Coalescence Type Based Confidence Warping for Agglutinative Language Keyword Spotting

Ji Xu
The Key Laboratory of Speech Acoustics and Content Understanding,
Chinese Academy of Sciences, Beijing, China
Email: xuji010@gmail.com

Yeming Xiao, Jielin Pan and Yonghong Yan
The Key Laboratory of Speech Acoustics and Content Understanding,
Chinese Academy of Sciences, Beijing, China
Email: {xiaoyeming, panjielin, yanyonghong} @hccl.ioa.ac.cn

***Abstract*—In agglutinative languages like Korean, words are formed by joining l affix morphemes to the stem, which leads to high OOV rate in dictionary building. Hence, subword units are usually used as basic language modeling units in Large-Vocabulary Continuous Speech Recognition (LVCSR) or LVCSR based applications such as keyword spotting. In this work, firstly a new word property called coalescence type is introduced, which is defined based on the result of word segmentation process and thus unique for agglutinative languages. A confidence warping approach is then proposed to adjust confidence measure for keyword candidates, with the additional linguistic level information. An evaluation on Korean telephone speech keyword spotting task shows that up to 2% improvement can be obtained in precision, which is significantly better than the baseline system.**

***Index Terms*—speech recognition, keyword spotting, coalescence type, confidence warping.**

I. INTRODUCTION

Linguistic typology studies and classifies languages according to their structural and functional features. Describing the common properties and the structural diversity of the world's languages is the aim of linguistic typology. There are two major categories in linguistic typology. The first one is analytic language, in which grammatical relationships are conveyed without using inflectional morphemes. A representative example of analytic language is Mandarin Chinese, since it has almost no inflectional affixes to convey grammatical relationships. The second one is synthetic language. As contrast to analytic language, inflectional morphemes are used frequently in synthetic language, which result in a high morpheme-per-word ratio. There are three subclasses in synthetic language, inflecting language, agglutinative language and polysynthetic language, which are classified by the degrees of synthesis. Languages of inflecting language class use the fewest inflectional affixes among synthetic languages. Modern English and Latin are two examples of Inflecting languages. Agglutinative language has the middle number of inflectional affixes. Some examples of this class are Korean and Japanese. The class has the most inflectional affixes are polysynthetic languages, which includes some Native American languages. As agglutinative languages are the middle ones, there is a common tendency for agglutinative languages to exhibit synthetic properties.

Korean is a representative agglutinative language, in which most words are formed by relative freely joined morphemes. There are two new challenges in Korean LVCSR system, compared to analytic languages or less inflected languages. The first challenge is in language modeling. The units determined by natural boundaries of Korean writing system are not appropriate units for language modeling, since such units will lead to high language model perplexity. That is to say, if commonly used Korean words are selected to create a dictionary, it causes high out-of-vocabulary (OOV) rate. And if a dictionary contains Korean words as many as possible, language modeling suffers from sample sparseness. Currently, many researches have been focused on this problem and great progress has been made [1-4]. In [3], vocabulary units in Korean LVCSR systems are determined using a data-driven approach. In [4], merged morphemes are used as the recognition units and pronunciation-dependent entries, which are obtained by merging pairs of short and frequent morphemes into larger units based on a rule-based method and a statistical method. The second challenge is in acoustic modeling. When morphemes are joined to form a Korean word, the pronunciation of the first and last phonemes may be changed depends on the morphemes it concatenated to. This phenomenon is called co-articulation effect, which leads to low discriminative power and high complexity in acoustic modeling. To solve this problem, some researches based on allophones are presented. Research [5] focuses on five major constraints in the Korean language. Its conclusion shows that allophone derived by linguistic knowledge brings great improvement while using context independent units but not that efficient while using context dependent units.

doi:10.4304/jsw.9.10.2699-2705

Besides the negative effect brings by agglutination of morphemes in Korean, there is some useful information that can be utilized to enhance Korean LVCSR. This work gives an initial attempt in using such information. Coalescence degree, which is defined by the result of word segmentation process, is considered as a property of keyword being searched in keyword spotting. Since word segmentation is a standard step in agglutination language LVCSR only [6-9], coalescence degree is a property that will not be acquired by analytic language LVCSR. In this work, the relationship between confidence measure and coalescence degree is studied, and warping functions are used to compensate the confidence measure difference between different coalescence types.

This paper is organized as follows. Section 2 gives an introduction to language characteristic of Korean and the concept of coalescence degree. Section 3 presents the system framework and the warping functions we used in this work. The experiments are conducted in Section 4 and the conclusion is finally drawn in section 5.

II. Characteristic Of Korean Language

A. Characteristic of Korean Language

Korean is an agglutinative language which uses agglutination extensively. There are three levels under the level of Korean sentence, Jamo (Korean orthographic phoneme) [10], Eumjeol (Korean character) and Eojeol (Korean word). Each Eumjeol consists of three Jamos at most. As the Korean writing system is alpha-syllabary [11] the three Jamo are named initial consonant, vowel and final consonant in the pronunciation order, respectively. Eojeol is a unit delimited by space in Korean sentence. It consists of one or several Eumjeols.

Since Korean is an agglutinative language, the number of Eumjeols an Eojeol contains can be much larger than the number of characters contained by a word in analytic languages or less inflected languages. Figure 1 shows some examples of Korean Eojeols. It can be seen that although some Eojeols are similar to the ones in analytic languages or less inflected languages, particular Eojeol can express a very complicated meaning even a whole sentence. Therefore, dictionary building becomes a problem in Korean language modeling as previously mentioned.

TABLE I.
Example Of Korean Eojeol

Korean Eojeol	English translation
시간	time
회사	company
다음주	cellphone
다큐멘터리	documentary
전화할테니까	I'll call you.
그저께갔습니다	I went the day before yesterday.

B. Word Segmentation and Coalescence Type

Word segmentation is introduced by previous researches to create appropriate dictionary units in agglutinative language LVCSR. Sub-word units like morphemes are generated using those approaches. This work adopts an unsupervised word segmentation method proposed in [12]. In this approach, a Dirichlet process model trained using Bayesian inference through block Gibbs sampling is utilized [13,14]. Thus, all Eojeols including dictionary units, corpora and keywords etc. are firstly segmented into morpheme like units before entering LVCSR system.

Coalescence degree is a new word property introduced by this work. It is defined by the quantity and length of sub-word units decomposed from a word through word segmentation process. Coalescence degree is unique for words of agglutinative languages since word segmentation step cannot be applied to analytic language LVCSR. Coalescence degree includes three main types: Total coalescence (Type I), Non coalescence (Type II) and Partial coalescence (Type III). Total coalescence refers to Eojeols that consist of only one morpheme or Eojeols which are fixed collocations themself. Eojeols with this coalescence type keep intact in word segmentation process. Non coalescence is opposite to Total coalescence, Eojeols with this coalescence type consist of several morphemes which are not close related, and are segmented into single Emujeols in word segmentation process. Partial coalescence is between the two extremes, Eojeols with this coalescence type are

TABLE II.
Examples Of Coalescence Type

Eojeol	Word Segmentation	Coalescence Type
그거	그거	Total Coalescence (Type I)
그러니까	그러니까	Total Coalescence (Type I)
그문	그 문	Non Coalescence (Type II)
시내로	시 내 로	Non Coalescence (Type II)
그래갖고	그래 갖고	Partial Coalescence (Type III)
가방만드는데	가방 만드는 데	Partial Coalescence (Type III)

segmented into a combination of Emujeols and morphemes in word segmentation process. Table 2 gives some examples of Eojeols and their coalescence types.

In particular, coalescence degree is related to the word segmentation method used in LVCSR system. That is to say, the same Eojeol can have different coalescence types when word segmentation method is changed. But in general, Eojeols that are commonly used usually have relative fixed coalescence types.

III. Confidence Warping

A. System Framwork

The framework of keyword spotting system with confidence warping is shown in Figure 1. In this framework, feature extraction and decoding belong to standard speech recognition steps, a lattice with rich information is generated for keyword searching [15-17]. Confidence rescoring is used after keyword candidate is generated [18,19], which modifies the confidence measure with extra information from acoustic level or linguistic level.

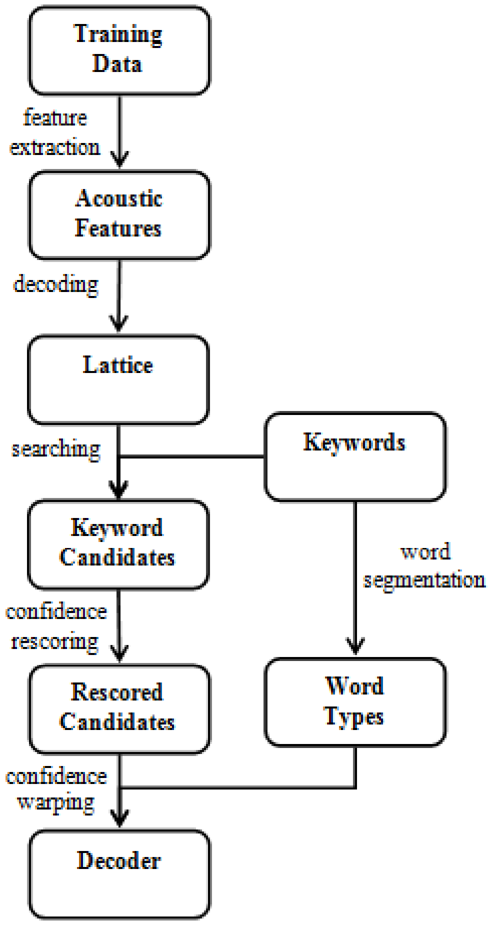

Figure 1. System framework.

The confidence warping approach proposed by this paper follows the step of confidence rescoring. The aim of this approach is to make the spotting precision of each coalescence type to be similar in a specific confidence region (operation region usually), while the total candidates of all types should not be modified. Although this goal can be realized by various implementations, in order to make the model simplified, keyword candidates of all coalescence types are using a unified warping function with only one parameter. The parameter is adjusted according to a training process using the data of develop set. In this process, the following criteria should be fulfilled:

$$\sum_{i=1}^{I}\sum_{k_i=1}^{N_i}\varphi\left(f(M_{k_i},\alpha_i)\right)=\sum_{i=1}^{I}\sum_{k_i=1}^{N_i}\varphi(M_{k_i}) \qquad (1)$$

where $i\in(1\cdots I)$ represents the coalescence type; $k_i\in(1\cdots N_i)$ represents the kth candidate of coalescence type i; f is the warping function and α_i is the corresponding parameter of coalescence type i. The Boolean function φ represents if confidence M belongs to a specific region θ, as following:

$$\varphi(M)=\begin{cases}1 & M\in\theta\\ 0 & M\notin\theta\end{cases} \qquad (2)$$

In practical, coalescence types are firstly obtained through word segmentation process, the warping function is then applied to compensate the confidence distribution difference between different coalescence types. Theoretically, as confidence warping modifies confidence measure using extra information, it can be considered as a kind of confidence rescoring. But since a warping function is used in our approach and this approach can coexist with any other confidence rescoring approaches, we refer the approach proposed in this work as an independent step, which is called confidence warping.

B. Confidence Warping

Two warping functions are selected for study in this paper, linear warping function and bilinear warping function.

Linear warping is a simple and straightforward warping function. It modifies the confidence measure linearly. For each coalescence type, a corresponding bias is added to the original confidence measure, which can be expressed as equation (3):

$$CM_{Type}^{warp}=CM_{Type}^{ori}+Bias_{Type} \qquad (3)$$

where CM_{Type}^{ori} is the original confidence measure for a certain coalescence type, and CM_{Type}^{warp} is the confidence measure after warping. $Bias_{Type}$ is the coalescence type based bias which determined by experiments.

Bilinear warping modifies the confidence measure in a certain region. While using this warping function, the confidence measure in the endpoints of the region stay the same after warping, but the confidence measure in the midpoint of the region changed greatly. If we use CM_{LL} and CM_{HL} to represent the lower and upper limits of confidence measure of keyword candidates, which are generated by a certain system, the unified confidence measure UCM can be calculated as equation (4):

$$UCM = -\frac{CM_{HL} - CM}{CM_{HL} - CM_{LL}} \cdot \pi \qquad (4)$$

The bilinear warping is then applied to the unified confidence measure, according to equation (5):

$$UCM_{Type}^{warp} = \tan^{-1} \frac{(1 - \alpha_{Type}^{2}) \sin UCM_{Type}^{ori}}{(1 + \alpha_{Type}^{2}) \cos UCM_{Type}^{ori} - 2\alpha_{Type}} \qquad (5)$$

where UCM_{Type}^{ori} and UCM_{Type}^{warp} are UCM before and after warping, α_{Type} is the warp factor. Figure 2 gives an intuitive understanding of bilinear warping function with different warping factors.

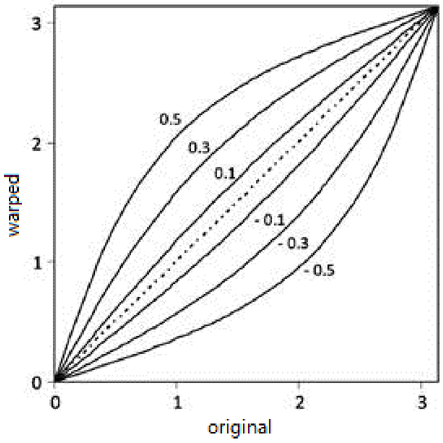

Figure 2. Bilinear warping.

VI. Experiments

Experiment conducted by this paper can be grouped into two parts. For the first part, the relationship between coalescence type and confidence measure distribution is studied and an analysis of characteristics of different coalescence types are given. For the second part, keyword spotting results are studied directly, core indicators like spotting precision are listed to reflect the actual performance of confidence warping.

A. *Experiment Setup*

Our experiments are conducted on a recorded dataset of telephone speech. The dataset includes three different parts, training set, develop set and test set, which are summarized in Table 3. The training set is used for model training. The develop set is used for adjusting parameters of different coalescence types. And the test set is used for performance evaluation. The text corpus used for language modeling is the labels of the training set plus some texts downloaded from internet. The downloaded texts are mostly from several famous BBS in Korean.

The phoneme set used in this paper are the same as the one used in [5], which is also used in [20,21]. Other phoneme set like IPA symbols are used in other systems [3], which is slightly modified from the one used in this work. Triphone HMMs are used for all the experiments conducted in this paper. The parameter of the acoustic model is 6000 states with 16 Gaussian mixtures for each state. Trigram language models built by SRILM Toolkit are adopted [22].

TABLE III.
The Data Set Used in Experiments

	Length	Sample Rate	Quantization	Channel
Training set	187h	8kHz	16bit	single
Develop set	36.0h	8kHz	16bit	single
Test set	36.7h	8kHz	16bit	single

B. *Confidence Measure Distribution of Different Coalescence Types*

Experiments of this section aim at study the confidence measure distribution of different coalescence types. Therefore, 50 keywords are selected for each coalescence type to draw a curve of keyword precision and confidence measure threshold. In the preliminary experiment, we found that the confidence measure distribution of coalescence Type II is a combination of two sub-distributions, which is defined by the length of keyword in Type II. Thus, we define two sub-classes Type II- and Type II+, which represents the length of keyword is equal to 2 (Type II-) or longer than 2 (Type II+). The keywords are reselected for the two sub-classes and make its quantity equal to other classes. That is to say, the total number of keywords selected for our experiments is 200. The warping function parameters of different coalescence types are then adjusted in the constraint of equation (1).

The curve which describes the relationship of keyword precision and confidence measure threshold is shown in Figure 3. It can be seen that the precision of Type II+ and Type III are much higher than Type I and Type II- in the same confidence measure threshold. For example, in the threshold of -1, the precision of Type I, Type II-, Type II+ and Type III are 48.74%, 56.10%, 86.42% and 81.00% respectively. Thus, it is not suitable to use a unified threshold for all the keywords, which will end in a terrible result. The curves after applying the confidence measure warping functions is shown in Figure 4 and Figure 5, for linear warping and bilinear warping respectively. It is not hard to see that the curves in Figure 4 and Figure 5 are closer than the ones in Figure 3.

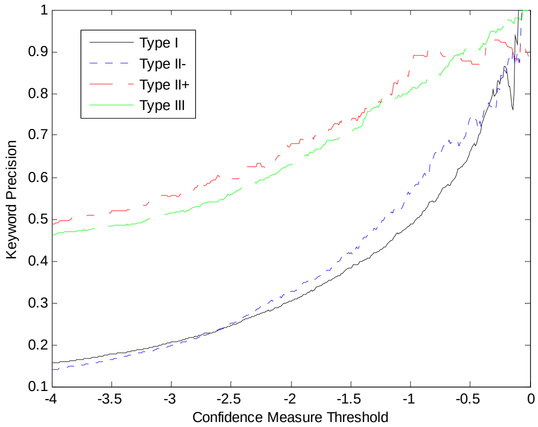

Figure 3. CM distribution of different coalescence types: Original.

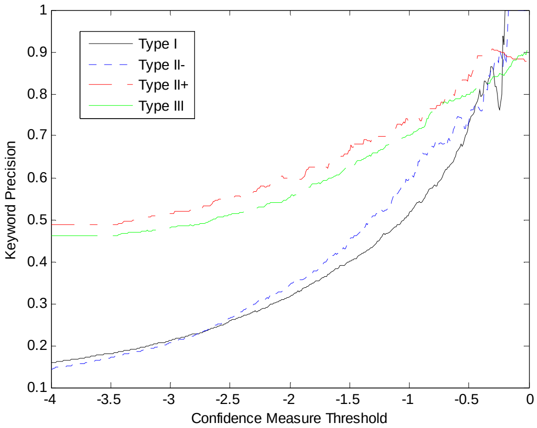

Figure 4. CM distribution of different coalescence types: linear warping.

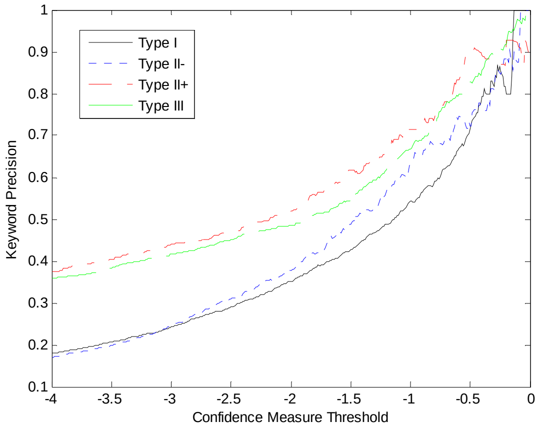

Figure 5. CM distribution of different coalescence types: bilinear warping.

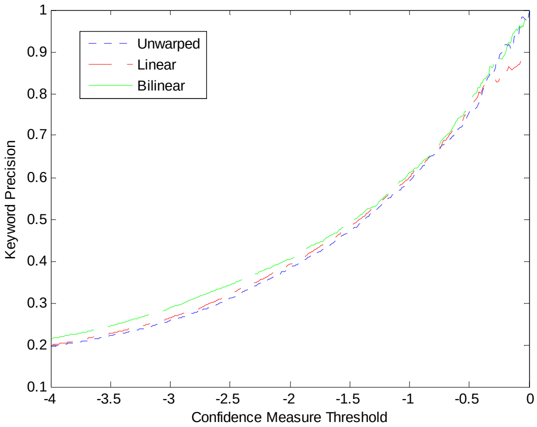

Figure 6. CM distribution of merged keyword list.

The curve of keyword precision and confidence measure threshold for merged keyword list (200 keywords) is shown in Figure 6, which reflects the overall distribution of the keyword spotting system. It is obvious to see that both linear warping and bilinear warping obtain higher precisions in regular operation region around -0.5, but the precision of linear warping is increased slowly when the threshold is above -0.3. The reason is that the confidence region after linear warping is not consistent with the original one, which leads to some confidence measures of warped keywords overflow the original confidence limits. But as the bilinear warping function does not have this problem, it obtains precision improvement in the whole region.

C. *Keyword Spotting Experiments*

Besides the confidence measure distributions studied in the previous section, core indicators like recall and keyword precision are more concerned by practical users. Therefore, another group of experiments is conducted in this section to show the performance of confidence warping on such indicators.

The targets studied in this experiment are 4 different types of keyword lists plus the mixed list in original and warped manner. The results are listed in Table 4. From the result, we can see that Type I has the best recall while Type II+ has the worst. This can be easily explained that Type I is the ones that keep intact in word segmentation process, so it belongs to in-dictionary keywords which usually have a good recall. And for Type II+, which is rather long and unfamiliar, a bad recall is indeed not surprising. The precision result shows that Type III has an outstanding high precision compared to all other Types. The reason is that although keywords of Type III belong to OOV, they are OOV that are familiar to the word segmentation model. Thus, they probably have a

TABLE IV.
KEYWORD SPOTTING EXPERIMENTS RESULT

	Total recall	Precision at 20% recall	Precision at 10% recall
Type I	78.50%	79.63%	86.00%
Type II-	42.66%	56.04%	73.49%
Type II+	32.41%	77.27%	92.86%
Type III	56.57%	97.70%	97.73%
Mix	55.69%	81.58%	91.44%
Mix Linear	55.69%	81.97%	90.37%
Mix Bilinear	55.69%	83.58%	92.22%

precision similar to in-dictionary keywords. Besides, through the definition of Type III, a keyword belongs to this type at least consists of three Emujeols and very possible to be a long keyword. It is known that long keywords usually have fewer false alarms due to the difficulty of matching. Therefore, keywords of Type III finally have a very high precision.

For the original and warped results of mixed keyword lists, bilinear warping obtained 2% precision improvement at 20% recall and 0.78% at 10% recall. But linear warping only obtained 0.39% precision improvement at 20% recall and decreased more than 1% at 10% recall. It is obvious to see that bilinear warping performs well than linear warping, and also more stable.

V. CONCLUSION

In this paper, a new idea of coalescence type is introduced in order to provide extra information for agglutinative language keyword spotting. Based on coalescence type, a novel confidence measure warping approach is presented to make the confidence measure distribution of different keywords to be similar in the operation region. The experiments show that both linear and bilinear warping functions have the ability to optimize the overall confidence measure distribution. And for practical indicators, the bilinear warping function over the linear warping function obtains the best improvement, which is 2% higher than the original one.

ACKNOWLEDGMENT

This work is partially supported by the National Natural Science Foundation of China (Nos. 10925419, 90920302, 61072124, 11074275, 11161140319, 91120001, 61271426), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant Nos. XDA06030100, XDA06030500), the National 863 Program (No. 2012AA012503) and the CAS Priority Deployment Project (No. KGZD-EW-103-2).

REFERENCES

[1] Lee, H. S., J. Park, and H. R. Kim. "An implementation of Korean spontaneous speech recognition system." 1996 International Conference on Signal Processing Applications & Technology. 1996.

[2] Kwon, Oh-Wook, Kyuwoong Hwang, and Jun Park. "Korean large vocabulary continuous speech recognition using pseudomorpheme units." *Eurospeech*. 1999.

[3] D. Kiecza, T. Schultz, and A. Waibel, “Data-driven determination of appropriate dictionary units for Korean speech recognition,” in Proc. ICSP, Seoul, Korea, 1999, pp. 323–327.

[4] O.-W. Kwon and J. Park, “Korean large vocabulary continuous speech recognition with morpheme-based recognition units,” Speech Communication, vol. 39, pp. 287–300, 2003.

[5] H. Hong, S. Kim, and M. Chung, "Effects of Allophones on the Performance of Korean Speech Recognition,” in Proc. of Interspeech 2008, Australia, pp. 2410 – 2413, 2008.

[6] Carki, Kenan, Petra Geutner, and Tanja Schultz. "Turkish LVCSR: towards better speech recognition for agglutinative languages." *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*. Vol. 3. IEEE, 2000.

[7] Vergyri, Dimitra, et al. "Morphology-based language modeling for arabic speech recognition." *INTERSPEECH*. Vol. 4. 2004.

[8] Hirsimäki, Teemu, et al. "Unlimited vocabulary speech recognition with morph language models applied to Finnish." Computer Speech & Language 20.4 (2006): 515-541.

[9] Bao, Feilong, et al. "Segmentation-based Mongolian LVCSR approach." Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013.

[10] Lee J., Kim B., Lee G.G “Hybrid approach to grapheme to phoneme conversion for Korean,” In Proc. of Interspeech 2009, Brighton, pp. 1291–1294.

[11] Taylor, I., “The Korean writing system: An Alphabet? A Syllabary? A Logography?”, In: Proc. of Visible Language 2, pp. 67-82, 1980.

[12] Sakriani Sakti, Andrew Finch, Ryosuke Isotani, Hisashi Kawai and Satoshi Nakamura, “Unsupervised determination of efficient Korean LVCSR units using a Bayesian Dirichlet process model” in Proc. ICASSP, Prague, Czech Republic, 2011.

[13] Ferguson, Thomas S. "A Bayesian analysis of some nonparametric problems." The annals of statistics (1973): 209-230.

[14] Neal, Radford M. "Markov chain sampling methods for Dirichlet process mixture models." *Journal of computational and graphical statistics* 9.2 (2000): 249-265.

[15] James, David A., and Steve J. Young. "A fast lattice-based approach to vocabulary independent wordspotting." Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on. Vol. 1. IEEE, 1994.

[16] Saraclar, Murat, and Richard Sproat. "Lattice-based search for spoken utterance retrieval." *Urbana* 51 (2004): 61801.

[17] Zhou, Zheng-Yu, et al. "Towards spoken-document retrieval for the internet: lattice indexing for large-scale web-search architectures." *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics, 2006.

[18] Wessel, Frank, et al. "Confidence measures for large vocabulary continuous speech recognition." *Speech and Audio Processing, IEEE Transactions on* 9.3 (2001): 288-298.

[19] Evermann, Gunnar, and Philip C. Woodland. "Large vocabulary decoding and confidence estimation using word posterior probabilities." Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on. Vol. 3. IEEE, 2000.

[20] Sakriani Sakti, Ryosuke Isotani, Hisashi Kawai, Satoshi Nakamura, “Utilizing a Noisy-Channel Approach for Korean LVCSR,” In Proc. of Interspeech 2010, Makuhari, Chiba, Japan

[21] M. Kim, Y.R. Oh, and H.K. Kim, “Non-native pronunciation variation modeling using an indirect data driven method,” in Proc. ASRU, Kyoto, Japan, 2007, pp. 231–236.

[22] Stolcke, Andreas. "SRILM-an extensible language modeling toolkit." *INTERSPEECH*. 2002.

Ji Xu received his B.E. and Master's degrees in 2008 and 2011 respectively, from the Department of Electronic Engineering, Tsinghua University. Now he is a doctor candidate of the Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics (IOA), Chinese Academy of Sciences (CAS). His research interests include Large Vocabulary Continuous Speech Recognition, Deep Learning and Speech synthesis.

Yeming Xiao received his B.E. degree in school of Electronic and Information Engineering from Beihang University in 2008. Now he is a doctor candidate of the Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics (IOA), Chinese Academy of Sciences (CAS). His research interests include Large Vocabulary Continuous Speech Recognition, Neural Networks, Deep Learning and Machine Learning.

Jielin Pan received his B.S. from Peking University in 1986 and his Master degree in Electronic Engineering from Tsinghua University in 1989. From Jan. 2000 to July 2001 he was working in Intel China Research Center as a senior researcher and speech recognition group manager. In Dec. 2002 he joined the Key Laboratory of Speech Acoustics and Content Understanding as a professor. His current research includes: LVCSR, speech analysis, acoustic model and search algorithm.

Yonghong Yan received his B.E. from Tsinghua University in 1990, and his PhD from Oregon Graduate Institute (OGI). He worked in OGI as Assistant Professor (1995), Associate Professor (1998) and Associate Director (1997) of Center for Spoken Language Understanding. He worked in Intel from 1998-2001, chaired Human Computer Interface Research Council, worked as Principal Engineer of Microprocessor Research Lab and Director of Intel China Research Center. Currently he is a professor and director of the Key Laboratory of Speech Acoustics and Content Understanding. His research interests include speech processing and recognition, language/speaker recognition, and human computer interface. He has published more than 100 papers and holds 40 patents.