Personalizing Your Speech Interface with Context Adaptive Deep Neural Networks

Marc Delcroix, Keisuke Kinoshita, Atsunori Ogawa, Shigeki Karita, Takuya Higuchi, and Tomohiro Nakatani

Abstract

This article introduces our recent progress in speaker adaptation of neural network based acoustic models for automatic speech recognition. Deep neural networks have greatly improved the performance of speech recognition systems, enabling the recent widespread use of speech interfaces. However, recognition performance still greatly varies from one speaker to another. To address this issue, we are pursuing research on novel deep neural network architectures that enable rapid adaptation of network parameters to the acoustic context, for example, the speaker voice characteristics. The proposed network architecture is general and can potentially be used to solve other problems requiring adaptation of neural network parameters to some context or domain.

Keywords: deep learning, automatic speech recognition, speaker adaptation

1. Introduction

Automatic speech recognition (ASR) is being used more and more in our everyday life. For example, it is now common to speak to our smartphones to ask for the weather forecast or the nearest restaurant. Communication agents such as home assistants and robots are also starting to enter our living rooms, suggesting that speech may become a common modality for accessing information in the near future.

The rapid expansion of ASR based products has been made possible by the significant recognition performance gains achieved through the recent introduction of deep neural networks (DNNs) [1]. However, simply using DNNs does not solve all the issues. Speech recognition performance can still greatly vary depending on the acoustic context such as the speaker voice characteristics or the noise environment.

In this article, we describe our approach to tackle this problem by making the ASR system adaptive to the acoustic context. To achieve this, we have developed a novel DNN architecture that we call context adaptive DNN (CADNN) [2]. A CADNN is a neural network whose parameters can change depending on the external context information such as speaker or noise characteristics. This enables us to rapidly generate an ASR system that is optimal for recognizing speech from a desired speaker, opening the way to better ASR performance.

In the remainder of this article, we briefly review how current ASR systems work, focusing on the acoustic modeling part. We then describe in more detail the proposed CADNN and a speaker adaptation experiment we conducted to confirm its potential. We conclude this article by discussing some outlooks on potential extensions of CADNNs to achieve online speaker adaptation and applications to other research areas.

2. Deep learning based acoustic modeling

A speech recognition system is composed of several components, as illustrated in **Fig. 1**. First, there is a feature extraction module, which extracts speech

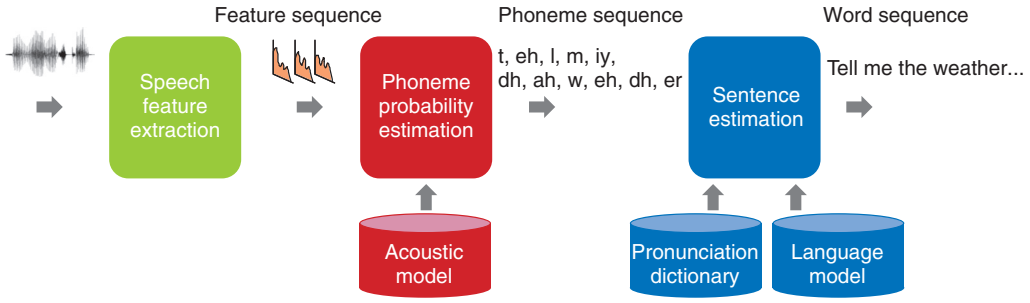


Fig. 1. Speech recognition system.

features from each short time frame of about 30 ms of a speech signal. Then, the acoustic model computes the probability that a speech feature corresponds to a given phoneme. Finally, the decoder finds the best word sequence given the input sequence of features by taking into account the phoneme probabilities obtained from the acoustic model, a pronunciation dictionary that maps the phoneme sequences to words, and scores obtained from the language model that outputs the probability of word sequences. In the remainder of this article, we focus our discussion on the acoustic model, and in particular on speaker adaptation.

Recently developed acoustic models use DNNs to map speech features to phoneme probabilities. An example of such an acoustic model is shown in **Fig. 2**. A DNN consists of several hidden layers that perform a nonlinear transformation of their input. With these stacked hidden layers, a DNN can model a complex mapping between its input features and its outputs. In the context of acoustic modeling, the inputs are speech features and the outputs are phoneme probabilities. Training such a DNN requires a large amount of speech data, from a few dozen hours to thousands of hours, depending on the task. The training data must also include the actual spoken phoneme sequences that can be derived from manual transcriptions of the utterances. With such training data, the acoustic model training follows the standard procedure for training DNNs such as error backpropagation with stochastic gradient descent.

To ensure that the acoustic model can well recognize speech in a variety of acoustic contexts such as for different speakers, the training data must contain speech from a large variety of speakers. Using such diverse training data enables us to obtain a good model on average. However, the DNN may not be optimal for a given speaker seen during the deployment of the recognition system because of the speaker's specific speaking style, which may result in poorer ASR performance for that particular speaker.

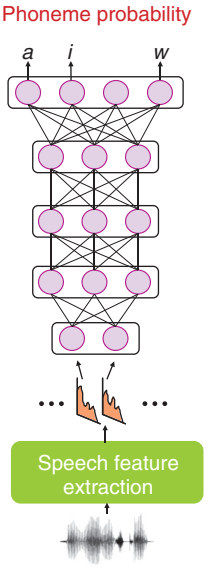


Fig. 2. DNN based acoustic model.

Solving this issue requires us to adapt the acoustic model to the desired speaker. However, adapting the acoustic model is challenging because it is often difficult to obtain the large amount of speech data with transcription that would be needed to train an acoustic model for the desired context. Specifically, it is impractical to require several hours of speech from each user to create a personalized acoustic model. In many applications, acoustic model adaptation should thus be *rapid*, that is, requiring a small amount of speech data such as a few seconds, and *unsupervised*, meaning it does not require transcribed data.

3. CADNN

Extensive research has been done to find approaches for adapting an acoustic model to speakers. A recent promising attempt consists of informing the DNN about the speaker by adding to its input an auxiliary feature describing the speaker characteristics. Such approaches have interesting properties because the speaker feature can be computed with only a few seconds of speech data, and they do not require transcriptions. However, simply adding an auxiliary feature to the input of a DNN has only a limited effect, as it can only partially adapt the DNN parameters. In this article, we describe an alternative way to exploit auxiliary information through a CADNN.

The idea behind CADNN is that a network trained for a given context should be optimal to recognize speech in that acoustic context. For example, we could build different networks to recognize speech from female and male speakers. Adaptation could then be realized simply by selecting the network corresponding to the target acoustic context. Such a naïve approach raises two issues. First, only part of the training data can be used for training each of the separate models. This would seem to be suboptimal because, for example, some speech characteristics are common to all speakers, and thus, better models could be trained when exploiting all the training data. Another issue is that it is unclear how to select the acoustic model in an optimal way.

The CADNN addresses these issues by making only part of the network dependent on the acoustic context. Moreover, we propose to select the model parameters using auxiliary features representing the acoustic context such as the speaker characteristics. A schematic diagram of a CADNN is shown in **Fig. 3** [3]. As illustrated in the figure, a CADNN has one hidden layer replaced by a context adaptive layer, that is, a layer that is split into several sublayers, each associated with a different acoustic context class.

For example, with two acoustic context classes, we could have a sublayer for male speakers and a sublayer for female speakers. The output of the hidden layer is obtained as a weighted sum of the output of each sublayer, with context weights derived from the auxiliary features. In our implementation, the context weights are computed from a small auxiliary network that has the auxiliary features as inputs. The outputs are the context weights that are optimal for recognizing speech for that acoustic context.

A CADNN has several interesting properties. The auxiliary network and the CADNN can be connected and trained jointly. This means that we can obtain context weights that are optimal for the acoustic context. Moreover, using such a joint training scheme, we do not need to explicitly define the acoustic context classes; they can be automatically learned from the training data during the training procedure. Finally, since except for the factorized layer, the rest of the network is shared among all the different acoustic context classes, all the training data can be used to train the parameters of the network.

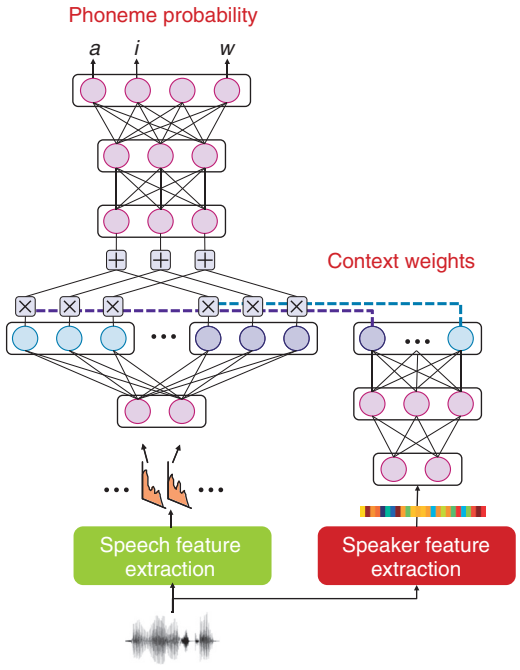


Fig. 3. Proposed CADNN for speaker adaptation.

4. Rapid speaker adaptation with CADNN

A CADNN can be used to achieve rapid speaker adaptation of acoustic models. The graph in **Fig. 4** shows the word error rate for recognition of English sentences read from the Wall Street Journal. Note that lower word error rates indicate better ASR performance. Our baseline system consists of a DNN with five hidden layers with ReLU (rectified linear unit) activations. The proposed CADNN uses a similar topology to that of the baseline DNN but has its second hidden layer replaced with a context adaptive layer, with four context classes. As auxiliary features, we use features representing speakers that are widely used for speaker recognition tasks. These auxiliary features were computed using a single utterance,

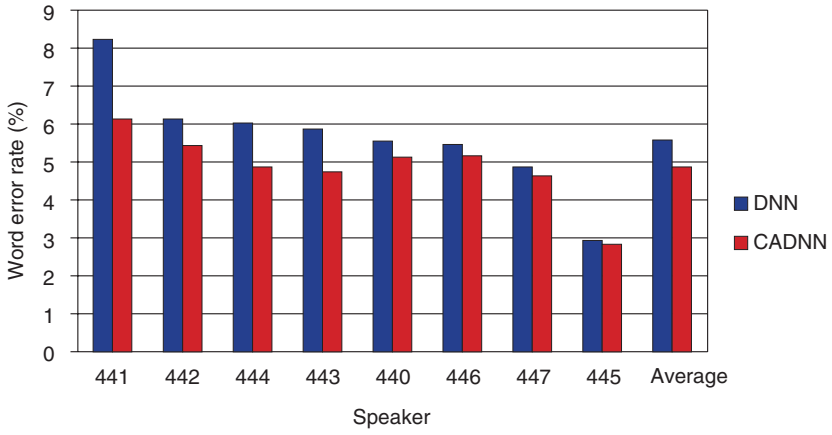


Fig. 4. Proposed CADNN for speaker adaptation.

which corresponds in this experiment to less than 10 s of speech data. Moreover, the speaker features can be obtained without transcriptions.

These results demonstrate that the proposed CADNN was able to significantly improve ASR performance, with a relative improvement of about 10% over the baseline. Since only a few seconds of speech data without transcriptions are sufficient to compute the auxiliary features, this experiment proves that CADNN can achieve rapid unsupervised speaker adaptation.

5. Outlook

The proposed CADNN appears promising for unsupervised rapid speaker adaptation of acoustic models. Potential further improvement could be achieved by developing better speaker representation for the auxiliary features [4]. Moreover, extension of the proposed scheme to online adaptation, where the adaptation process could start with even less data, is also a challenging research direction [5].

Finally, the proposed CADNN architecture is general and could be applied to other problems. For example, we are currently exploring the use of the same principle to extract a target speaker from a mixture of speakers [6]. We also believe that the proposed CADNN could be employed in other fields requiring context or domain adaptation of DNNs.

References

[1] Y. Kubo, A. Ogawa, T. Hori, and A. Nakamura, "Speech Recognition Based on Unified Model of Acoustic and Language Aspects of Speech," NTT Technical Review, Vol. 11, No. 12, 2013. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201312fa4.html

[2] M. Delcroix, K. Kinoshita, T. Hori, and T. Nakatani, "Context Adaptive Deep Neural Networks for Fast Acoustic Model Adaptation," Proc. of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4535–4539, South Brisbane, Australia, Apr. 2015.

[3] M. Delcroix, K. Kinoshita, C. Yu, A. Ogawa, T. Yoshioka, and T. Nakatani, "Context Adaptive Deep Neural Networks for Fast Acoustic Model Adaptation in Noisy Conditions," Proc. of ICASSP 2016, pp. 5270–5274, Shanghai, China, Mar. 2016.

[4] K. Veselý, S. Watanabe, K. Žmolíková, M. Karafiát, L. Burget, and J. Černocký, "Sequence Summarizing Neural Network for Speaker Adaptation," Proc. of ICASSP 2016, pp. 5315–5319, Shanghai, China, Mar. 2016.

[5] T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa, T. Asami, S. Katagiri, and T. Nakatani, "Cumulative Moving Averaged Bottleneck Speaker Vectors for Online Speaker Adaptation of CNN-based Acoustic Models," Proc. of ICASSP 2017, New Orleans, USA, Mar. 2017.

[6] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, "Speaker-aware Neural Network Based Beamformer for Speaker Extraction in Speech Mixtures," Interspeech 2017, Stockholm, Sweden, Aug. 2017.

Marc Delcroix
Senior Research Scientist, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

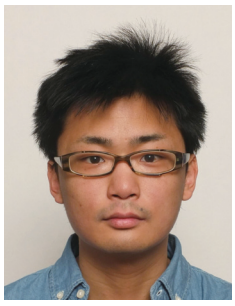
He received an M.Eng. from the Free University of Brussels, Brussels, Belgium, and the Ecole Centrale Paris, Paris, France, in 2003 and a Ph.D. from the Graduate School of Information Science and Technology, Hokkaido University, in 2007. He joined NTT Communication Science Laboratories as a research associate in 2007 and became a permanent research scientist in 2012. His research interests include robust multi-microphone speech recognition, acoustic model adaptation, integration of speech enhancement front-end and recognition back-end, speech enhancement, and speech dereverberation. He took an active part in the development of NTT robust speech recognition systems for the REVERB and CHiME 1 and 3 challenges, which all achieved best performance results in the tasks. He was one of the organizers of the REVERB challenge 2014 and of ASRU 2017 (Automatic Speech Recognition and Understanding Workshop). He is a visiting lecturer in the Faculty of Science and Engineering, Waseda University, Tokyo. He received the 2005 Young Researcher Award from the Kansai branch of the Acoustical Society of Japan (ASJ), the 2006 Student Paper Award from the Institute of Electrical and Electronics Engineers (IEEE) Kansai branch, the 2006 Sato Paper Award from ASJ, the 2015 IEEE ASRU Best Paper Award Honorable Mention, and the 2016 ASJ Awaya Young Researcher Award. He is a senior member of IEEE and a member of ASJ.

Keisuke Kinoshita
Senior Research Scientist, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received an M.Eng. and a Ph.D. from Sophia University, Tokyo, in 2003 and 2010. Since joining NTT in 2003, he has been engaged in research on speech and audio signal processing. His research interests include single- and multichannel speech enhancement and robust automatic speech recognition. He received Institute of Electronics, Information and Communication Engineers (IEICE) Paper Awards (2006), ASJ Technical Development Awards (2009), an ASJ Awaya Young Researcher Award (2009), Japan Audio Society Award (2010), and the Maejima Hisoka Award (2017). He is a member of IEEE, ASJ, and IEICE.

Atsunori Ogawa
Senior Research Scientist, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received a B.E. and M.E. in information engineering and a Ph.D. in information science from Nagoya University, Aichi, in 1996, 1998, and 2008. He joined NTT in 1998. He is involved in researching speech recognition and speech enhancement. He received the ASJ Best Poster Presentation Awards in 2003 and 2006. He is a member of IEEE, the International Speech Communication Association (ISCA), IEICE, the Information Processing Society of Japan, and ASJ.



Shigeki Karita
Research Scientist, Media Information Laboratory, Signal Processing Research Group, NTT Communication Science Laboratories.

He received a B.Eng. and M.Eng. from Osaka University in 2014 and 2016. He joined NTT in 2016. His research interests include end-to-end speech recognition and speech translation. He received a Young Researcher Award from IEICE in 2014. He is a member of ASJ and ISCA.

Takuya Higuchi
Research Scientist, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received a B.E. and M.E. from the University of Tokyo in 2013 and 2015. He has been with NTT since 2015, where he has been working on acoustic signal processing, array signal processing, blind source separation, and noise robust automatic speech recognition. He is a member of IEEE and ASJ.

Tomohiro Nakatani
Group Leader and Senior Distinguished Researcher, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received a B.E., M.E., and Ph.D. from Kyoto University in 1989, 1991, and 2002. Since joining NTT as a researcher in 1991, he has been investigating speech enhancement technologies for intelligent human-machine interfaces. He spent a year at Georgia Institute of Technology, USA, as a visiting scholar in 2005. He has also been a visiting assistant professor in the Department of Media Science, Nagoya University, since 2008. He received the 2005 IEICE Best Paper Award, the 2009 ASJ Technical Development Award, the 2012 Japan Audio Society Award, the 2015 IEEE ASRU Best Paper Award Honorable Mention, and the 2017 Maejima Hisoka Award. He was a member of the IEEE SP Society Audio and Acoustics Technical Committee (AASP-TC) from 2009 to 2014 and served as the chair of the AASP-TC Review Subcommittee from 2013 to 2014. He is a member of the IEEE SP Society Speech and Language Processing Technical Committee (SL-TC). He served as an associate editor of the IEEE Transactions on Audio, Speech and Language Processing from 2008 to 2010 and was a chair of the IEEE Kansai Section Technical Program Committee from 2011 to 2012, a Technical Program co-Chair of IEEE WASPAA-2007, a Workshop co-Chair of the 2014 REVERB Challenge Workshop, and a General co-Chair of the IEEE Automatic Speech Recognition and Understanding Workshop. He is a member of IEEE, IEICE, and ASJ.