International Journal of Speech Technology (2019) 22:971–977
https://doi.org/10.1007/s10772-019-09637-2

A usage of the syllable unit based on morphological statistics in Korean large vocabulary continuous speech recognition system

Hyok-Chol Ri[1]

Received: 4 March 2019 / Accepted: 16 September 2019 / Published online: 25 September 2019

Abstract
In large vocabulary continuous speech recognition (LVCSR), it is important in improving the system's performance to determine reasonably the recognition unit. In Korean continuous speech recognition, a morph rather than a word is used basically as the recognition unit due to Korean's agglutinative property and a good performance is provided by combining high-frequency morph sequences, which leading to an increase of vocabulary size and high out-of-vocabulary (OOV) rate. Sub-lexical units such as a syllable and a graphone are widely used for inflectional languages, while they have not been introduced successfully for Korean speech recognition, due to a weakness of their linguistic information. In this paper, we investigate a usage of a syllable unit to resolve a mismatch problem between the recognition unit and vocabulary size that have occurred frequently in Korean large vocabulary speech recognition. We apply the local segmentation into syllables based on morphological statistics and perform experiments using the language model (LM) constructed from mixed unit types of morpheme, combined morpheme and syllable. By the proposed model, an absolute reduction of around 0.4% in word error rate (WER) is obtained compared to a traditional LM consisting of morphemes and combined morphemes.

Keywords Recognition unit · Language model · Morpheme · Syllable

1 Introduction

Recognition units for large vocabulary continuous speech recognition (LVCSR) are different in languages. European languages such as English have the inflectional property, and thus are characterized by high lexical variety. This morphological richness leads to high OOV rates, and causes a data sparseness problem and high LM perplexities. For such languages the use of sub-lexical units for LVCSR becomes a natural choice. For Korean continuous speech recognition, a morph is used basically as the recognition unit due to a agglutinative property and a good performance is provided by combining high frequency morph sequences (Kurimo et al. 2006). But this causes an increase of vocabulary size and high OOV rate, thus it is inevitable to use shorter sub-lexical units with respect to LVCSR system design.

On the other hand, the open vocabulary LVCSR tasks require the number of recognizable words to almost be infinite. Therefore, the recognition of OOV words is a major challenge for such systems. To improve the OOV recognition rate, sub-lexical LMs are good candidates. The sub-lexical units can be properly combined to produce a wide range of words, achieving better lexical coverage and thus fitting the task of open vocabulary speech recognition.

One of the main issues of sub-lexical language modeling is the proper choice of the sub-word type. A non-careful choice of the sub-word type could increase the WER.

A possible type of sub-word is the morpheme which is the smallest linguistic component of the word that has a semantic meaning. Normally, morphemes are generated from the full-words by applying word decomposition based on supervised or unsupervised approaches. Supervised methods rely on carefully built morphological analyzers based on lexical and syntactic knowledge (El-Desoky et al. 2009; Byrne et al. 2000; Kneissler and Klakow 2001; Diehl et al. 2012). Although the supervised decomposition is normally optimized for high performance, it requires labor-intensive work and still suffers from the so-called unknown word problem. On the other hand, unsupervised approaches (Adda-Decker 2003; Ordelman et al. 2003; Rotovnik et al. 2007; Larson et al. 2000; Creutz et al. 2007; Creutz 2006) are statistical

✉ Hyok-Chol Ri
info3@ryongnamsan.edu.kp

[1] College of Information Science, KIM IL SUNG University, Pyongyang, Democratic People's Republic of Korea

and data driven approaches, and are language independent as they do not require any language specific knowledge and can be applied to any language. In Korean continuous speech recognition, the word is unreasonable unit for recognition due to a agglutinative property, therefore the morph is used basically as recognition unit. But short morphs such as prefix and suffix are difficult to reflect exactly the pronouncing change between their boundaries and become a basic cause of insertion and deletion errors in the speech recognition. When the compound words are produced from combining morphs and used as a recognition unit, then they can not only reflect a pronouncing change between boundaries of combined morphs but give the effectiveness in a high order n-gram locally in terms of the language model (Stolke 2006; Hirsimaki 2006; Huet 2010). But in large vocabulary continuous speech recognition applications dealing with many fields, using all morphs and compound words as it leads to an increase of vocabulary size, therefore data type of word index becomes larger and it results in a 'fat' model. Of course, various methods for compressing and reducing a model have been proposed, but they caused the lowering of performance. And the method to cut off low-frequency words locally drops coverage rate.

Another type of sub-word is the syllable which is a phonological building block of words. Syllable based LMs are successfully used for languages like Chinese (Xu et al. 1996), Polish (Piotr 2008), and English (Schrumpf et al. 2005). Syllable unit is not used independently because of it's short length and weak linguistic constraint for Korean language. And previous works have proposed the method to represent a training corpus as variable-length syllable sequences in the manner of maximum likelihood segmentation and estimate model parameters (Zitoni 2003). By the iteration of segmentation and re-estimation, this method could produce variable-length syllable sequences and control the produced number with a convergence condition, resulting in some improvement of model's perplexity. But many sequences lost a linguistic meaning due to combining only statistically, thus it gives negative effect to automatic speech recognition (ASR) performance.

Another type of unit is the graphone which is a combination of the graphemic sub-word with its context dependent pronunciation forming one joint unit. A set of graphones is used for OOV words in an English ASR task, where the graphones are constructed based on fixed-length sub-words without any linguistic considerations (Bisani and Ney 2005). While a set of graphones based on morphemes derived from data driven segmentation is used to model OOV words in a German LVCSR system. Graphones are mainly used to model OOV words (El-Desoky et al. 2010; Shaik et al. 2011).

So far, we conducted extensive research about various sub-lexical units. In this work, we investigate the use of hybrid lexicons and LM based on three mixed types of sub-lexical units for building a Korean LVCSR system. Frequent words are represented with combined morphemes and morphemes as lexical unit. While, for less frequent words cut off from the vocabulary and OOV words, syllables are used. This mixture of units is hypothesized as a more reliable methodology to achieve better lexical coverage and experimented for a LVCSR system. The experimental results show significant improvements in WERs and OOV rates.

2 Hybrid lexicon consisting of three types of sub-lexical units

2.1 Morpheme based sub-words

We perform word decomposition using morphological analysis technology adopted for natural language processing (NLP) applications, according to word decomposition criterion suitable for ASR. The aim of morphological analysis for word decomposition in ASR is neither to analyze a language nor to check a spelling error, but is just to decompose a word into proper linguistic units so as to improve the ASR performance.

If results of Korean morphological analysis based on rules (supervised method) are used directly as lexical unit, they are confusable in decoding due to their short lengths, and thus they are combined according to some criterion. For example, in the result 《나 + 는 + 대학 + 으로 + 가 + ㄴ + 다》 of morphological analysis on the sentence 《나는 대학으로 간다》 ("I go to college"), 《ㄴ》 in itself is unreasonable lexical unit, and so if we combine resulting morphemes according to proper criterions like 《나 + 는 + 대학 + 으로 + 가 + ㄴ 다》, then it is efficient in building a LM and enhancing the overall performance.

Based on the morpheme corpus in which results of morphological analysis are combined according to some criterions, we construct a word decomposition model and then decompose words in the statistical method (unsupervised method).

2.2 Combined morpheme (compound word) based sub-words

Generally, a substantial number of short phrase have a very high frequency in natural languages (Stolke 2006). For Korean language, the short morphemes have very high frequency and combined morphemes (we named it as compound word) are often used as single linguistic units. These compound words play an important role in the improvement of language model.

In selecting compound words by statistical method, it is important to determine proper measure for evaluating them. A standard that pairs of morpheme can be compound word is determined as follows.

First, compound words must be pairs of morpheme with high frequency in training corpus. Pairs of morpheme with low frequency have not to be selected as compound words, because adding these pairs of morpheme with low frequency to vocabulary can cause acoustic confusability with other words similar to them during decoding.

Second, the morphemes within compound word have to occur frequently together and more rarely in the pair context of other morphemes. In case that a short morpheme with high frequency occur together with other several pairs of morpheme with high frequency, if all these pairs were to be added to the vocabulary then the confusability between them would be increased. This will result in insertions or deletions of errors.

Based on above standard, we combine highly frequent morpheme pairs and repeat this procedure until there are no candidates. Here we determine experimentally threshold of combining count, as about 300.

2.3 Syllable based sub-words

Morpheme pairs occurring frequently in a training corpus are often selected as compound words. If we add low-frequency morph pair as a compound word to vocabulary, then confusing with acoustically similar words could happen in decoding procedure.

Finally, there are compound words and morphemes in the training corpus and statistic data by Part-of-Speech (POS) information are given in Table 1. Though specific details depend on a training corpus, the overall statistics retain a generality. Here, we introduce only typical POSs comprising large proportion in the vocabulary and omit POSs such as prefix, suffix, interjection that depend little on a corpus.

Table 1 Statistics of typical POSs in vocabulary

POS	Percent (%)
Compound word	10.25
General noun	21.59
Proper noun	53.10
Adjective	1.66
Verb	2.42
Adverb	2.82
Combined particle	2.71

Table 2 Frequency characteristics of proper and general noun dictionaries

Frequency threshold	Percent in proper noun dictionary (%)	Percent in general noun dictionary (%)
1 and less	41.50	11.42
5 and less	69.75	25.59
10 and less	78.48	33.36
20 and less	85.18	41.45
50 and less	91.83	52.76
100 and less	95.23	61.22
200 and less	97.39	69.76

As seen in Table 1, proper nouns including loanwords and general nouns comprise about 75% of the total vocabulary, and the difference between their frequencies is great. Frequency characteristics of proper nouns and general nouns are given in Table 2 and Fig. 1.

As seen in Fig. 1, proper nouns show partially saturated feature, but general nouns show nearly linear feature. Almost of proper nouns are loanwords and they influence individually little on the recognition performance of system because of low frequencies. Therefore, it is pretty possible for low-frequency proper nouns to be excluded in the vocabulary, but the number of words to be excluded is fairly great (even about 50% of total vocabulary), resulting in considerable losses on the coverage rate of vocabulary and overall performance of system. It is same as for general nouns and only the frequency characteristic differs from the former.

(example) proper noun : 킹스타운(Kingstown) → 킹 + 스 + 타 + 운

동쥬앙(Don Juan) → 동 + 쥬 + 앙

로제타석(Rosetta stone) → 로 + 제 + 타 + 석

알라닌(alanine) → 알 + 라 + 닌

general noun : 반살미(reception after the wedding ceremony) → 반 + 살 + 미

여락(leftover pleasure) → 여 + 락

항다반(matter of common occurrence) → 항 + 다 + 반

방구리(water jar) → 방 + 구 + 리

Here, low-frequency words are usually unfamiliar for common users and thus it is quite possible to pronounce such Korean words by one syllable.

Based on such statistical analysis and phonetic characteristic of Korean language, while we use compound words and morphemes as it is, we are going to segment low-frequency words of proper and general noun dictionaries into syllables that hold a large proportion in total vocabulary and depend strongly on a topic and to represent them as syllable n-gram, thus resulting in solving all problems of coverage rate, vocabulary size and recognition performance. From this viewpoint, we would verify the efficiency of proposed method through experiments.

3 Experimental setup

3.1 Acoustic model

To train Hidden Markov Models (HMMs), 400 h of audio corpora are recorded by 120 men and 60 women in sampling frequency 22.05 kHz.

Acoustic model is trained by the HMM toolkit (Young et al. 2006). The script for modeling includes 48 monophones and 23,000 triphones and trained monophone and triphone models have averagely 60 Gaussian distributions and 24 ones per a state, respectively.

The acoustic model includes around 4400 states and 104,000 Gaussian distributions. The acoustic feature vector is 29-order one, which is obtained from applying HLDA transform to 39-order one that consists of 13-order MFCC parameters and their derivatives of the first and second order. A width of window for speech analysis is 25.6 ms and 100 frames are produced per a second.

Audio data for testing is read data of 1176 sentences by 20 men and 20 women from text corpus that is not used for LM training. Total of recordings takes about 5 h.

3.2 Language model

Here, we use experimental LMs constructed from text corpora such as 《Rodong Sinmun》 and ones of sociopolitical and cultural fields, including economy, military affairs, philosophy, history and law. Text corpora are databases appended by POS tags for each morph, of which the size is 2.3 GByte and the vocabulary size is 130K. Test data are 1176 sentences taken from a corpus of common sense field that is not used for training LM.

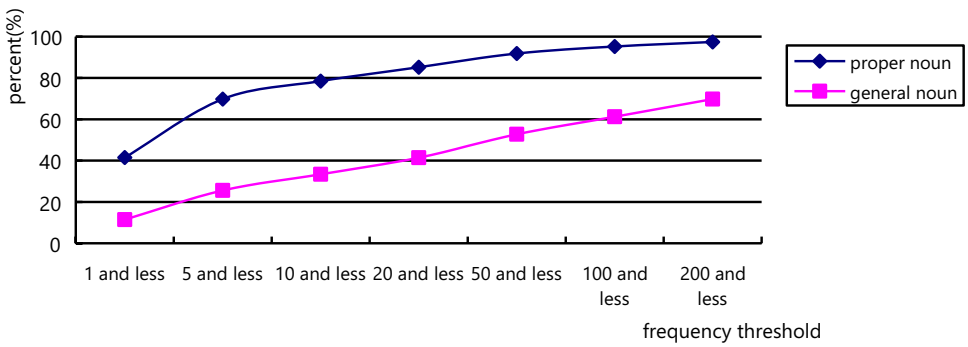

Fig. 1 Frequency characteristic graph of proper and general noun dictionaries

Table 3 Baseline recognition results using LMs based on morphemes and compound words

LM	Lexical unit	Vocabulary size	Training data		Test data	
			OOV rate (%)	WER (%)	OOV rate (%)	WER (%)
LM0	Compound word, morpheme	65k	0.27	4.65	0.21	4.78
LM1	Compound word, morpheme	130k	0	4.18	0.17	4.36

Table 4 Recognition results for baseline LM and LMs using a syllable

LM	Lexical unit	Vocabulary size (k)	Training data		Test data	
			OOV rate (%)	WER [%]	OOV rate (%)	WER (%)
LM0	Compound word, morpheme	65	0.27	4.65	0.21	4.78
LM1	Compound word, morpheme	130	0	4.18	0.17	4.36
LM2	Variable-length syllable sequence	65	0	4.72	0	4.80
LM3	Compound word, morpheme, syllable	65	***0***	***4.27***	***0***	***4.34***

Trigram LM is constructed by SRILM toolkit (Stolcke 2002) from training data. Here, LMs are estimated in the modified Kneser–Ney method, and then model reduction is performed using the relative entropy measure.

To evaluate the gain of recognition performance against the increase of vocabulary size and model size, we prepare LMs for 2 types of vocabulary size: 65k vocabulary (the size of data type of word index is 2 Bytes), 130k vocabulary (the size of data type of word index is 4 Bytes).

3.3 Decoder

Korean speech recognizer "RyongNamSan"is two-pass LVCR that in the first pass candidate hypotheses are generated by a synchronous Viterbi beam search, and in the second pass they are rescored to produce a final result. To enhance the recognition accuracy and speed, triphone HMMs and trigram LM are used in decoding.

4 Experiments

In this section, we explain our open vocabulary recognition experiments. First, we introduce our baseline experiments. Then, we present results using hybrid sub-lexical language models based on mixed unit types as discussed in Sect. 2. At the end, we analyze the advantages and disadvantages of our approach.

4.1 Baseline experiments (open vocabulary ASR)

In Table 3, we show the results of our baseline recognition experiments using traditional morpheme and compound word LMs. As the result of experiments, OOV rates and WERs are shown on training data and test data.

We consider the system of 65k morphemes and compound words as a reference baseline, while the system of 130k is listed for comparison purposes.

Baseline model LM0 is word 3-gram model with the vocabulary of 65k words in which proper and general nouns with 20, 10 frequencies and less are excluded from total vocabulary of 130 k, respectively. Word 3-gram model including total vocabulary of 130 k words is selected as baseline model LM1 to compare relatively between a model size and the recognition performance.

Although the OOV rate of training data is higher compared with one of test data for LM0, WER is lower. The reason is considered that a large number of low frequency words have been cut off and n-gram contexts differ partly between test and training data. Increases of WER due to a difference of n-gram contexts between test and training data are common for all LMs.

When comparing LM1 with LM0, it is obvious that although both of LM0 and LM1 use same mixed type of sub-lexical units, LM1 is superior in terms of OOV rate and WER by using a total of uncut off vocabulary.

4.2 Comparative experiments (open vocabulary ASR)

In Table 2, we summarize results of recognition experiments using syllable based LMs. We distinguish two main types of experiments: the one where the basic sub-lexical unit is compound word, morpheme and syllable, and the one where the basic sub-lexical unit is variable-length syllable sequence.

The vocabulary size is fixed to 65k.

We consider 65k variable-length syllable 3-gram model as comparative model LM2. Variable-length syllable model represents in-vocabulary words in form of combination of syllables, therefore it is selected as a comparative model from the aspect of using a syllable.

Graphones, which take account of only context dependent pronunciations without any linguistic considerations, are partly used for recognition of OOV words and are not suitable as recognition unit, thus we exclude from comparison in our experiments (we have verified it through preliminary experiments).

LM3 (proposed model) is 3-gram model of mixed types of sub-lexical units of compound word, morpheme and syllable, where words excluded from baseline model LM0 are segmented into syllables and trained. Among 65k entries, 12k entries are compound words, 51k morphemes, and rest 2k syllables. The vocabulary is fully covered by a mixture of compound words, morphemes and syllables. Any OOV words can be represented with syllable sequences, thus any sentences can be fully covered by mixed types of compound words, morphemes and syllables.

Results of comparative experiments are detailed in Table 4. For comparison, results of baseline models are given with them.

As shown in Table 4, both of LM2 and LM3 can resolve the OOV problem with the ability of syllable to represent a word. But LM3 retains linguistic syntactic components like a compound word and a morpheme, and therefore it achieves 9.5% relative (0.45% absolute) decrease in WER compared to LM2. This shows that the proposed model is superior absolutely to variable-length syllable model.

4.3 Experimental analysis

First, let us compare our model with baseline model LM0.

LM0 causes 0.27% OOV words due to cutting off low frequency words in training data, giving negative effects on recognition performance. So it results in WER decline of 0.38% absolute than proposed model LM3 that is representative of low frequency words in syllable unit. Moreover, with high coverage ability, LM3 achieves improvement in WER of 0.44% absolute than LM0 on test data.

Next, let compare our model with baseline model LM1.

Although LM1 gives the improvement in WER of about 0.1% absolute on training data than our model by holding total vocabulary in forms of compound word and morph unit, but the relative improvement of recognition performance is slight, compared with a growth of model size (by about 2.7 times). But LM1 reduces the performance on test data by about 0.2% due to OOV words, so our model is superior a little rather than LM1.

On the whole, we can consider that the proposed model is superior to all baseline models with respect to all of model size, OOV rate and recognition performance.

Next, we evaluate dependencies of each model on test data. For variable-length syllable model and our model possessing a high coverage ability, the degrees of WER decrease between training data and test data are respectively 0.08% and 0.07%, which are less than baseline models (LM0: 0.13%, LM1: 0.18%). This shows that corpus dependencies, task dependencies of syllable based models are small relatively.

So, we verified the superiority of a proposed method through experiments.

We can adjust experimentally the number of compound words and the frequency threshold for selecting words to segment into syllables, according to a scope of recognition system and applications.

5 Conclusion

In this paper, we proposed a usage of a syllable based on a statistic of POSs as well as a compound word and a morpheme as sub-lexical units to resolve the recognition unit and vocabulary size arising in applications of Korean large vocabulary speech recognition system.

Through experiments we demonstrated that proposed method was superior to previous methods from the practical viewpoint. In experiments, proposed model achieves 9.5% relative (0.45% absolute) decrease in WER compared to variable-length syllable model, showing that it is superior absolutely to comparative model. Then it resulted in WER improvements of 0.38% and 0.44% absolute on training data and test data respectively compared to compound word and morpheme based baseline model of same vocabulary size. Moreover, we verified the superiority of our model to 130 k baseline model holding total vocabulary in forms of compound word and morph unit with respect to all of model size, OOV rate and recognition performance, and then evaluated a degree of task dependency compared with baseline models.

Acknowledgements We appreciate the helpful discussions with Dr. Kim and Prof. Ri, anonymous reviewers and editors for many invaluable comments and suggestions to improve this paper.

References

Adda-Decker, M. (2003). A corpus-based decompounding algorithm for German lexical modeling in LVCSR. *Proceedings European Conference on Speech Communication and Technology* (pp. 257–260). Geneva, Switzerland.

Bisani, M., & Ney, H. (2005). Open vocabulary speech recognition with flat hybrid models. *Interspeech* (pp. 725–728), Lisbon, Portugal.

Byrne, W., Hajič, J., Ircing, P., Krbec, P., & Psutka, J. (2000). Morpheme based language models for speech recognition of Czech. *Text, Speech and Dialogue, ser. Lecture Notes in Computer Science, 1902* (pp. 139–162). Berlin: Springer.

Creutz, M. (2006). Induction of the morphology of natural language: Unsupervised morpheme segmentation with application to automatic speech recognition. *Ph.D. dissertation*, Helsinki University of Technology, Finland, 2006.

Creutz, M., Hirsimäki, T., Kurimo, M., Puurula, A., Pylkkönen, J., Siivola, V., et al. (2007). Morph-based speech recognition and modeling of out of- vocabulary words across languages. *ACM Transactions on Speech and Language Processing, 5*(1), 3.

Diehl, F., Gales, M., Tomalin, M., & Woodland, P. (2012). Morphological decomposition in Arabic ASR systems. *Computer Speech and Language, 26,* 229–243.

El-Desoky, A., Gollan, C., Rybach, D., Schlüter, R., & Ney, H. (2009). Investigating the use of morphological decomposition and diacritization for improving Arabic LVCSR. *Interspeech* (pp. 2679–2682), Brighton, UK.

El-Desoky, A., Shaik, M., Schlüter, R., & Ney, H. (2010). Sub-lexical language models for German LVCSR. *IEEE Workshop on Spoken Language Technology* (pp. 159–164), Berkeley, CA, USA, Dec. 2010.

Hirsimaki, T. (2006). Unlimited vocabulary speech recognition with morph language models applied to Finish. *Computer Speech and Language, 20,* 515–541.

Huet, S. (2010). Morpho-syntactic post-processing of N-best lists for improved French automatic speech recognition. *Computer Speech and Language, 24,* 663–684.

Kneissler, J., & Klakow, D. (2001). Speech recognition for huge vocabularies by using optimized sub-word units. *Proceedings of the European Conference on Speech Communication and Technology*, 1, (pp. 69–72). Aalborg, Denmark.

Kurimo, M., Puurula, A., Arisoy, E., Siivola, V., Hirsimäki, T., Pylkkönen, J., Alumäe, T., & Saraclar, M. (2006). Unlimited vocabulary speech recognition for agglutinative languages. *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL* (pp. 487–494).

Larson, M., Willett, D., Köhler, J., & Rigoll, R. (2000). Compound splitting and lexical unit recombination for improved performance of a speech recognition system for German parliamentary speeches. *Proceedings of the International Conference on Spoken Language Processing*, Beijing, China.

Ordelman, R., Hassen, A. V., & Jong, F. D. (2003). Compound decomposition in Dutch large vocabulary speech recognition. *Proceedings of the European Conference on Speech Communication and Technology* (pp. 225–228), Geneva, Switzerland.

Piotr, M. (2008). Syllable based language model for large vocabulary continuous speech recognition of polish. *Text, Speech and Dialogue, ser. Lecture Notes in Computer Science, 5246,* 397–401.

Rotovnik, T., Maučec, M. S., & Kačič, Z. (2007). Large vocabulary continuous speech recognition of an inflected language using stems and endings. *Speech Communication, 49*(6), 452–537.

Schrumpf, C., Larson, M., & Eickeler, S. (2005). Syllable-based language models in speech recognition for English spoken document retrieval. *Proceedings of the 7th International Workshop of the EU Network of Excellence DELOS on AVIVDiLib* (pp. 196–205). Cortona, Italy.

Shaik, M., El-Desoky, A.,Schlüter, R., & Ney, H. (2011). hybrid language models using mixed types of sub-lexical units for open vocabulary German LVCSR. *Interspeech* (pp. 28–31). Florence, Italy.

Stolcke, A. (2002). SRILM—an extensible language modeling toolkit. *Proceedings of the International Conference on Spoken Language Processing, 2* (pp. 901–904). Denver, Colorado, USA.

Stolke, A. (2006). Morphology-based language modeling for conversational Arabic speech recognition. *Computer Speech and Language, 20,* 589–608.

Xu, B., Ma, B., Zhang, S., Qu, F., & Huang, T. (1996). Speaker independent dictation of Chinese speech with 32K vocabulary. *Proceeding of Fourth International Conference on Spoken Language Processing* (Vol. 4, pp. 2320 – 2323), Philadelphia, PA, USA.

Young, S., et al. (2006). *The HTK book version 3.4*. Cambridge: Cambridge University.

Zitoni, I. (2003). Statistical language modeling based on variable-length sequences. *Computer Speech and Language, 17,* 27–41.