UNSUPERVISED DETERMINATION OF EFFICIENT KOREAN LVCSR UNITS USING A BAYESIAN DIRICHLET PROCESS MODEL

Sakriani Sakti, Andrew Finch, Ryosuke Isotani, Hisashi Kawai, Satoshi Nakamura

Spoken Language Communication Research Group, MASTAR Project,
National Institute of Information and Communications Technology (NICT), Japan
{sakriani.sakti,andrew.finch,ryosuke.isotani,hisashi.kawai,satoshi.nakamura}@nict.go.jp

ABSTRACT

Korean is an agglutinative language that does not have explicit word boundaries. It is also a highly inflective language that exhibits severe coarticulation effects. These characteristics pose a challenge in developing large-vocabulary continuous speech recognition (LVCSR) systems. Many existing Korean LVCSR systems attempt to overcome these difficulties by defining a set of "word" units using morphological analysis (rule-based) or statistical methods. These approaches usually require a great deal of linguistic knowledge or at least some explicit information about the statistical distribution of the units. However, exceptions or uncommon words (e.g., foreign proper nouns) still exist that cannot be covered by rules alone. In this paper, we investigate the use of an unsupervised, nonparametric Bayesian approach to automatically determining efficient units for a Korean LVCSR system. Specifically, we utilize a Dirichlet process model trained using Bayesian inference through block Gibbs sampling. Our approach provides a principled way of learning units without explicit linguistic knowledge or any static parameters. Experiments were conducted on a travel domain corpus, which includes many foreign words and proper nouns. In our experiments we compared our method to a set of state-of-the-art baseline systems that relied on either morphological analysis or segmentation heuristics. Our system was able to produce a considerably more compact set of "word" units than the best baseline system (the lexical dictionary was approximately half the size), with a recognition accuracy 5.89% higher in terms of the relative word error rate than the best baseline system.

Index Terms— Korean language, large-vocabulary continuous speech recognition, unsupervised segmentation, nonparametric Bayesian approach, Dirichlet process model, Gibbs sampling.

1. INTRODUCTION

Most state-of-the-art large-vocabulary continuous speech recognition (LVCSR) systems typically choose words as the basis for recognition units. This is basic for Indo-European languages (e.g. English), since the number of word forms is relatively small, and the boundaries between adjacent words are clearly separated by a white space. However, this choice becomes problematic in languages that do not have explicit word boundaries like Korean. Although a space exists in the Korean writing script, it is used to separate two adjacent word-phrases (*eojeol*), which generally correspond to two or three words in English in a semantic sense. This word-phrase is represented by one or more *Hangul* characters of an orthographic syllable (*eumjeol*) unit, where in a linguistic sense it is basically an agglomerate of morphemes. This agglutinative process may combine one or more stem morphemes with one or more functional morphemes (e.g., tenses, suffixes, or honorifics). Consequently, there may be thousands of distinct *eojeol* that can be generated from a given word root depending on their usage. Thus, developing a LVCSR system with *eojeol* as the basic recognition unit leads to high language model perplexity and out-of-vocabulary (OOV) rates. On the other hand, choosing an *eumjeol* as the basic recognition unit results in high acoustic confusability because of the severe phonological phenomena and coarticulation effects. A more detailed discussion of Korean phonological phenomena can be found in [1, 2].

Many existing Korean LVCSR systems attempt to overcome these difficulties by creating a set of new units that lie between these two *eojeol* and *eumjeol* units. One response is to choose a morpheme as a basic recognition unit, and this approach has often been used in many agglutinative languages [3, 4]. One study [5] shows that this morpheme-based approach still requires an additional cross-word phone variation lexicon to deal with the severe coarticulation problem. Another study [6] has proposed merging several morphemes into a basic unit and defining it as a word. Starting from the original morpheme units defined in Korean morphology, pairs of short and frequent morphemes are merged into larger units by combining both rule-based and statistical methods. Another method [7], is to determine appropriate vocabulary units using a data-driven approach in which the consecutive units are merged based on the frequency of their pronunciation transition. However, these approaches either require a great deal of linguistic knowledge or at least some explicit information about the statistical distribution of the units which is often difficult to estimate for uncommon foreign words or proper nouns.

This paper investigates the use of an unsupervised, nonparametric Bayesian approach to automatically determine efficient word-units for Korean LVCSR systems. Specifically, we utilize a Dirichlet process model [8] trained using Bayesian inference through block Gibbs sampling [9]. This unsupervised approach has been known in Bayesian statistics for more than three decades, but has only recently gained attention in the natural language processing field [10, 11, 12]. Consequently, investigations into its use for improving LVCSR performance are still very rare. The advantage of this approach is that it provides a principled way of learning units without explicit linguistic knowledge, while controlling the efficiency level of the resolution, which corresponds to unit distribution in the data. The term "nonparametric" here means that the model does not have a fixed set of parameters. It naturally generates a compact set of units without the overfitting problems typically encountered when using maximum likelihood training. Thus, in this study, the number of Korean units is learned along with their identities. To properly model the phonological phenomena and coarticulation effects of the units, we apply the proposed approach to joint sequences consisting of Korean orthography together with its pronunciation.

In the following sections, we describe the Bayesian Dirichlet process model and the inferencing process using Gibbs sampling. Then, details on the experiments are presented in Section 4 and conclusions are drawn in Section 5.

2. BAYESIAN DIRICHLET PROCESS MODEL

2.1. The General Framework

A Dirichlet process is a prior used in nonparametric Bayesian models of data [8]. It is a stochastic process that can be thought of as a probability distribution whose domain is itself a random distribution. Each draw from a Dirichlet process is itself a distribution. We denote a Dirichlet process as $DP(\alpha, G_0)$ with two arguments: α and G_0, where $\alpha \in \mathbb{R}$ is called the concentration parameter, and G_0 is a base distribution.

Assuming we are required to determine K random variables $\theta = [\theta_1, \theta_2, ..., \theta_K]$, where each θ_k is distributed according to G, and G itself is a random measure drawn from a Dirichlet process

$$\begin{aligned} \theta_k|G &\sim G \\ G|_{\alpha,G_0} &\sim DP(\alpha, G_0) \end{aligned} \tag{1}$$

Intuitively, G_0 is basically the mean of the DP and thus $E[G] = G_0$, while α controls the variance of G. The larger α is, the smaller the variance, and the DP will concentrate more of its mass around the mean or in other words G will be similar to G_0.

To determine each θ_k, we never deal with G directly, since it is represented by an infinite-dimensional Dirichlet distribution. Instead, we sample from G by casting the problem as a Chinese restaurant process (CRP) [13]. In this paradigm, one considers a restaurant having an infinite number of tables, each with infinite seating capacity. Every customer who enters the restaurant chooses a table according to the following random process

$$P(\theta_k|\theta_{-k}) = \frac{n(\theta_k) + \alpha G_0(\theta_k)}{N + \alpha}, \tag{2}$$

where N is the number of customers so far, $n(\theta_k)$ is the number of customers already sitting at the kth table and $G_0(\theta_k)$ is the probability of generating a new table θ_k. In this process,

- The first customer always chooses the first table.
- The n_{th} customer chooses:
  - a new or the first unoccupied table with probability $\frac{\alpha G_0(\theta_k)}{N+\alpha}$, or
  - the occupied table θ_k with probability $\frac{n(\theta_k)}{N+\alpha}$.

Finally, the probability of the sequence of random variables $\theta = [\theta_1, \theta_2, ..., \theta_K]$ is obtained by the product of K observations:

$$P(\theta_1^K) \approx \prod_{k=1}^{K} P(\theta_k|\theta_{-k}) \tag{3}$$

2.2. DP Model for Determining Korean LVCSR Units

Applying the DP model for our Korean LVCSR task, we focus on the determination of the Korean "word" unit sequence given a joint sequence $\langle c, p\rangle$ of *Hangul* characters of the Korean orthographic syllables (*eumjeol*) sequence $c = [c_1, c_2, ..., c_N]$ and the corresponding phonemic syllables of the actual pronunciation sequence $p = [p_1, p_2, ..., p_M]$. In most cases, the length of *eumjeol* sequence is equal to the length of the phonemic syllables ($M = N$). Thus, we can simplify our task by defining the joint sequence as a tuple of $\langle c, p\rangle = [\langle c, p\rangle_1, ..., \langle c, p\rangle_N]$, where the goal is to determine word units $W\langle c, p\rangle = [W_1\langle c, p\rangle, ..., W_K\langle c, p\rangle]$.

Following the theoretical framework described in the previous section, random variables θ_k are currently our Korean "word" unit $W_k\langle c, p\rangle$, and G is a discrete probability distribution over all possible word units according to a Dirichlet process prior. The restaurant tables in the CRP paradigm correspond to the generated lexical entries, and the seating arrangement thus specifies a distribution over Korean "word" units with each customer representing one token unit. The base distribution G_0 is the prior probability over words and α controls the generation of novel word units (to avoid overfitting problems). Here, we use a *spelling model* that assumes that the word unit $W_k\langle c, p\rangle$ with word length L has a Poisson distribution with a mean λ. The probability of a new word unit $W_k\langle c, p\rangle$ is therefore assigned according to the following distribution:

$$G_0(W_k\langle c, p\rangle) = \frac{\lambda^L}{L!} e^{-\lambda} u^{-L} \tag{4}$$

where u is the vocabulary size of all *eumjeol* character-pronunciation tuples $\langle c, p\rangle$ in the document.

3. GIBBS SAMPLING INFERENCE

After defining our generative model, it is important to determine the posterior distribution $P(H|\Theta)$ of our hypothesis H given the data corpus Θ. However, it is generally impossible to find the most likely segmentation of Korean LVCSR units from among all possible segmentations given the data using exact inference in our nonparametric Bayesian model, because the hidden variables of the word segmentation do not allow for exact computation of the integrals.

Instead, we apply a block-wise version of Gibbs sampling, as motivated by the work in [11]. Gibbs sampling is one of the simplest Markov chain Monte Carlo (MCMC) algorithms [14], in which word segmentations are repeatedly sampled (in this case, block-wise for each sentence) from their conditional posterior distribution given the current values of all other variables in the model. The sampling algorithm for our study is shown in Alg. 1.

Algorithm 1: The block Gibbs sampling algorithm

```
foreach i=1 to NumIterations do
    foreach sentence s ∈ randperm(S) do
        if i > 1 then
            Remove customers of W⟨c,p⟩ exist in s from Θ
        end
        Generate all possible candidates W⟨c,p⟩ in s;
        foreach candidate W⟨c,p⟩ do
            Compute probability P(W⟨c,p⟩|Θ) using DP
            model (Eq. 2-4)
        end
        Draw W⟨c,p⟩ according to P(W⟨c,p⟩|Θ);
        Add customers of W⟨c,p⟩ to Θ
    end
end
```

In practice, block-wise sampling is done for each *eojeol* to avoid segmentations that cross word-phrase (*eojeol*) units. The computation of the probability of all possible $P(W\langle c, p\rangle|\Theta)$, as well as the sampling of $W\langle c, p\rangle$ according to $P(W\langle c, p\rangle|\Theta)$, is implemented using the forward filtering/backward sampling (FFBS) dynamic programming algorithm [11, 12]. The FFBS algorithm operates directly on the segmentation graph (each node represents a set of partial segmentation hypotheses, and each arc is labeled with the probability of adding a segment to the set hypotheses), and has two steps. The *forward filtering* step, calculates for each node in the graph the total probability of a subgraph leading to that node. The *backward sampling* step samples a derivation of the $P(W\langle c, p\rangle|\Theta)$ according to $P(W\langle c, p\rangle|\Theta)$, using values stored in the graph by the forward

filtering process. The sampling procedure is performed backwards staring from the sink node of the graph, and the procedure is applied recursively on the tail of each sampled arc until the source node of the graph is reached. A more detailed explanation of the application of the FFBS algorithm to bilingual sequence co-segmentation can be found in [12].

4. EXPERIMENTAL EVALUATION

4.1. Corpora

The experiments were conducted on the travel domain using the Korean Basic Travel Expression Corpus (BTEC) [15] which contains many foreign words and proper nouns (eg. names of famous places, restaurants or streets). The BTEC text material used for training consists of about 900,000 sentences. The available BTEC speech material, which has been developed by the Electronics and Telecommunication Research Institute (ETRI) in Korea, only contains 510 BTEC sentences spoken by 40 speakers (20 males, 20 females). We therefore used it for evaluation.

The training speech material used was based on the large-vocabulary continuous Korean speech database developed by the Speech Information Technology and Industry Promotion Center (SiTEC) [16]. It consists of: (1) phonetically-balanced sentences selected from a large Korean text corpus that contain high frequency morphemes; and (2) dictation application sentences containing words and morphemes of high frequency that were generated for a dictation application. There are about 200 speakers (100 males, 100 females) for the phonetically-balanced sentences and 800 speakers (400 males, 400 females) for the dictation application sentences. Each speaker uttered about 100 sentences, resulting in a total of 100,000 utterances (about 70 hours of speech). Orthographic transcriptions annotated with pronunciation were available for the whole corpus.

4.2. Baseline LVCSR System

A sampling frequency of 16 kHz, a frame length of a 20-ms Hamming window, a frame shift of 10 ms, and 25 dimensional feature parameters consisting of 12-order MFCC, Δ MFCC and Δ log power were used as feature parameters. The full Korean phoneme set, as defined in [17], contained a total of 40 phoneme symbols. These consisted of 19 consonants and 21 vowels (including nine monophthongs and 12 diphthongs). One silence symbol was added during acoustic model training. Three states were used as the initial hidden Markov model (HMM) for each phoneme. Then, a shared state HMnet topology was obtained using a successive state splitting (SSS) algorithm based on the minimum description length (MDL) criterion to gain the optimal structure in which triphone contexts are shared and tied at the state level. Details about MDL-SSS can be found in [18]. The resulting context-dependent triphone model had 2,231 states in total with an optimum 15 Gaussian mixture components per state. The decoding engine is a time-synchronous Viterbi beam search system that operates on a Weighted Finite State Transducer (WFST) search space [19]. Thus, all components were compiled into $C \circ L \circ G$, a recognition cascade, where C is the context-dependency acoustic model, L is the lexicon dictionary and G is the language model.

A number of different LVCSR units were explored and investigated for our baseline system: (1) U1-ChrBase is based on syllable *eumjeol* units; (2) U2-WrdBase is based on the "word" unit generated by a morphological analysis tool and a morpheme dictionary; and (3) U3-PhrBase is based on word-phrase *eojeol* units. To take into consideration the phonological phenomena and coarticulation effects on the units, we also constructed a joint sequence model that used tuple units consisting of Korean orthography together with its pronunciation: (4) U4-ChrPair is based on a tuple unit of syllable *eumjeol* units together with phonemic syllables of the actual pronunciation; (5) U5-WrdPair is based on tuple units of the "word" unit generated by the morphological analysis tool and the morpheme dictionary together with its pronunciation; and (6) U6-PhrPair is also based on tuple units of the word-phrase *eojeol* unit together with its pronunciation. The dictionary size and the perplexity of the trigram language models are summarized in Table 1.

Table 1. The dictionary size and the perplexity of trigram language models based on various different LVCSR units

	Dictionary Size	**LM Perp. (OOV)**
U1-ChrBase	4,524	11.2 (0%)
U2-WrdBase	52,718	30.3 (<0.5%)
U3-PhrBase	216,030	263.1 (5%)
U4-ChrPair	4,524	11.6 (0%)
U5-WrdPair	52,718	31.8 (<0.5%)
U6-PhrPair	216,030	267.4 (5%)
DP100-WrdPair	23,963	35.6 (<0.1%)
DP300-WrdPair	23,826	36.2 (<0.1%)

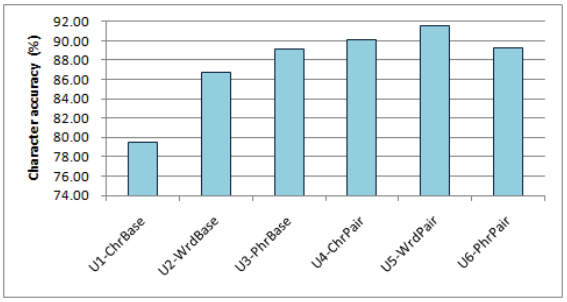

Fig. 1. *Character accuracy of the baseline LVCSR systems.*

The performance of the baseline system with different LVCSR units is shown in Fig. 1. Since each system has a different basic unit length, only the character accuracy of each system is presented here. The performance of the baseline U3-PhrBase outperformed both the U1-ChrBase and U2-WrdBase baselines, because it has longer units that are likely to reduce the acoustic confusability. However, it is difficult to use this baseline system in real applications because of its high perplexity and OOV rates. The use of tuple units consisting of Korean orthographic units together with their pronunciation improved system performance further: the best baseline system was the U5-WrdPair, which achieved 91.55% character accuracy.

4.3. Proposed LVCSR System

Using the same amount of training text material, we trained our DP model and extracted the resulting new "word" units from the DP segmentation of the data. The convergence of the algorithm during training procedure is shown in Fig. 2, which plots the log-probability of the sampled derivation at the end of each pass through the training corpus (iteration) against the iteration number. It can be seen from the graph that the system rapidly improves from the poor initial segmentation, and thereafter continues to gradually improve. In this study, we took two different samples of corpus segmentation hypotheses from iterations 100 and 300, and used their co-segmentations to build a joint sequence model based on tuples of the new "word" units together with their pronunciation. These systems are denoted DP100-WrdPair and DP300-WrdPair, respectively.

When integrating these models into the LVCSR system, the feature parameters and acoustic model are identical to the baseline.

The lexicon dictionary and language model are generated based on DP100-WrdPair or DP300-WrdPair. The dictionary size and the perplexity of the trigram language models for both DP100-WrdPair and DP300-WrdPair are summarized in Table 1. The performance of the proposed system in terms of both character accuracy and word accuracy in comparison with the best baseline system U5-WrdPair is shown in Fig. 3. Additionally, we also include U5b-WrdPair in which we reduced the U5-WrdPair dictionary size by selecting only the most frequent words. The results show that by reducing the dictionary size of the baseline system, the performance dropped from 83.71% to 81.56%. However, the proposed approach produced a model with half the number of lexical units of that provided by the morphological analysis tool, whilst at the same time improving the recognition accuracy of the LVCSR system. The best proposed system is DP300-WrdPair, which was able to provide a 16.86% relative word error rate reduction over U5b-WrdPair given the same dictionary size, and a 5.89% relative word error rate reduction over U5-WrdPair (the best baseline system).

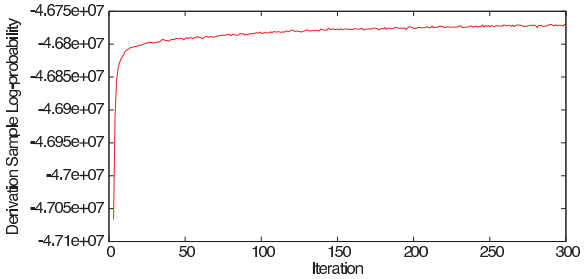

Fig. 2. *The evolution of the log-probability of the Gibbs sampled derivation with respect to training iteration.*

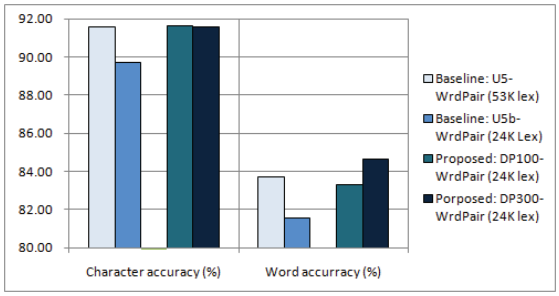

Fig. 3. *Character and word accuracy of the proposed LVCSR system in comparison with the best baseline system.*

5. CONCLUSION

In this paper, we propose a novel method of determining Korean LVCSR units in unsupervised manner using a non-parameteric Bayesian approach. Specifically, we utilize a Dirichlet process model trained using Bayesian inference through block Gibbs sampling. This approach provides a principled way of learning units, whilst controlling model complexity and avoiding the overfitting issues typically associated with maximum likelihood approaches. To model phonological phenomena and the coarticulation between the units, we apply the proposed approach to tuple sequences consisting of Korean orthographic units together paired with their pronunciation. Experiments were conducted on a travel domain corpus that included many foreign words and proper nouns. The results clearly expose the key advantages of our approach. First, in terms of model complexity, our technique yields a model with far fewer parameters than the baseline systems: the number of lexical units in the model being about one half of those produced by any of the baseline systems. Secondly, in terms of the recognition accuracy of the LVCSR system: the proposed system was able to achieve a 16.86% relative word error rate reduction over U5b-WrdPair given the same dictionary size, and a 5.89% relative word error rate reduction over U5-WrdPair (the best baseline system). The final advantage of our approach is that the entire segmentation process is performed automatically in an unsupervised manner from two token sequences. It relies on no explicit linguistic knowledge, and should therefore be applicable to other languages with little or no modification, removing the need for morphological analysis tools for segmentation.

6. ACKNOWLEDGEMENTS

The authors would like to thank Paul Dixon for his support and useful discussion regarding the WFST framework.

7. REFERENCES

[1] K. Yoon and C. Brew, "A linguistically motivated approach to grapheme-to-phoneme conversion for Korean," *Computer Speech & Language*, vol. 20, no. 4, pp. 357–381, 2006.

[2] S.-C. Song, *201 Korean Verbs - Fully Conjugated in All Forms*, Baron's Educational Series, 1988.

[3] G. Choueiter, D. Povey, S.F. Chen, and G. Zweig, "Morpheme-based language modeling for Arabic LVCSR," in *Proc. ICASSP*, Tolouse, France, 2006, pp. 1053–1056.

[4] K. Kirchoff and R. Sarikaya, "Processing morphologically-rich languages," in *INTERSPEECH Tutorial*, Antwerp, Belgium, 2007.

[5] H.-J. Yu, H. Kim, J.-M. Hong, M.-S. Kim, and J.-S. Lee, "Large vocabulary Korean continuous speech recognition using a one-pass algorithm," in *Proc. ICSLP*, Beijing, China, 2000, pp. 278–281.

[6] 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.

[7] D. Kiecza, T. Schultz, and A. Waibel, "Data-driven determination of appropriate dictionary units for Korean LVCSR," in *Proc. ICSP*, Seoul, Korea, 1999, pp. 323–327.

[8] T. S. Ferguson, "A bayesian analysis of some nonparametric problems," *Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[9] R. M. Neal, "Markov chain sampling methods for dirichlet process mixture model," *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 249–265, 2000.

[10] J. Xu, J. Gao, K. Toutanova, and H. Ney, "Bayesian semi-supervised Chinese word segmentation for statistical machine translation," in *Proc. COLING*, Morristown, NJ, USA, 2008, pp. 1017–1024.

[11] D. Mochihasi, T. Yamada, and N. Ueda, "Bayesian unsupervised word segmentation with nested Pitman-Yor language model," in *Proc. ACL-IJCNLP*, Singapore, 2009, pp. 100–108.

[12] A. Finch and E. Sumita, "A Bayesian model of bilingual segmentation for transliteration," in *Proc. IWSLT*, Paris, France, 2010, pp. 259–266.

[13] D. J. Aldous, "Exchangeability and related topics," in *École d'été de probabilités de Saint-Flour, XIII-1983, Lecture Notes in Math*, vol. 1117, pp. 1–198. Springer, Berlin, 1985.

[14] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer, 2004.

[15] G. Kikui, S. Yamamoto, T. Takezawa, and E. Sumita, "Comparative study on corpora for speech translation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1674–1682, 2006.

[16] B.W. Kim, D.L. Choi, Y.I. Kim, K.H. Lee, and Y.J. Lee, "Current state and future plants at SiTEC for speech corpora for common use," *Malsori*, vol. 46, pp. 175186, 2003.

[17] 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.

[18] T. Jitsuhiro, T. Matsui, and S. Nakamura, "Automatic generation of non-uniform HMM topologies based on the MDL criterion," *IEICE Trans. Inf. & Syst.*, vol. E87-D, no. 8, pp. 2121–2129, 2004.

[19] P. R. Dixon, D. A Caseiro, T. Oonishi, and S. Furui, "The Titech large vocabulary WFST speech recognition system," in *Proc. ASRU*, Kyoto, Japan, 2007, pp. 1301–1304.