Speech Recognition Based on Unified Model of Acoustic and Language Aspects of Speech

Yotaro Kubo, Atsunori Ogawa, Takaaki Hori, and Atsushi Nakamura

Abstract

Automatic speech recognition has been attracting a lot of attention recently and is considered an important technique to achieve natural interaction between humans and machines. However, recognizing spontaneous speech is still considered to be difficult owing to the wide variety of patterns in spontaneous speech. We have been researching ways to overcome this problem and have developed a method to express both the acoustic and linguistic aspects of speech recognizers in a unified representation by integrating powerful frameworks of deep learning and a weighted finite-state transducer. We evaluated the proposed method in an experiment to recognize a lecture speech dataset, which is considered as a spontaneous speech dataset, and confirmed that the proposed method is promising for recognizing spontaneous speech.

Keywords: speech recognition, deep learning, spontaneous speech

1. Introduction

Automatic speech recognition refers to the technology that enables a computer to extract and identify the words contained in given speech signals. Recently, systems using speech recognition technology have become increasingly common and have been used in several real world applications.

In addition to the development of systems and applications, basic techniques to enable accurate recognition have also been intensively investigated. In the first era of speech recognition technology, speech recognizers were only able to recognize speech uttered by one preregistered person. In the last decade, however, speech recognizers have become more powerful, and this has enabled recognition of speech from unknown persons if the input speech signals are appropriately uttered. Consequently, the current state-of-the-art technologies focus mainly on recognition of *spontaneous speech.*

Recognition of spontaneous speech is difficult because the assumption we introduced above, that speech is *appropriately uttered*, no longer holds in this case. The main objective of research on spontaneous speech recognition is to recognize speech signals even though they are inappropriately uttered but are still possible to be perceived by humans. Such inappropriate speech is characterized by several fluctuations in the input signals. For example, utterances such as *ah, well, uh, or hmm*, which are called fillers, are frequently inserted, and sometimes several articles and/or particles are deleted. Since speech recognizers recognize speech signals in accordance with linguistic rules, these fluctuations lead to recognition errors. Furthermore, fluctuations in pronunciation also affect speech recognizers. For example, even though a human may perceive that speech signals have been pronounced correctly, the computer analysis results often fluctuate for several reasons such as vowel omission or unstable vocal tract control. These fluctuations in acoustic and linguistic aspects of speech make recognition of spontaneous speech

difficult.

Conventionally, linguistic fluctuations of human speech are expressed by using probabilistic models called *language models*, and acoustic fluctuations are expressed separately by *acoustic models.*

However, as mentioned above, because the fluctuations that occur with spontaneous speech often appear in both the acoustic and linguistic aspects, the recognition of spontaneous speech by using such probabilistic models based on a divide-and-conquer strategy is considered to be difficult.

Deep learning techniques, which integrate signal processing models and acoustic models, have recently demonstrated significant improvement over conventional speech recognizers that have separate signal processing and acoustic models [1]. Deep learning suggests that optimizing several models in a unified way is important in order to overcome such difficult phenomena in spontaneous speech recognition. However, even with these advanced techniques, the fluctuations that span both the acoustic and linguistic aspects of speech have not yet been sufficiently expressed since deep learning techniques do not optimize linguistic aspects of recognizers.

In this article, we describe a method that enables joint optimization of the acoustic and language models of speech recognition, and we explain how this technique improves speech recognizers for spontaneous speech by focusing on speech recognition of a lecture video.

2. Weighted finite-state transducers

Weighted finite-state transducers (WFSTs) are commonly used as a core software component of several automatic speech recognizers including our proposed speech recognizers. WFSTs are abstract machines that represent rules to convert one type of sequence into another type of sequence, for example, to convert a speech feature sequence into a word sequence in automatic speech recognition. All the probabilistic models used in automatic speech recognizers can be expressed by using WFSTs, and therefore, WFSTs are used as a unified representation of automatic speech recognizers [2]. The speech recognition technique we describe in this article enabled joint acoustic and linguistic representation by extending WFSTs.

An illustration of an example WFST that converts phoneme sequences to word sequences is shown in **Fig. 1(a)**. The circles in the figure represent internal states of the abstract machine, and arrows indicate that the state may change in the direction of the arrow. The numerical values annotated to the arrows denote probabilities of the state transition corresponding to the arrow, and the symbols annotated to the arrow (for example, "ow"/"go") mean that the machine is expected to read the symbol "ow" from the input sequence during this state change, and to write the symbol "go" to the output sequence. The conversion process starts from the initial state (state 1), follows the arrow repeatedly while reading from the input sequence and writing to the output sequence, and ends when the state reaches the final state (state 7).

One of the main advantages of using WFSTs is that they have advanced composition algorithms. A WFST that accepts the word sequences that can be assumed as system inputs is shown in **Fig. 1(b)**. Even though this WFST actually does no conversion (i.e., it only outputs the same sequence as the input), this kind of probabilistic acceptance can also be represented in a WFST. The composition algorithm processes these two WFSTs (Figs. 1(a) and (b)) and constructs a composite WFST as in **Fig. 1(c)**. The composite WFST is constructed to represent the cascade connection of the input WFSTs. In other words, the WFST represents all possible conversion patterns obtained if the output sequences of the WFST in Fig. 1(a) are used as input sequences of the WFST in Fig. 1(b). The probability corresponding to each conversion pattern is simply denoted as a product of these two internal transductions.

The entire probabilistic process of automatic speech recognition can be represented by a large WFST that converts acoustic pattern sequences representing a short-time spectral pattern of acoustic signals to word sequences. This large WFST can be obtained by applying the composition algorithm to elemental WFSTs that convert acoustic patterns to interim representations called phoneme-states, the phoneme-states to phonemes, the phonemes to words (Fig. 1(a)), and the words to word sequences (Fig. 1(b)), respectively. Automatic speech recognition is subsequently achieved by finding the path on the WFST that maximizes that probability.

Estimating the elemental WFSTs (Figs. 1(a) and (b) in this case) is the central problem in constructing speech recognizers. Typically, these WFSTs are constructed by converting probabilistic models corresponding to each element into a WFST representation.

However, the strategy based on combining elemental WFSTs that are estimated separately cannot sufficiently express the phenomena in spontaneous

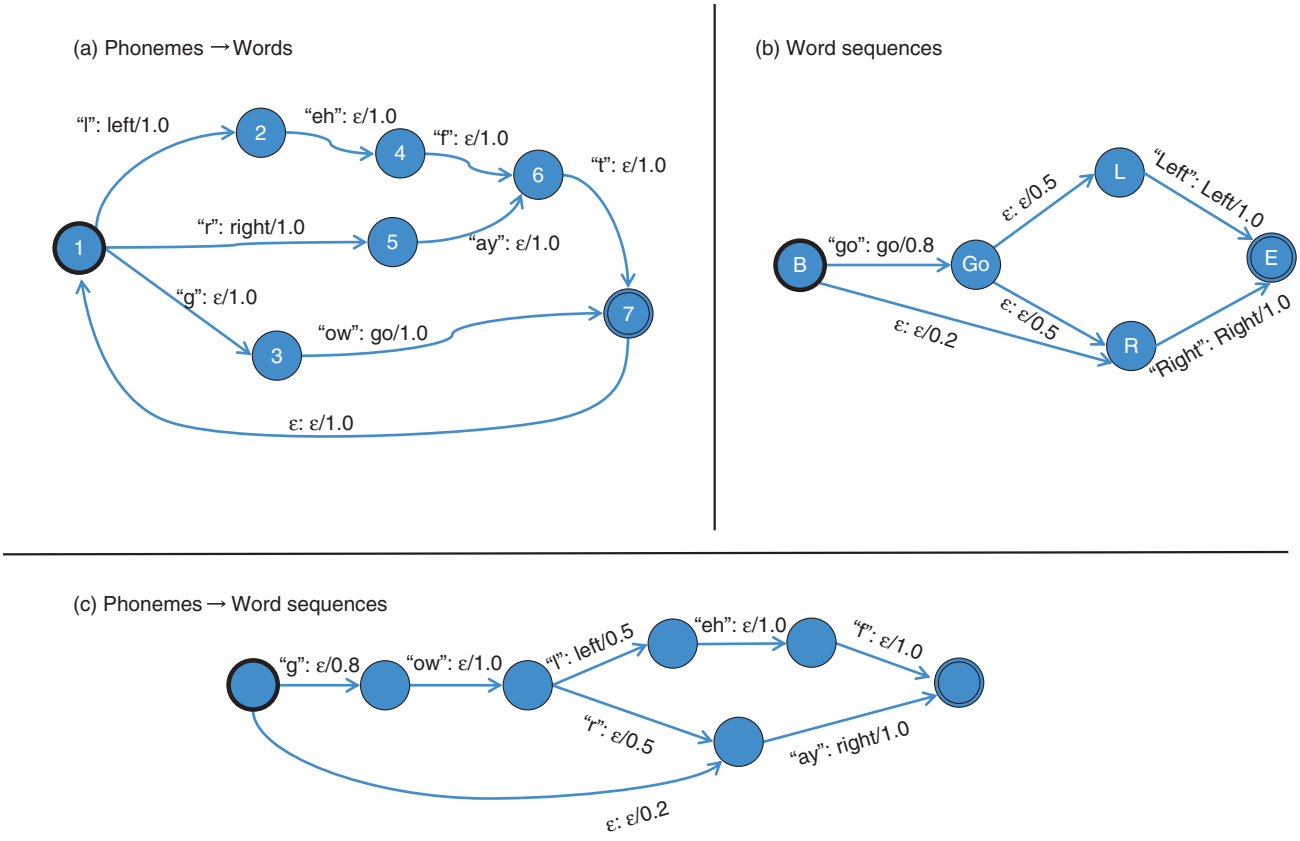

Fig. 1. Example of WFSTs.

speech because these phenomena often span several elemental WFSTs. Thus, it is necessary to consider the interdependency of these elemental WFSTs in order to model these fluctuations.

3. Acoustic model based on deep learning

Recently, researchers at Toronto University applied a method called *deep learning* to achieve accurate acoustic modeling. It has been demonstrated that deep learning methods can achieve accurate speech recognition without the need for complex acoustic pattern normalization techniques. Deep learning is a general term that refers to advances in the study of neural networks that have relatively deep architectures. Even though research on neural networks has been going on for over 30 years, the practical use of models with deep architectures was impossible before deep learning was developed.

The neural networks we focused on compute the output vector $y = (y_1, y_2, \ldots, y_{D'})^T$ with the given input vector $x = (x_1, x_2, \ldots, x_D)^T$ by using the following equation

$$y_j(x) = h_j^{(L)}(x),$$

$$h_j^{(\ell)}(x) = f\left(\sum_{i=1}^{D^{(\ell)}} w_{i,j}^{(\ell)}\ h_j^{(\ell-1)}(x) + b_j^{(\ell)}\right),$$

$$h_j^{(0)}(x) = x_i,$$

$$f(z) = \frac{1}{1 + e^{-z}},$$

where e is Napier's constant (the base of the natural logarithm).

By optimizing the parameters of the above equation ($w_{i,j}^{(\ell)}$ and $b_j^{(\ell)}$) so that the equation represents the given examples of x and y, we can use this equation to predict y that corresponds to the unseen example x. In automatic speech recognizers, x typically denotes the vectors that represent speech signals, and y typically denotes the probability of appearance of each acoustic pattern. The above equation includes L recursions, and L should be adjusted manually.

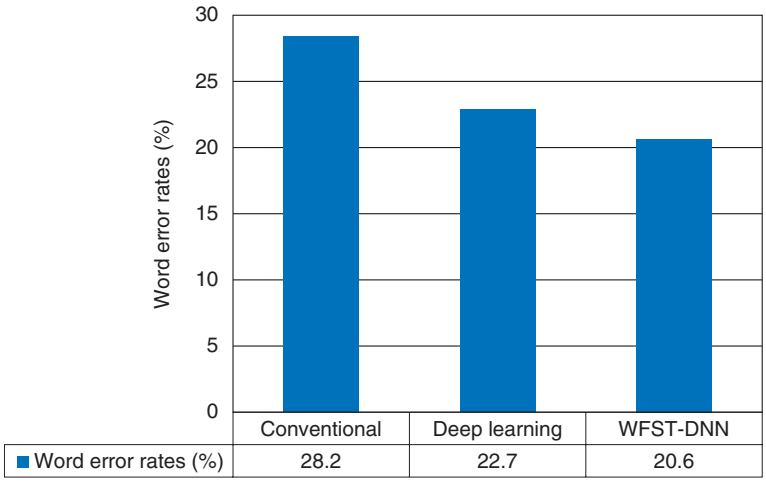

Fig. 2. Word error rates of three methods in English lecture speech recognition experiment.

Deep learning techniques enable optimization of neural networks with large L (such as $L > 3$) that are conventionally considered difficult to optimize. To enable such optimization, deep learning introduces an additional training procedure called pretraining. In this procedure, the parameters of neural networks are optimized so that the input vectors x in training examples are accurately expressed in the neural networks. Performing this pretraining procedure before the actual optimization procedure, which optimizes networks so that correspondence of x and y is expressed in the networks, makes it possible to optimize neural networks with large L, which are called deep neural networks (DNNs).

Since DNNs can be viewed as composite models of pattern correctors, the optimization of DNNs can be viewed as a method to achieve unified modeling of pattern normalization and classification.

4. Unified model based on deep learning

An entire conversion process from acoustic patterns to word sequences can be represented in a unified form by using large WFSTs. Furthermore, recent advances in deep learning have led to the development of a unified algorithm for acoustic pattern normalization and acoustic pattern classification. To take advantage of both approaches, we developed a unified modeling technique called WFST-DNN that integrates pattern normalization, pattern classification, acoustic models, and language models defined by unifying WFST and deep learning technologies [5].

In WFST-DNN, we enhanced the probabilities annotated to each arc of the WFSTs. In conventional methods, these probabilities are computed as a product of the probabilities of each elemental WFST. We enhanced these probabilities by defining them as an output of DNNs. Specifically, we defined and optimized y in the above equation to represent the probability of arcs. By applying this enhancement, the implicit assumption introduced in the conventional method, which is that the fluctuations appear independently in each elemental WFST, can be prevented. This enhancement is straightforward in that the speech fluctuations caused by phenomena of spontaneous speech span both the acoustic and language models. Further, since the proposed approach does not change the structure of conventional WFSTs, advanced methods, for example, computationally efficient recognition techniques developed for WFST-based speech recognizers can also be applied to a speech recognizer with the proposed technique.

We applied this method to a lecture recognition task and found that it performed better than a conventional method. The word error rates of the proposed method and the conventional method are shown in **Fig. 2.** Here, *conventional method* denotes a conventional state-of-the-art method before the introduction of deep learning techniques [6], and *deep learning* is of course the method based on deep learning. The word error rate achieved using the deep learning method was surprisingly high. WFST-DNN denotes the proposed method. It is clear from the results that the proposed method exhibited improved performance compared to the advanced conventional system and the results of the deep learning method.

5. Future outlook

The two main objectives of our future research are as follows. The first one is to achieve a deeper understanding of deep learning techniques. Deep learning techniques involve difficulties in terms of mathematical analysis, and therefore, the advantages of deep learning have only been shown through empirical and experimental results. However, investigating the advantages of deep learning and understanding how deep learning achieves advanced acoustic modeling would be very useful in order to apply these techniques in many other fields.

The second objective is to develop more computationally efficient techniques. Even though the current WFST-DNN recognizer can output recognition results in an acceptable time frame by exploiting graphic processing units (GPUs), it is important to be able to compute the results in a personal computer without fast GPUs. Parameter optimization requires a very long time even with acceleration based on GPUs. This long optimization time would also be problematic when customizing the system for a specific application.

We will pursue these objectives as we continue to investigate speech recognition techniques, with the ultimate goal of enabling application to various fields and achieving accurate recognition.

References

[1] F. Seide, G. Li, and D. Yu, "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks," Proc. of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011), Florence, Italy.

[2] T. Hori and A. Nakamura, "Speech Recognition Algorithms Using Weighted Finite-State Transducers," Morgan & Claypool Publishers, 2013.

[3] A.-R. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic Modeling Using Deep Belief Networks," IEEE Trans. on Audio, Speech, and Language Processing, Vol. 20, No. 1, pp. 14–22, 2012.

[4] C. M. Bishop, "Pattern Recognition and Machine Learning," Springer, 2006.

[5] Y. Kubo, T. Hori, and A. Nakamura, "Integrating Deep Neural Networks into Structured Classification Approach Based on Weighted Finite-State Transducers," Proc. of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH 2012), Portland, OR, USA.

[6] E. McDermott, S. Watanabe, and A. Nakamura, "Discriminative Training Based on an Integrated View of MPE and MMI in Margin and Error Space," Proc. of 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 4894–4897, Dallas, TX, USA.

Yotaro Kubo
Research Engineer, Media Information Processing Laboratory, NTT Communication Science Laboratories.

He received the B.E., M.E., and Dr.Eng. degrees from Waseda University, Tokyo, in 2007, 2008, and 2010, respectively. He was a visiting scientist at RWTH Aachen University, Aachen, Germany, from April to October 2010. In 2010, he joined NTT and has been with NTT Communication Science Laboratories. His research interests include machine learning and signal processing. He received the Awaya Award and the Itakura Prize Innovative Young Researcher Award from the Acoustical Society of Japan (ASJ) in 2010 and 2013, respectively, the Yamashita SIG Research Award from the Information Processing Society of Japan (IPSJ) in 2011, and the IEEE Signal Processing Society Japan Chapter Student Paper Award in 2011. He is a member of the International Speech Communication Association, ASJ, IPSJ, the Institute of Electronics, Information and Communication Engineers (IEICE), and IEEE.

Atsunori Ogawa
Research Engineer, Media Information Processing Laboratory, NTT Communication Science Laboratories.

He received the B.E. and M.E. degrees in information engineering, and the Ph.D. degree in information science from Nagoya University, Aichi, in 1996, 1998, and 2008, respectively. Since joining NTT in 1998, he has been engaged in research on speech recognition. He received the ASJ Best Poster Presentation Award in 2003 and 2006, respectively. He is a member of ASJ, IPSJ, IEICE, and IEEE.

Takaaki Hori
Senior Research Scientist, Signal Processing Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received the B.E. and M.E. degrees in electrical and information engineering and the Ph.D. degree in system and information engineering from Yamagata University in 1994, 1996, and 1999, respectively. He joined NTT in 1999 and began researching spoken language processing at NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories). He moved to NTT Communication Science Laboratories in 2002. He was a visiting scientist at the Massachusetts Institute of Technology, Cambridge, MA, USA, from 2006 to 2007. He received the 22nd Awaya Prize Young Researcher Award from ASJ in 2005, the 24th TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2009, the Kiyasu Special Industrial Achievement Award from IPSJ in 2012, and the 58th Maejima Hisoka Award from the Tsushinbunka Association in 2013. He is a member of ASJ, IEICE, and IEEE.

Atsushi Nakamura
Senior Research Scientist, Supervisor, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received the B.E., M.E., and Dr.Eng. degrees from Kyushu University, Fukuoka, in 1985, 1987, and 2001, respectively. In 1987, he joined NTT, where he engaged in R&D of network service platforms, including studies on application of speech processing technologies to network services at Musashino Electrical Communication Laboratories. From 1994 to 2000, he was a Senior Researcher at Advanced Telecommunications Research (ATR) Institute, Kyoto, Japan, where he was engaged in spontaneous speech recognition research, construction of a spoken language database, and development of speech translation systems. Since April 2000, he has been with NTT Communication Science Laboratories. His research interests include acoustic modeling of speech, speech recognition and synthesis, spoken language processing systems, speech production and perception, computational phonetics and phonology, and application of learning theories to signal analysis and modeling. He received the IEICE Paper Award in 2004, and twice received the Telecom-technology Award of The Telecommunications Advancement Foundation, in 2006 and 2009. He is a senior member of IEEE and serves as a member of the IEEE Machine Learning for Signal Processing (MLSP) Technical Committee, and as the Chair of the IEEE Signal Processing Society Kansai Chapter. He is also a member of ASJ and IEICE.