learning a mapping rule from the distant talk feature space to its close talk counterpart. The main advantage of this approach is that it leaves the feature extraction and the back-end untouched, as the mapping is performed after feature extraction and prior to decoding. Therefore, the technique can be easily integrated with any existing ASR systems. This work was firstly realized by employing a *multi-layer perceptron* (MLP) via mapping multiple channel array speech to clean speech. Then, it was extended by using *recurrent neural networks* (RNNs) [16] for the 2nd CHiME challenge [17], where reduction in word error rates were observed.

Long short-term memory recurrent neural networks (LSTM-RNNs) [18], a more sophisticated form of RNNs, nowadays have been successfully applied to a variety of pattern recognition tasks, especially to sequential pattern tasks, i.e., handwriting recognition [19], continuous speech recognition [20], and driver distraction detection [21]. Compared with 'classic' RNNs, LSTM neural networks adopt memory blocks to replace the individual artificial neurons. Therefore, these networks can learn an optimized range of contextual information, aiming at overcoming the vanishing gradient problem of conventional RNNs [18], [22]. The superiority of LSTM neural networks (especially the bidirectional type of BLSTM) when compared to DNNs and conventional RNNs have been empirically confirmed in several recent comparative studies [20], [23]. Moreover, in 2013 the effectiveness of LSTM networks to handle nonstationary noisy speech was first demonstrated [24] and later extended to enhance reverberated noisy speech [25].

In this paper, the BLSTM-RNNs are explored to learn the nonlinear feature mapping rule. In comparison with the work done previously [25], this work contributes to (1) evaluating the BLSTM dereverberation approach by executing extensive experiments on realistic and synthesized reverberated speech, and comparing the approach with other traditional network structures like MLP and (B)RNN in order to exploit the potential value of memory networks; (2) proposing the *differential* feature vectors between the distant talk (reverberant/distorted) speech and close talk (clean) speech as training targets, which differs with the previous work [25] where only the *absolute* feature vectors of close talk speech are adopted as training targets; (3) comparing and integrating our feature enhancement methods with the widely used adaptation algorithms like MLLR and CMLLR; and (4) accessing the robustness of the techniques in the scenarios of mismatched recording environments between training and evaluation sets.

The remainder of this paper is organized as follows. Section II describes a framework of a feature dereverberation system by neural networks, which are trained by either absolute or differential targets as given successively. Then, the details of BLSTM structure are presented in Section III. Section IV mainly focuses on investigating the effectiveness of our methods by conducting a large-scale experiments in various scenarios, after a short description of our databases and experimental setups. Finally, conclusions are drawn and possible future directions are pointed out in Section V.

II. FEATURE DEREVERBERATION BY NEURAL NETWORK

A. System Overview

The framework of BLSTM models for dereverberation in distant talk ASR is illustrated in Fig. II-A. The clean talk signal $s(t)$ is corrupted by convolutional noise $r(t)$ and additive noise $n(t)$ when transmitting through space channel. So, the observed distant talk signal $\hat{s}(t)$ at the microphone can be written as:

$$\hat{s}(t) = s(t) * r(t) + n(t). \tag{1}$$

For the sake of simplification, additive noise is ignored in this article. Thus, equation (1) becomes

$$\hat{s}(t) = s(t) * r(t). \tag{2}$$

The total length of RIR can be denoted as T60 which represents the time taken for the energy in the impulse response to decay by 60dB compared to the direct sound. The RIR $r(t)$ can be divided into two portions: The early reflection $r_e(t)$ that includes several strong reflections, and the late reverberation $r_l(t)$ that consists of a series of numerous indistinguishable reverberation. This is,

$$r(t) = r_e(t) + r_l(t), \tag{3}$$

where

$$r_e(t) = \begin{cases} r(t) & 0 \le t \le T \\ 0 & \text{otherwise,} \end{cases} \quad r_l(t) = \begin{cases} r(t+T) & 0 \le T \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

and T is the length of the spectral analysis window (20-30 ms). Thus, equation (2) can be changed into

$$\hat{s}(t) = s(t) * r_e(t) + s(t-T) * r_l(t). \tag{5}$$

When the length of RIR T60 is much shorter than the analysis window size T, $r(t)$ is equal to $r_e(t)$, which only affects the speech affects the speech signals within a frame (analysis window). This linear distortion in the spectral

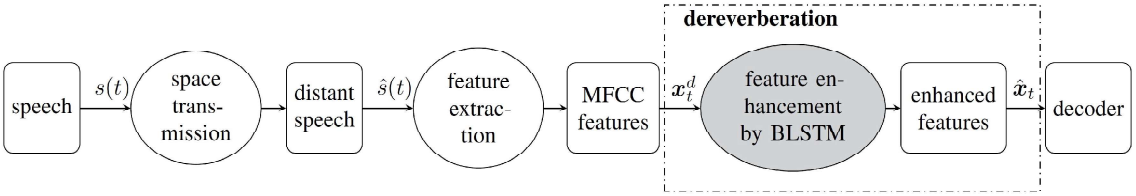

Fig. 1. Framework of BLSTM models for dereverberation in distant talk ASR.

domain can be effectively mitigated by conventional techniques like CMN [7]. For most applications (e.g., occurring in typical office and home environment), however, the reverberation time T60 ranges from 200 to 1000 ms [26] that is much longer than the analysis window size, resulting in an undesirable influence on the following speech frames. For example, if the duration of a RIR is 1 s (T60) and a feature frame is extracted every 10 ms, one RIR would smear across the following 100 frames. Therefore, this distorted speech, after applying short-time discrete Fourier transform (STDFT), can be formulated by:

$$\hat{S}(t,f) = S(t,f)R_e(t,f) + \sum_{d=1}^{D-1} S(t-d,f)R_l(t-d,f), \tag{6}$$

where $R(d,f)$ denotes the part of $R(f)$ (i.e., STDFT of RIR $r(t)$) corresponding to frame delay d. In this case, the channel distortion is no more of multiplicative nature in a linear spectral domain – rather it is convolutional.

Assuming the phases of different frames are non-correlated for simplification, the power spectrum of (6) can be approximated as

$$\begin{aligned} |\hat{S}(t,f)|^2 \approx & |S(t,f)|^2 |R_e(t,f)|^2 + \\ & \sum_{d=1}^{D-1} |S(t-d,f)|^2 |R_l(t-d,f)|^2 . \end{aligned} \tag{7}$$

To extract the standardized feature vectors in cepstral domain for ASR, logarithms and discrete cosine transform (DCT) are executed over the above spectral signals. So,

$$\begin{aligned} \mathcal{D}(ln|\hat{S}(t,f)|)^2 \approx & \mathcal{D}(ln|S(t,f)|^2) + \mathcal{D}(ln|R_e(t,f)|^2) \\ & + \mathcal{D}(ln|M(t,f)|^2), \end{aligned} \tag{8}$$

where $\mathcal{D}$ denotes the discrete cosine transformation matrix, and

$$\begin{aligned} |M(t,f)|^2 &= 1 + \frac{\sum_{d=1}^{D-1} |S(t-d,f)|^2 |R_l(t-d,f)|^2}{|S(t,f)|^2 |R_e(t,f)|^2} \\ &= \frac{|\hat{S}(t,f)|^2}{|S(t,f)|^2 |R_e(t,f)|^2}. \end{aligned} \tag{9}$$

If the speech signal transmission channel is invariable within the sentence period, the second term of $\mathcal{D}(ln|R_e(t,f)|^2)$ in (8) can be treated as a constant, and can be theoretically removed just by subtracting the cepstral mean over each utterance [7]. Therefore, the objective of our strategy is to get rid of the third term of $\mathcal{D}(ln|M(t,f)|^2)$ which is the proportion of the power spectrum of the whole observed distorted speech and the distorted speech only convoluted by early reverberation (cf. (9)). The specific way to realize such a strategy in this article is to apply neural networks to map the feature vectors $\boldsymbol{x}_t^d$ that are extracted from the distant talk speech signals $\hat{s}(t)$ to the target ones frame by frame. Finally, the enhanced feature vectors $\hat{\boldsymbol{x}}_t$ will be fed into the ASR decoder.

B. Differential vs. Absolute Targets for Training Neural Networks

From (8) and (9), one can observe that the term of $M(t,f)$ is not only relative to the early reflection, but also convoluted to the late reverberation of previous speech signals. Such highly nonlinear and nonstationary characteristic makes dereverberation an extremely challenging task [5], [26]. To this end, using a *nonlinear* system to predict this term might be a potentially promising approach. On the other hand, the close relationship of $M(t,f)$ with the numerous previous speech frames also implies the possibility of compensating for the late reverberation by leveraging the *long-term acoustic context*. That is, to exploit the sequence of reverberant feature vectors preceding the current ones might be also beneficial for mitigating the late reverberation. The traditional way to capture such contextual information is to use triphone HMMs, which is empirically proved not sufficient for this task [17].

Motivated by these analyses, an approach is explored based on a nonlinear and more efficient context-learning-ability neural network [18] – BLSTM-RNN – to remove such convoluted late reverberation in the cepstral domain. More specifically, two ways could be applied according to (8) via transforming the distorted feature vectors $\boldsymbol{x}_t^d$ from the distant speech signal $\hat{s}(t)$ into:

1) the corresponding *absolute* (clean) ones $\boldsymbol{x}_t^c$ from close talk speech signals $s(t)$ by minimizing the following objective function of the mean squared error (MSE):

$$J(\theta) = \sum_{n=1}^{N} (x_t^c - \hat{x}_t^c)^2, \tag{10}$$

where $\hat{x}_t^c$ is the predicted close talk feature, and N is the dimensionality of the feature vector. This *direct* channel mapping strategy has already been investigated previously [24], [25].

2) the corresponding *differential* (delta) ones $\boldsymbol{x}_t^\Delta$ which are obtained from later reverberation of $M(t,f)$ (cf. (9)). Before training the neural network, the differential vectors are calculated by subtracting the feature vectors of distant talk $\boldsymbol{x}_t^{\mathrm{d}}$ from those of the corresponding close talk $\boldsymbol{x}_t^c$. When training the neural networks, the parameters are optimized by minimizing:

$$J(\theta) = \sum_{n=1}^{N} (x_t^\Delta - \hat{x}_t^\Delta)^2, \tag{11}$$

where $\hat{x}_t^{\Delta}$ is the predicted differential feature. After that, these mapped differential vectors are added to the original distant talk feature vectors $\boldsymbol{x}_t^d$ frame by frame, so as to compensate the distortion by reverberation. This *indirect* channel mapping strategy is firstly proposed and investigated in this work.

III. Bidirectional Long Short-Term Memory Neural Network

As discussed in Section II, a nonlinear system with the capability of learning long-term contextual information is preferred to tackle with the nonlinear, nonstationary, and highly convoluted late reverberation. The conventional MLP propagates the input signals unidirectionally layer-by-layer with sigmoid activations without any recurrent connection, and needs to stack several successive feature vectors as input. Nevertheless, the capability of capturing context information is still limited by the chosen context [27]. Another method to address this problem is to employ RNNs, where the output of a previous time step is looped back and used as additional input. However, research shows that standard RNNs can not access long-range context since the backpropagated error either blows up or decays over time (the *vanishing gradient problem*) [22].

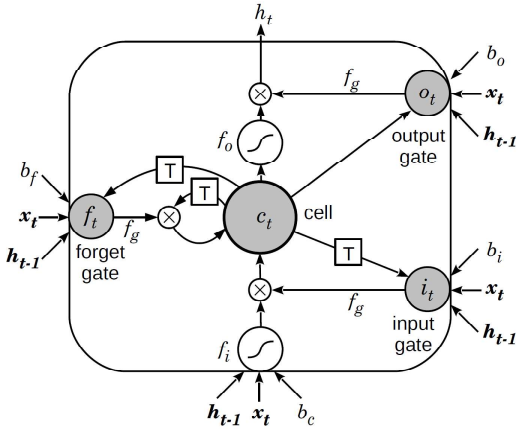

Fig. 2. LSTM memory block. The symbols f_g, f_i, and f_o denote logistic sigmoid, tanh, and tanh activation functions, respectively; i_t, o_t, and f_t are the activations of the input, output, and forget gates at time *t*, respectively; x_t, h_t, and c_t represent input, output, and cell values of the memory block at time *t*, respectively; b is a bias.

To overcome this limitation, [18] introduced LSTM networks, which are able to store information in memory cells over a long period of time. LSTM networks can be interpreted as RNNs in which the traditional neurons are replaced by scaled *memory blocks* (shown in Fig. 2). Similar to the cyclic connections in RNNs, these memory blocks are recurrently connected. Every memory block consists of self-connected linear memory cells and three multiplicative gate units: *input*, *output*, and *forget* gate. The input and output gates scale the input and output of the cell while the forget gate scales the internal state. In other words, the three gates are responsible for writing, reading, and resetting the memory cell values, respectively. For example, if the forget gate is open and the input gate is closed (i.e., the input gate activation is close to zero), the activation of the cell will not be overwritten by new inputs, and therefore the information from previous time *t* can be accessed at the following arbitrary time steps by opening the output gate. (Please refer to [18] and [19] for more details.)

In particular, for a memory block, the activation of the input gate i_t is composed of four components:

$$i_t = f_g(\boldsymbol{W}_{xi}\boldsymbol{x}_t + \boldsymbol{W}_{hi}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{ci}\boldsymbol{c}_{t-1} + b_i), \tag{12}$$

where f_g denotes the logistic sigmoid function of the input unit, $\boldsymbol{W}$ is a weight matrix of the connections from all input gates, output gates, or forget gates in the same hidden layer to the input unit, $\boldsymbol{x}_t$ is the input vector, $\boldsymbol{h}_t$ is the hidden vector, and b_i is the unit bias. The activation of the forget gate f_t follows the same principle, and can be written as

$$f_t = f_g(\boldsymbol{W}_{xf}\boldsymbol{x}_t + \boldsymbol{W}_{hf}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{cf}\boldsymbol{c}_{t-1} + b_f). \tag{13}$$

The memory cell value c_t is the sum of the inputs at time step t and its previous time step activations that are multiplied by forget gate activation, and updated by:

$$c_t = i_t \cdot f_i(\boldsymbol{W}_{xc}\boldsymbol{x}_t + \boldsymbol{W}_{hc}\boldsymbol{h}_{t-1} + b_c) + \boldsymbol{f}_t \cdot \boldsymbol{c}_{t-1}, \tag{14}$$

where f_i is the tanh activation function. Finally, the output of the memory cell is controlled by the output gate activations of

$$o_t = f_g(\boldsymbol{W}_{xo}\boldsymbol{x}_t + \boldsymbol{W}_{ho}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{co}\boldsymbol{c}_t + b_o), \tag{15}$$

and delivered by

$$h_t = o_t \cdot f_o(c_t), \tag{16}$$

where f_o is also a tanh activation function.

Note that each memory block can be regarded as a separate, independent unit. Therefore, if each memory block includes one memory cell, the activation vectors $\boldsymbol{i}_t$, $\boldsymbol{o}_t$, $\boldsymbol{f}_t$, and $\boldsymbol{c}_t$ are all of same size as $\boldsymbol{h}_t$, i.e., the number of memory blocks in the hidden layer. And from the formulas given above, it can be seen that the values of all memory cells and block outputs in the previous time step *t-1* will certainly affect the activations of all input gates, output gates, forget gates, even the input units in the current time step *t* in the same layer, except the case between memory cell and output gate – it is the current state of memory cell $\boldsymbol{c}_t$ rather than the state from previous time step that contributes to forget gate activation.

Overall, the LSTM memory cell can store and access information over long temporal range and thus avoid the vanishing gradient problem [18]. Therefore, LSTM could also be regarded as a natural extension of DNNs for temporal sequence data, where the deepness comes from layers through time.

Standard RNNs have access to past but not to future context. To exploit both, past and future context, RNNs can be extended to bidirectional RNNs, where two separate recurrent hidden layers scan the input sequences in opposite directions [28]. The network calculates its forward hidden layer activations $\boldsymbol{h}_t^f$ from the beginning to the end of the sequence, and its backward hidden layer activations $\boldsymbol{h}_t^b$ from the end to the beginning of the sequence, then updates the output layer by

$$\boldsymbol{y}_t = \boldsymbol{W}_{fy}\boldsymbol{h}_t^f + \boldsymbol{W}_{by}\boldsymbol{h}_t^b + \boldsymbol{b}_y, \tag{17}$$

where $\boldsymbol{W}_{fy}$ and $\boldsymbol{W}_{by}$ are the forward and backward weight matrices, and $\boldsymbol{b}_y$ is the hidden bias vector. The forward and backward directed layers are connected to the same output layer, which therefore can access the whole context.

IV. Experiments and Results

A. Databases

To demonstrate the effectiveness of the proposed methods, two databases – a *French* and a *English* corpus were recorded beforehand in a realistic acoustic space environment. Both databases are collected for *speech controlled TV application*. This application is designed to enable the user to change the TV controls (volume, brightness, etc.) or browse the programs using her voice. Table I shows the statistics of the two databases. The French corpus is recorded in a living room with furniture, where one microphone near the mouth records the close talk, and another microphone array consisting of 16 channels records the distant talk. 22 native French speakers (11 females) were asked to speak naturally so as to control the TV as their wish, i.e., "je veux un film avec Cameron Diaz (I want a movie with Cameron Diaz)." Finally, 8.3 h recordings are obtained, including about 7 k sentences and 45 k words in total. The distant talk data obtained from a 16-channel microphone array is grouped into four disjoint sets (1-4, 5-8, 9-12, and 13-16). The four channel speech in each of the sets is beamformed and noise reduced to get a single speech signal. As a result, the amount of distant talk training/test data is four times its close talk counterpart. The whole database was then divided into training and test set speaker-independently and equally.

TABLE I
Distribution of Speakers, Sentences, Words, and Recording Time of Close Talk per Partition of French and English Corpora.

	French		English	
	train	test	train	test
# speakers (f/m)	11 (5/6)	11 (6/5)	9 (5/4)	11 (5/6)
# sentences	2 231	4 619	1 430	1 801
# words	15 148	30 094	7 886	9 907
time (hours)	4.1	4.2	2.9	3.4

Likewise, 6.3 h of recordings were captured for the English corpus which comprises 20 speakers (10 females), and approximate 3 k sentences, 18 k words in total. For French, the training and test data sets were recorded in the same room, but for English, these data sets were recorded in different rooms. The details of the French and English corpus are shown in Table I. In the ongoing, the proposed techniques is mainly evaluated on the French corpus. The English database is used to study the impact of mismatch in acoustic (room) environments between training and testing conditions.

B. Experimental Setup

The stereo training (close talk and distant talk) feature vectors are time aligned such that the Pearson product-moment correlation coefficient (PCC) is maximized between the MFCC-0 time series. The training utterances with maximum PCC coefficient lower than 0.9 were dropped to avoid utterances with severe channel distortions.

The mapping techniques were evaluated on the standard MFCCs. The 12 dimensional static MFCCs were appended to their first, second, and third order regression coefficients, resulting in a feature vector of size 48. The feature vectors of $\boldsymbol{x}_t^c$ and $\boldsymbol{x}_t^d$ are extracted from the close and distant talk signals, respectively, every 10 ms using a window size of 25 ms. Then, the differential feature vectors of $\boldsymbol{x}_t^{\Delta}$ are acquired by $\boldsymbol{x}_t^c - \boldsymbol{x}_t^d$. Furthermore, before training the neural networks, the global means and variances are calculated over the close talk, distant talk, and their differential feature vectors of the whole neural network's training sets. Then, mean and variance normalization are performed over the network inputs and targets (i.e., the absolute or the differential feature vectors) using the means and variances from the corresponding sets, respectively.

For the neural networks, both input and output node numbers are equal to the dimension of the feature vector (48 in our case) except that stacked frames are used as input. And one hidden layer with 200 neurons is chosen. Particularly, for the LSTM memory block, input and output gates adopt hyperbolic tangent (tanh) activation functions, and the forget gates take logistic activation functions.

During network training, gradient descent is implemented with a learning rate of 10^{-5} and a momentum of 0.9. Zero mean Gaussian noise with standard deviation 0.1 is added to the input activations in the training phase in order to improve generalization. All weights are randomly initialized in the range from -0.1 to 0.1. Finally, the early stopping strategy is used as no improvement of the MSE on the evaluation set has been observed during 20 epochs.

C. Speech Recognition Evaluation

The effectiveness of different mapping strategies and neural network configurations was evaluated on a research ASR system available off-the-shelf. The acoustic models were trained on mobile data collected on hand held devices. The performance of the ASR is measured and compared in terms of word error rate (WER) and its relative reduction (WERR) metrics, and the baselines for the close talk and distant talk of the French corpus are 11.8% and 19.41% WERs, respectively.

1) Neural Network Architectures: A performance comparison is shown in Table II between BLSTM networks and other networks such as MLP, and recurrent networks without memory (RNNs) or with memory (i.e., LSTM) Note that, according to the empirical experience, the best performance for training MLP and (B)RNN was achieved by a learning rate of 10^{-6}, as opposed to 10^{-5} for the (B)LSTM networks.

From Table II, it can be seen that, when no context is used at the input of the MLP, there is an increase in WER compared to the baseline. Whereas, the recurrent neural networks (standard RNN and more sophisticated LSTM) show lower WERs. This is because of their ability to capture the contextual information implicitly. When the temporal

context is increased at the input of the MLP, there is a steady decrease in WERs and for 600 hidden nodes and a context of 7 frames, a WERR of 7% is delivered over the baseline system.

TABLE II
PERFORMANCE OF THE BASELINE RECOGNIZER AND DEREVERBERANT SYSTEMS BY ADOPTING VARIOUS NEURAL NETWORK ARCHITECTURES LIKE MLP, RNN, BRNN, LSTM, AND BLSTM WITH DIFFERENT NUMBER OF HIDDEN NEURONS AND STACKED FEATURE FRAMES. FR: FRAMES; WGT: WEIGHTS.

network	# neurons	#fr	# wgt	WER[%]	WERR[%]
w/o mapping (close talk)				11.81	
w/o mapping (distant talk)				19.41	
MLP	200	1	19 k	24.15	-24.4
	200	5	58 k	18.74	3.5
	200	7	77 k	18.72	3.6
	200	9	96 k	18.74	3.5
	600	7	230 k	18.06	7.0
RNN	200	1	59 k	19.07	1.8
BRNN	200	1	118 k	17.42	10.3
LSTM	200	1	180 k	18.47	4.8
BLSTM	200	1	360 k	**16.38**	15.6
BLSTM	200	7	590 k	16.43	15.4
BLSTM	144-200-144	1	1 M	**16.32**	15.9

RNN and LSTM models capture only the past information. However, for dereverberation, it is important to learn the temporal smearing in the future frames because the distant talk signal is delayed (future) and attenuated version of the close talk signal (cf. Section II-A). The bidirectional RNN and LSTM yield significant (one-side z-test, $p<0.001$) reduction in WERs compared to the corresponding unidirectional models capturing past information.

It can also be seen that both uni- and bi- directional LSTM models give lower WERs compared to the simple RNN models. This can be attributed to the sophisticated architecture of the individual neurons compared to the simple neuron. Previous acoustic information can be stored in the memory cell until the input gates and the forget gates allow to (partly) change it (cf. Section III).

Moreover, as seven successive frames are simultaneously fed into BLSTM networks, no improvement is observed from this side (see Table II). Hence, the BLSTM seems to learn context better if feature frames are presented one by one and the increased size of the input layer rather harms recognition performance. In addition, when increasing one hidden layer with 200 neurons to three hidden layers with 144-200-144 neurons, the performance improvement is not obvious. In the following experiments, one hidden layer with 200 neurons is kept as the BLSTM network's architecture on the French corpus.

To visualize the mapping learned by the BLSTM model, the trajectories of MFCC-0 for two randomly selected utterances are plotted in Fig. 3. The figure shows three trajectories – close talk (red), distant talk (green), and mapped (or estimated) close talk (blue). It can be seen that the MFCC-0 curves of mapped close talk speech (by BLSTM networks) are closer to the original one than the distant talk speech during the

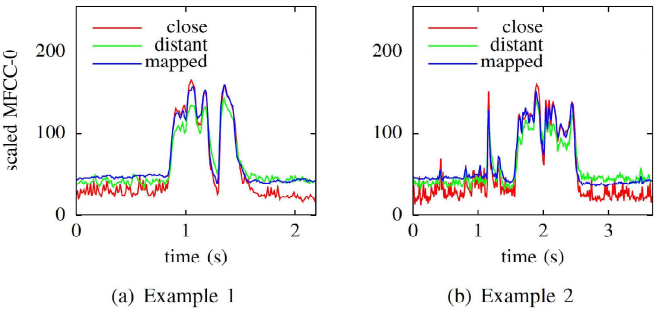

Fig. 3. The scaled MFCC-0 (0-255) of a close talk utterance (red), a distant talk one (green), and a mapped close talk one (blue) for two examples.

speaking period, and are smoother during the silence period. This indicates that the reverberant signals and channel noise are successfully suppressed. Such a feature enhancement phenomenon can be further confirmed over the entire training set and the whole feature vectors. Fig. 4 presents the PCCs of the 48 MFCCs between distant talk utterances (hollow circle and dotted line)/mapped utterances (solid circle and line) and close talk utterances over the whole training set. Obviously, the PCCs are boosted after reverberated features are enhanced, which could demonstrate the performance improvement of ASR by using a BLSTM dereverberation model.

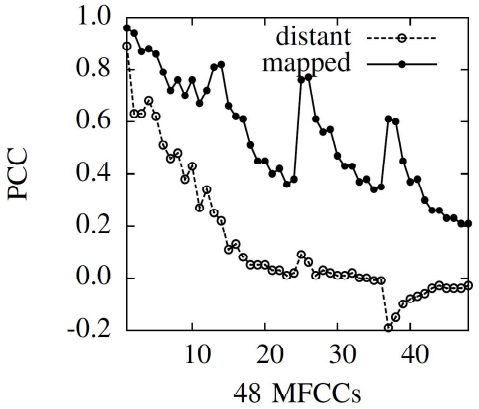

Fig. 4. Pearson product-moment Correlation Coefficient (PCC) of 48 MFCCs between distant talk utterances (hollow circle and dotted line)/mapped close talk utterances (solid circle and line) and close talk utterances over the whole training set.

2) Training on Differential Targets: As discussed in Subsection II-B, there are two ways to obtain the enhanced features from distant talk, either by direct way (training networks with absolute targets) or by indirect way (training networks with differential targets). Table III compares the performance of the two mapping ways in ASR system.

By checking three types of BLSTM network structure, the BLSTM dereverberation models trained on differential targets perform better than the models trained on the absolute targets when the network structure is simpler. It can be seen that a gain of about 3% relative WERR (at the 0.05 significance level in a one-side z-test) is achieved when only 144 neurons are used in only one hidden layer, compared to using absolute targets.

TABLE III
PERFORMANCE COMPARISON BY USING *ABS*OLUTE TARGETS AND *DIFF*ERENTIAL TARGETS.

targets	# neurons	WER [%]	WERR [%]
abs.	144	17.04	12.2
diff.	144	16.52	14.9
abs.	200	16.38	15.6
diff.	200	16.43	15.4
abs.	144-200-144	16.32	15.9
diff.	144-200-144	16.29	16.1

To find out the rationale behind this phenomenon, the distribution of globally normalized log energies (MFCC-0) on the absolute targets (a) and the differential targets (b) over the whole French corpus is plotted in Fig. 5. Obviously, the differential targets have a symmetrical unimodal distribution which is centered around zero. In contrast, the absolute target has a bimodal distribution which could be harder to learn. Therefore, the simpler the neural networks are, the higher a gain would be obtained via training on the differential targets. Such superiority of differential targets-based learning can further be verified in Subsections IV-C3 and IV-C4.

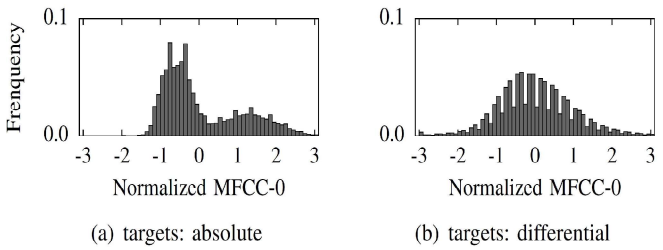

Fig. 5. Distribution of normalized log energy (MFCC-0) of absolute targets (a) and differential targets (b).

3) Incorporating CMLLR and MLLR: As the distant talk is passing through the BLSTM dereverberation models, its feature vectors are transformed (almost) to the clean target, which most preexisting acoustic models are trained on. Thus, this technique could also be considered as a sort of feature adaptation. It is interesting to see whether incorporating back-end adaptation techniques like CMLLR and MLLR can further enhance the ASR performance.

As expected, without our mapping technique the WERs for distant talk decrease from 19.41%, over 19.01%, to 17.19% with no adaptation, CMLLR, and CMLL + MLLR, respectively (as shown in Table IV). The WERs drop further to 16.43%, 16.34%, and 15.68% when integrating with our suggested mapping technique, which results in 15.4%, 13.8% and 7.8% relative WERR, respectively (All improvements are at the 0.001 significance level in a one-side z-test). Overall, the best result is achieved by combining both mapping and adaptation (CMLLR + MLLR) techniques, with 8.8% and 19.2% performance improvement in WERR in comparison with adaptation techniques only and the baseline (w/o adaptation and mapping), respectively. Additionally, Table IV also shows that if the close talk was falsely detected as distant talk and fed into the mapping and adaptation systems, the WER will increase about 10% relatively.

TABLE IV
ASR EVALUATION ON DISTANT TALK AND CLOSE TALK SETS BY COMBINING BLSTM DEEVERBERATION AND ADAPTATION (CMLLR AND MLLR) TECHNIQUES. ABS./DIFF.: ABSOLUTE/DFFERENTIAL TRGETS.

[%] adaptation	tar-gets	distant talk WER	distant talk WERR	close talk WER	close talk WERR
w/o adaptation					
w/o mapping		19.41		11.81	
w/ mapping	abs.	16.38	15.6	14.47	-22.5
w/ mapping	diff.	16.43	15.4	14.02	-18.7
w/ CMLLR					
w/o mapping		19.01	2.0	11.78	-0.3
w/ mapping	abs.	16.14	16.8	13.70	-16.0
W /mapping	diff.	16.34	15.8	13.46	-13.9
w/ CMLLR+MLLR					
w/o mapping		17.19	11.4	11.63	-1.5
w/ mapping	abs.	15.70	19.1	13.33	-12.9
w/ mapping	diff.	**15.68**	19.2	13.04	-10.4

4) Inter Room Evaluation: In the above experiments, the data set used for training the dereverberation model is recorded in the same room with the evaluation set. In the real-life application, however, the evaluation scenarios are always unpredictable. That is, the acoustic environments (i.e., room size, type) for creating the training data normally mismatch with the evaluation scenarios. To cope with this problem, several artificially reverberant corpora were synthesized on the close talk set of French by convolving various RIRs and adding a little noise. The rooms to create the RIRs are different with the ones for creating the French corpus. When generating the simulated corpora, three elements were taken into account: positions variation of the speakers w.r.t. the microphones, the weights of the reverberation signal and the weights of the noise signal. The first column of Table V shows the four scenarios of simulated speech. The second to sixth columns represent the WER and WERR for each simulated corpus without mapping, mapping to the absolute targets, and mapping to the differential targets, respectively. As observed from the table, the BLSTM dereverberant ASR systems prevail over the systems without dereverberation, which overall leads to a reduction of WER with 3.3% relatively by the usage of absolute targets and 6.6% relatively by the usage of differential targets.

TABLE V
ASR EVALUATION ON THE *ARTIFICIAL* DISTANT TALK SET USING THE BLSTM DEREVERBERANTION MODELS TRAINED ON THE *NATURAL* DISTANT TALK SET. POS: POSITION OF SPEAKERS W.R.T. MICROPHONES; R/N: REVERBERANT/NOISY SIGNAL WEIGHTS (DB); W/O: WITHOUT MAPPING; ABS./DIFF.: ABSOLUTE/DIFFERENTIAL TARGETS.

[%]	w/o WER	abs. WER	abs. WERR	diff. WER	diff. WERR
Pos-1,R:-100,N:-30	20.86	20.06	3.8	19.24	7.8
Pos-1,R:-30,N:-100	21.28	20.59	3.2	19.97	6.2
Pos-2,R:-100,N:-30	20.24	19.25	4.9	18.78	7.2
Pos-2,R:-30,N:-100	19.89	19.70	1.0	18.71	5.9
average	20.57	19.90	3.3	**19.18**	6.8

In addition, the experiments were repeated on a realistic English corpus, of which the training and test sets are recorded in totally different rooms (cf. Section IV-A). The baselines of the distant talk of English corpus are WERs of 18.30% and

18.77% for the training and test sets, both of which almost double the baseline of close talk (WERs of 9.27% and 9.48% for the training and test sets). As expected, a high gain is obtained for the training set when applying channel mapping. Nevertheless, such high gain is not observed for the test set. Only when using the differential targets to train neural networks, a gain can be obtained by 5.5% of WERR on the mismatched test set, and can be enlarged to 7.7% WERR when the utterance level CMN is implemented [7]. In this experiment, it can also be noticed that the indirect mapping way (using differential targets for networks training) significantly overcomes the direct mapping way (using absolute targets for networks training).

TABLE VI
ASR EVALUATION ON THE TRAINING AND TEST SETS OF THE ENGLISH CORPUS BY USING THE BLSTM (ONE HIDDEN LAYER WITH 128 NEURONS) FEATURE DEREVERBERATION MODEL TRAINED ON THE TRAINING SET. ABS./DIFF.: ABSOLUTE/DIFFERENTIAL TARGETS. CMN (UTT.): UTTERANCE LEVEL CEPSTRAL MEAN NORMALIZATION.

[%]	tar-gets	training set		test set	
		WER	WERR	WER	WERR
w/o mapping (close talk)		9.27		9.48	
w/o mapping (distant talk)		18.30		18.77	
BLSTM	abs.	15.80	13.7	20.67	-10.0
BLSTM	diff.	15.26	16.6	**17.73**	5.5
BLSTM+CMN(utt.)	abs.	14.61	20.2	19.38	-3.0
BLSTM+CMN(utt.)	diff.	14.96	18.3	**17.32**	7.7

From the above two experiments, the results imply that the inter-room scenario is more challenging when compared to the intra-room scenario shown in Section IV-C1 to IV-C3. On the one hand, the performance improvement on both training and test sets indicates that different rooms share some common reverberation information. These shared information can be learned by the BLSTM networks. On the other hand, the different gains obtained by the training and test sets suggest that the networks probably learn too much information from a specific acoustic environment.

V. CONCLUSIONS

In this study, a feature-based dereverberation method was proposed and investigated for realistic hands-free voice controlled devices. The basic idea is to use bidirectional long short-term memory (BLSTM) neural networks for channel mapping – from distant talk cepstal feature space to its close talk counterpart.

In such application scenario, the speech signal at each frame time will impact the subsequent frames in a long-term. This consequentially requires a learning algorithm which could not only access long-term context information but also make use of the future information. The bidirectional structure (past and future) of LSTM neural networks is capable of dealing with these problems. The experimental results on a French corpus show a word error rate reduction (WERR) of more than 16% for ASR, which significantly outperform the 'conventional' networks Multilayer Perceptron (MLP) (one-side z-test, $p<0.001$) and bidirectional recurrent neural networks (BRNNs) (one-side z-test, $p<0.05$). Such effectiveness of our feature mapping method is further confirmed by integrating widely used adaptation techniques of maximum linear likelihood regression (MLLR) or/and constrained MLLR (CMLLR), which yields the best performance of about 20% of WERR. And it is also confirmed in the scenario of inter-room evaluation, as the mismatched evaluation sets in acoustic environment also obtain a gain via channel mapping when using BLSTM.

This study also presents another indirect way for channel mapping – the differential feature vectors (between the distant talk speech and the close talk speech) as network targets, then adding the estimated differential feature vectors to the counterpart of original distant talk. The results based on a rich number of experiments show that this indirect mapping strategy can compete with the previously used direct mapping strategy, particularly in some cases like using a simple network structure and evaluating mismatched data sets. All these cases are quite welcome for real-life applications.

Due to a gain gap between matched and mismatched evaluation cases, future work will focus on the further exploitation of joint acoustic information across different rooms with the goal of 'blind' dereverberation application. On a way to achieve this is to train the networks by a vast amount of reverberant speech collected in a variety of rooms. Further, one can apply to the objective functions some generalization terms such as weight decay. In addition, it seems also beneficial to develop a way of selecting predefined mapping models for different room categories, in order to ultimately explore the advantages of the room-specific models.

REFERENCES

[1] K. Fujita, H. Kuwano, T. Tsuzuki, Y. Ono, and T. Ishihara, "A new digital TV interface employing speech recognition," *IEEE Trans. Consum. Electron.*, vol. 49, no. 3, pp. 765–769, 2003.

[2] T. Giannakopoulos, N.-A. Tat1las, T. Ganchev, and I. Potamitis, "A practical, real-time speech-driven home automation front-end," *IEEE Trans. Consum. Electron.,* vol. 51, no. 2, pp. 514–523, 2005.

[3] P. Ding, L. He, X. Yan, R. Zhao, and J. Hao, "Robust mandarin speech recognition in car environments for embedded navigation system," *IEEE Trans. Consum. Electron.*, vol. 54, no. 2, pp. 584–590, 2008.

[4] T. Virtanen, R. Singh, and B. Raj, Techniques for Noise Robustness in Automatic Speech Recognition. New York, NY: John Wiley & Sons, 2012.

[5] P. A. Naylor and N. D. Gaubitch, Speech Dereverberation. Berlin: Springer, 2010.

[6] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoustics, Speech and Signal Process.*, vol. 36, no. 2, pp. 145–152, 1988.

[7] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoustics, Speech and Signal Process.*, vol. 29, no. 2, pp. 254–272, 1981.

[8] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, 1994.

[9] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, 1994.

[10] C. J. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, no. 2, pp. 171–185, 1995.

[11] M. J. Gales, "Maximum likelihood linear transformations for HMMbased speech recognition," *Comput. Speech Lang.*, vol. 12, no. 2, pp. 75–98, 1998.
[12] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process Mag.*, vol. 29, no. 6, pp. 82–97, 2012.
[13] T. Yoshioka, X. Chen, and M. J. Gales, "Impact of single-microphone dereverberation on dnn-based meeting transcription systems," *in proc. International Conference on Acoustics, Speech and Signal Processing*, Florence, Italy, 2014, pp. 5564–5568.
[14] M. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," *in proc. International Conference on Acoustics, Speech and Signal Processing*, Vancouver, Canada, 2013, pp. 7398–7402.
[15] W. Li, J. Dines, and M. Magimai.-Doss, "Robust overlapping speech recognition based on neural networks," Martigny, Switzerland, Tech. Rep. IDIAP-RR-55-2007, 2007.
[16] A. L. Maas, T. M. O'Neil, A. Y. Hannun, and A. Y. Ng, "Recurrent neural network feature enhancement: The 2nd CHiME challenge," *in proc. CHiME Workshop,* Vancouver, Canada, 2013, pp. 79–80.
[17] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: Datasets, tasks, baselines," *in Proc. International Conference on Acoustics, Speech and Signal Processing*, Vancouver, Canada, 2013, pp. 126–130.
[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[19] A. Graves, M. Liwicki, S. Fernandez, H. Bertolami, R. andBunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 855–868, 2009.
[20] C. Plahl, M. Kozielski, R. Schlüter, and H. Ney, "Feature combination and stacking of recurrent and non-recurrent neural networks for LVCSR," *in proc. International Conference on Acoustics, Speech and Signal Processing*, Vancouver, Canada, 2013, pp. 6714–6718.
[21] M. Wöllmer, C. Blaschke, T. Schindl, B. Schuller, B. Farber, S. Mayer, and B. Trefflich, "Online driver distraction detection using long short-term memory," *IEEE Trans. Intell Transp. Syst.*, vol. 12, no. 2, pp. 574–582, 2011.
[22] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient flow in recurrent nets: The difficulty of learning long-term dependencies," in A Field Guide to Dynamical Recurrent Neural Network, S. C. Kremer and J. F. Kolen, Eds. New York, NY: IEEE Press, 2001, pp. 1–15.
[23] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," arXiv preprint arXiv:1402.1128, 2014.
[24] M. Wöllmer, Z. Zhang, F. Weninger, B. Schuller, and G. Rigoll, "Feature enhancement by bidirectional LSTM networks for conversational speech recognition in highly non-stationary noise," *in proc. International Conference on Acoustics, Speech and Signal Processing*, Vancouver, Canada, 2013, pp. 6822–6826.
[25] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The Munich feature enhancement approach to the 2nd CHiME challenge using BLSTM recurrent neural networks," *in proc. CHiME Workshop*, Vancouver, Canada, 2013, pp. 86–90.
[26] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Process Mag.*, vol. 29, no. 6, pp. 114–126, 2012.
[27] O. Vinyals, S. V. Ravuri, and D. Povey, "Revisiting recurrent neural networks for robust ASR," *in proc. International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, 2012, pp. 4085–4088.
[28] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, 1997.

BIOGRAPHIES

Zixing Zhang received his master degree (2010) in telecommunications from Beijing University of Posts and Telecommunications, Beijing, China. He is currently pursuing his Ph.D. degree as a Researcher in the Machine Intelligence & Signal Processing (MISP) Group at the Institute for MMK at Technische Universität München (TUM) in Munich, Germany. His current research focuses on efficient machine learning algorithms for robustness of automatic speech recognition and computational paralinguistics.

Joel Pinto is a Research Manager at Nuance Communications in Aachen, Germany. He holds a PhD from École Polytechnique Fédérale de Lausanne, Switzerland (2010) and a Master in Engineering degree from the Indian Institute of Science, India (2003), both in Electrical Engineering. During his doctoral studies, he was with Idiap Research Institute, Switzerland working on neural network based acoustic modeling for automatic speech recognition. Between 2003-2005, he was with Hewlett Packard Labs India working in the area of speech and language technology.

Christian Plahl received the diploma degree in computer science from the University of Bielefeld, Bielefeld, Germany, in 2005 and the Ph.D. degree in Computer Science from the RWTH Aachen University, Aachen, Germany, in 2014. Since 2013 he is working as a Research Scientist at Nuance Communications, Inc., Aachen, Germany. His research interests cover speech recognition, signal analysis, deep learning and artificial neural networks.

Björn Schuller (M'05) received his diploma in 1999, and his doctoral degree in 2006 and his habilitation in 2012, all in electrical engineering and information technology from TUM. He is tenured head of the MISP Group at TUM since 2006, senior lecturer at the Imperial College London's Department of Computing in the UK, CEO of audEERING UG (limited), Visiting Professor of HIT in Harbin/China, Associate of CISA in Geneva/Switzerland and Joanneum Research in Graz Austria since 2013. Best known are his works advancing Machine Intelligence for Speech Analysis. He is the president of the Association for the Advancement of Affective Computing (AAAC), and member of the IEEE and its Speech and Language Processing Technical Committee (SLTC), ACM, and ISCA and (co-)authored five books and more than 400 publications in the field leading to more than 6000 citations – his current h-index equals 39.

Daniel Willett received his diploma of Computer Science from the Technical University in Darmstadt in 1994 and Ph.D. in Electrical Engineering from Duisburg University in 2000, both in Germany. He was Postdoc at NTT Communication Science Laboratories in Kyoto, Japan, from 2000 to 2002. In 2002, he joined Harman-Becker in Ulm, Germany, where he worked on in-car speech recognition until 2006, when he joined Nuance Communications in Aachen, Germany, where he has since been working on large vocabulary automatic speech recognition and nowadays runs a research team that focuses on cloud-based speech recognition. With nearly 20 years in the speech recognition area, he contributed by numerous publications to the field as well as innovations to real-world speech recognition systems as widely deployed today.