Unsupervised Acoustic Model Training for the Korean Language

Antoine Laurent, William Hartmann, Lori Lamel

Spoken Language Processing Group
CNRS-LIMSI, BP 133 91403 Orsay cedex, France
laurent@limsi.fr, hartmann@limsi.fr, lamel@limsi.fr

Abstract

This paper investigates unsupervised training strategies for the Korean language in the context of the DGA RAPID Rapmat project. As with previous studies, we begin with only a small amount of manually transcribed data to build preliminary acoustic models. Using the initial models, a larger set of untranscribed audio data is decoded to produce approximate transcripts. We compare both GMM and DNN acoustic models for both the unsupervised transcription and the final recognition system. While the DNN acoustic models produce a lower word error rate on the test set, training on the transcripts from the GMM system provides the best overall performance. We also achieve better performance by expanding the original phone set. Finally, we examine the efficacy of automatically building a test set by comparing system performance both before and after manually correcting the test set.

Index Terms: speech recognition, unsupervised training, korean, under-resourced language

1. Introduction

For languages with limited resources, building Large Vocabulary Continuous Speech Recognition (LVCSR) systems is a challenge. Typical systems are built from dozens or even hundreds of hours of transcribed audio data and written text containing millions of words [1]. Many languages, such as Korean, do not have this amount of supervised training data widely available such as via data providers such as LDC or ELRA. Although large amounts of acoustic data can be found online in the form of television shows and podcasts in the Korean language, the data is largely unstructured and without accurate transcripts. Lightly supervised approaches attempt to use quick transcripts or closely related texts [2, 3]. Other approaches do not rely on any transcribed data.

Most of these unsupervised methods rely on a strong language model to guide the training process [4]. The many varied approaches differ in their reliance on confidence scores [5, 6, 7], the use of iterative training [8], and on the level of supervision [9]. A more detailed analysis of supervised and unsupervised approaches can be seen in [10, 11].

This study is the continuation of the one initialized in [12]. While the previous study focused on the development of the Korean corpus, we extend the results in several ways. Previous work was done using an uncorrected test corpus. The new experiments present a comparison between the corrected and uncorrected test set. The comparison serves to illustrate that while the uncorrected corpus is not as accurate, it does serve as a proxy for determining the efficacy of certain techniques.

Two acoustic modeling techniques are explored (classical GMM and DNN [13, 14, 15, 16]) for the unsupervised training and for the decoding. We show that cross-model adaptation—training on transcripts produced from a different system—provides a significant improvement. Further, our results demonstrate that using the most accurate system for unsupervised transcription does not necessarily produce the best performance. The best unsupervised system provides a 23% relative reduction in character error rate compared to a state-of-the-art supervised DNN system. We also see further improvements by expanding the original phone set.

System development is very lightly supervised. We used a small annotated corpus of Korean Broadcast News from VOA distributed by the LDC to bootstrap the language and acoustic models. Additional audio data without any transcripts were then used to improve the acoustic models, and language models were built using several sources of text data (also from LDC or web downloads).

We will present a brief overview of the characteristics of the Korean language (more details in [12]), followed by a description of the training and testing corpora. Section 3 presents the speech recognition system, including a description of the acoustic units, acoustic models, and language models. The unsupervised training approach is described in Section 4. Section 5 describes the experimental results and conclusions are presented in Section 6. This speech recognition system will be used for the RAPMAT (Speech translation) project [1].

2. Korean Language Data

2.1. Korean Language

Historically, the Korean language was written with adapted Chinese characters. More recently a writing system known as *Hangeul* is used, although, many Chinese characters are still used too. Each *Hangeul* character represents a syllable consisting of one or more phonetic components. Through the combination of the phonetic components, over 11,000 *Hangeul* syllables are possible. While many of these combinations are never used or are illegal phonetically, they still occasionally appear in online data.

While a single syllable can represent a word, words are typically composed of strings of syllables—often referred to as word phrases [17]. A single word phrase in Korean would typically correspond to several words in English. Given the large number of syllables and the fact that word phrases can consist of sequences of several syllables, the effective vocabulary of Korean is quite large [18]. For example, where a 40 million word English corpus contains about 190000 distinct words [1], the 95 million word Korean corpus used in this work contains about 2 million distinct words.

[1] This work was partially financed by the DGA RAPID project RAPMAT. http://www.limsi.fr/tlp/rapmat.html

2.2. Training Corpus

We only located a limited number of small corpora for Korean via LDC [19] or ELDA [20]. Larger studies typically use undistributed internal data. We instead use a combination of a small amount of LDC data with a subset of the unsupervised corpus described in [12].

From the LDC, we use a 9-hour corpus of transcribed broadcast news speech. This 9-hour corpus represents the total amount of supervised data used for acoustic model training. The supervised acoustic audio data was supplemented with an additional 100 hours of untranscribed available from Korean news websites (VOA[2], RFA[3], and NHK[4]). For language model training, we also used the LDC corpus Korean newswire second edition (LDC2010T19, which includes newswire first edition, composed of 55M words) and the transcripts from the LDC Korean telephone conversations corpus (LDC2003T08, 230k words) for language model training.

2.3. Testing Corpus

For development, we also use a single hour of acoustic data from the LDC. However, the development set is not ideal due to both its small size and close match to the supervised training data. In order to create an actual test set, we attempted to build an unsupervised corpus. Approximately 3.5 hours of audio data coming from RFA, VOA and NHK was used.

This corpus was automatically transcribed using our bootstrap system, and a DTW algorithm was used to align the automatic speech recognition outputs with corresponding HTML page content, discarding parts in which no words were aligned.

The original HTML pages contain 18k words. We did a forced alignment between those words and the automatic transcription, obtaining a confidence alignment score for each word. We also partitioned the audio file into segments. We only kept segments containing words aligned with a mean confidence measure greater than or equal to 40%. The remaining "approximate" corpus contains about 11k words.

This corpus was further manually corrected by a native Korean speaker. After correction, the corpus contains about 16k words. The original HTML pages contained many English words—much of the data came from an English learning show—which were discarded in the "approximate" corpus. Other sentences were just an abstract of the speech. We report results using both the unsupervised and corrected test set.

Unlike Japanese and Chinese, Korean writing places spaces between adjacent word phrases, providing an explicit word boundary segmentation. However, it seems that the Korean language allows some flexibility in the location of word separators. Two transcribers will not necessarily segment the text in the same way. An example is shown in Figure 1. In this example, we can see that spaces are not placed at the same locations in REF (original LDC transcript) and HYP (the references made by our transcriber). There is a 50% WER (Word Error Rate) and 8% CER (Character Error Rate) difference between the two manual transcriptions. While word segmentation is meaningful in Korean, this inconsistency in annotator agreement leads us to believe CER is a better evaluation metric for this corpus.

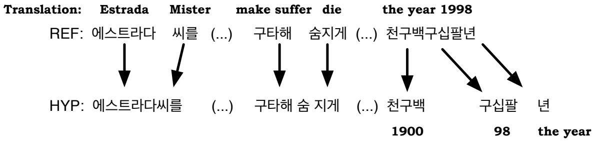
Translation: Estrada Mister make suffer die the year 1998
REF: 에스트라다 씨를 (...) 구타해 숨지게 (...) 천구백구십팔년
HYP: 에스트라다씨를 (...) 구타해 숨 지게 (...) 천구백 구십팔 년
1900 98 the year

Figure 1: Example of differences between LDC (REF) and Korean transcriber (HYP)

Table 1: Korean small phone set.

Type	Phones (Sampa format)
non speech	silence, filler, breath_noise
consonants	p, t, k, C, s, h, w, y, r, l, m, n, G
vowels diphthongs	i, e, a, o, u E, O, A, U

3. Automatic Speech Recognition System

3.1. Phone set and acoustic units

Each of the Korean syllables describes a sequence of phonetic units. These phonetic units consist of 16 basic consonants and 5 double consonants (formed by doubling one the basic consonants). There are also 9 basic vowels and 12 complex vowels. Each complex vowel is either a diphthong of two basic vowels or a basic vowel followed by a semi-vowel offglide. Loan words and Chinese characters may not follow this structure, but we still represent them with the same set of phonetic units in this corpus.

In this work, we consider two possible phone sets that differ only in their use of doubled consonants. While the Korean written language describes strong consonants by doubling them, there is no corresponding symbol in the IPA. Our small phone set (Table 1) simply replaces each double consonant with a single consonant. Our large phone set (Table 2) maintains the distinction by representing each double consonant with a special symbol. The final phone set consists of 9 vowels, 13 consonants (or 21 consonants in the large phone set), and 3 extra units for silence, breath, and filler words.

3.2. Acoustic Models

Prior to using the supervised training set, the LDC BN data needed to be segmented and aligned with the original transcripts. Since we did not have access to initial Korean models, the first set of acoustic models were initialized through language transfer. The Korean phones were matched to their closest counterpart in English. Context-independent English phone models served as seed models. An initial segmentation of the

Table 2: Korean large phone set.

Type	Phones (Sampa format)
non speech	silence, filler, breath_noise
consonants doubled	b, p, d, t, g, k, C, s, h, w, y, r, l, m, n, G pp, tt, kk, CC, ss
vowels diphthongs	i, e, a, o, u E, O, A, U

[2] http://www.voakorea.com/

[3] http://www.rfa.org/korean/

[4] http://www.nhk.or.jp/korean/

data was produced by the seed models. This process was iterated several times, gradually increasing the size of the final models. To verify the process produced reasonable segmentations, the development data was decoded, giving a WER of 35.5 (12.4% CER).

Once the training data had been segmented, we built new acoustic models using the Kaldi speech recognition toolkit [21]. The models were trained from a flat start. For acoustic features standard cepstral features (perceptual linear prediction - PLP) were used. The PLP feature vector has 39 cepstral parameters: 12 cepstrum coefficients and the log energy. An additional estimated pitch feature was also included. The features are transformed using LDA (with a nine frame window) and MLLT is also applied.

The acoustic models are tied-state, left-to-right context-dependent, HMMs with Gaussian mixtures. The triphone-based context-dependent phone models are word-independent, but word position-dependent. The supervised model has 2016 tied states and the model trained with the unsupervised transcripts has 4893. Speaker adaptive training (SAT) is also applied. Decoding is performed using a two-pass approach with a 4-gram language model.

We also experimented with a DNN-based acoustic model. Training was performed using the DNN training recipe within Kaldi [22]. In all cases, a nine frame input window was used, giving a input layer of 360 features. When training on only the 9 hours of transcribed data, a smaller 2-layer DNN was used. The network had approximately 1.5 million parameters (570 nodes per hidden layer). Given the additional unsupervised transcripts, the DNN was expanded to four layers and 7 million parameters (910 nodes per hidden layer).

3.3. Language Models

While we were unable to align all of the acoustic data downloaded from the NHK, VOA, and RFA websites with the associated HTML text, we could still use that text for building language models. The following text data from the LDC was also used: newswire (Newswire 2), telephone conversation transcripts (TEL), and broadcast news transcripts (BN). An initial two million word vocabulary was identified by pooling all the data together. For each separate set of data, 2,3,and 4-gram language models were built. Combined language models were trained by interpolating the language models from each set of data. Mixture weights were automatically determined through an EM algorithm that minimized the perplexity of a set of held-out development data. Final interpolation weights can be seen in Table 3.

The full two million word vocabulary is quite large and results in many n-grams with low counts. Since our ultimate measure was CER, capturing the word phrases was not required, and infrequent words are also likely to be erroneous. By reducing the vocabulary to 200k words, we were able to both reduce the size of the models and increase the counts for higher-order n-grams. Based on the unigram counts, the most common 200k words were selected. Any word that was not a member of the reduced vocabulary set was decomposed into smaller units from the 200k vocabulary. The most probable decomposition based on the original 4-gram LM was selected. A small subset of words (36k) in the original two million word vocabulary did not have a valid decomposition, so the remained out-of-vocabulary. Given the mapping and reduced vocabulary, all words in the training data were mapped to the reduced vocabulary set.

The LDC distributes a 25251-entry lexicon (LDC2003L02 Korean Telephone Conversations Lexicon) covering the words in the corpus of telephone conversations (LDC2003T08 Korean Telephone Conversations Transcripts). In addition to this lexicon, the LDC also generates a tool to generate phonetic pronunciations for unseen words. We used this tool to generate pronunciations for any words in the original lexicon. We also used the tool to identify illegal symbols and sequences of symbols.

Table 3: Amount of training texts and interpolation weights for the component language models.

Data source	#words	4-gram
Newswire 2	55M	.334
NHK	5.5M	.459
RFA+VOA	70k	.039
LDC BN	70k	.163
LDC TEL	230k	.003

4. Unsupervised Training Approach

Our approach to unsupervised training is similar to previous work [1, 9, 8, 23]. Given an initial model trained on 9 hours of transcribed speech, the untranscribed audio is automatically transcribed with the speech recognition system to produce an "approximate" transcript. All training was performed in a single batch. An alternative is to iteratively train and decode using the untranscribed data with increasing amounts of data. The main drawback to this approach is the additional computational cost. Also, the iterative approach was tried in a previous study with this data [12], but did not produce an improvement over training on all of the data at once.

We do not train on all of the untranscribed data. Instead we only use the data that passes a certain confidence threshold. For each word in the transcript, there is an associated confidence value. We determine the confidence for the entire segment by taking the geometric mean of the individual word confidence scores. Three thresholds were tested: 0.7 (88% of the corpus), 0.8 (80%), 0.9 (60%).

Table 4 shows the results for both supervised and unsupervised training using the GMM and DNN acoustic models. In all cases, the transcripts for the unsupervised training are coming from the supervised DNN system and uses the reduced phone set. Based on these results, we use a confidence threshold of 0.7 for all GMM systems and 0.9 for all DNN systems in the remainder of the paper. We can also notice that the CER computed on the "approximate" transcripts follows the CER computed one the "manual" ones.

Table 4: CER using the 200k words LM

Audio trn Sources	Threshold	Acoustic model	manual CER	approx. CER
LDC	-	GMM	17.4	25.7
LDC	-	DNN	15.2	26.0
LDC+Web	0.7	GMM	17.0	24.5
LDC+Web	0.8	GMM	17.2	24.5
LDC+Web	0.9	GMM	17.4	24.9
LDC+Web	0.7	DNN	15.0	23.4
LDC+Web	0.8	DNN	14.8	23.4
LDC+Web	0.9	DNN	14.2	23.1

5. Experimental results

We tested four different models for generating the "approximate transcripts" by varying the acoustic model and phone set. The approaches are referred to in the remainder of the paper as

- *unsup1*: DNN acoustic model with the small phone set,
- *unsup2*: GMM acoustic model with the small phone set,
- *unsup3*: DNN acoustic model with the large phone set,
- *unsup4*: GMM acoustic model with the large phone set.

We wanted to compare results obtained with the "approximate" transcripts with the results obtained with the corrected ones. Results are presented in Table 5 and 6.

The first two lines refer to models trained only on the transcribed 9-hour LDC corpus. The acoustic model refers to the type of acoustic model used during decoding. In all cases, the CER refers to the corrected test set.

Table 5: CER using the 200k words LM on the approximate test corpus

Unsupervised decoding	Acoustic Model	CER with small phone set	CER with large phone set
-	DNN	25.7	25.0
-	GMM	26.0	25.8
unsup1	DNN	23.9	23.5
unsup1	GMM	25.6	25.4
unsup2	DNN	23.4	23.0
unsup2	GMM	25.4	24.6
unsup3	DNN	23.7	23.2
unsup3	GMM	25.6	24.7
unsup4	DNN	23.2	**22.8**
unsup4	GMM	25.2	25.2

Table 6: CER using the 200k words LM on the corrected test corpus

Unsupervised decoding	Acoustic Model	CER with small phone set	CER with large phone set
-	DNN	15.2	14.5
-	GMM	17.4	17.0
unsup1	DNN	14.2	13.3
unsup1	GMM	17.0	16.3
unsup2	DNN	12.7	11.5
unsup2	GMM	16.0	15.0
unsup3	DNN	14.1	13.3
unsup3	GMM	17.0	16.4
unsup4	DNN	12.5	**11.2**
unsup4	GMM	15.7	14.9

As expected, the DNN acoustic models outperform the GMM models, regardless of the phone set. The large phone set also always outperforms the small phone set. While the DNN models are more accurate, all systems perform better when trained on "approximate" transcripts from the GMM system. Best results are obtained using the DNN acoustic model with the large phone set trained on the GMM generated "approximate" transcripts. Overall improvement in CER with the best unsupervised system is 3.3% absolute compared to the best supervised system.

The tendency is exactly the same using the "approximate" and the corrected test corpus.

Our initial hypothesis was that cross-model adaptation—training on transcripts generated from a different acoustic model—would provide an improvement in over all performance. The DNN results match this hypothesis, but not the GMM results. The GMM systems also perform best when using "approximate" transcripts generated from a GMM system. Since the GMM and DNN systems makes different errors, perhaps the errors made by the GMM system are still more similar acoustically to the correct result. We will investigate this further in future work.

6. Conclusion

In this work, we use the Korean dataset originally developed in [12]. We improve upon the results in the previous work in several ways. The previous work used an automatically generated test set; we instead report results on a manually corrected version of this test set. While the relative results are similar, the manually corrected test set provides a more accurate gauge for system performance. We use a larger phone set in this work that significantly improves performance for the final unsupervised systems.

Multiple methods for generating "approximate" transcripts for unsupervised training were explored. Both traditional GMM acoustic models and state-of-the-art DNN models were used. While the DNN systems were the most accurate, the best results were obtained by using the transcripts generated by the GMM systems for unsupervised training. We plan to further investigate this phenomenon in future work.

7. References

[1] L. Lamel and B. Vieru, "Development of a speech-to-text transcription system for finnish," in *Workshop on Spoken Languages Technologies for Under Resourced Languages (SLTU 2010)*, Penang, Malaysia, 2010, pp. 62–67.

[2] O. Kimball, C. Kao, R. Iyer, T. Arvizo, and J. Makhoul, "Using quick transcriptions to improve conversational speech models," in *Proceedings of International Conference on Spoken Language Processing (ISCA, Interspeech)*, 2004, pp. 2265–2268.

[3] C. Cieri, D. Miller, and W. K., "The fisher corpus: a resource for the next generations of speech-to-text," in *Language Evaluation and Resources Conference (LREC)*, 2004, pp. 69–71.

[4] L. Lamel, J.-L. Gauvain, and G. Adda, "Investigating lightly supervised acoustic model training," in *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP)*, vol. 1, 2001, pp. 477–480.

[5] C. Collan, M. Bisani, S. Kanthak, R. Schlüter, and H. Ney, "Cross domain automatic transcription on the tc-star epps corpus," in *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP)*, vol. 1, 2005, pp. 825–828.

[6] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," in *Speech Communication*, vol. 50, 2008, pp. 434–451.

[7] F. Wessel and H. Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," in *IEEE Transactions on Speech and Audio Processing*, 2005, pp. 23–31.

[8] J. Ma and R. Schwartz, "Unsupervised versus supervised training of acoustic models," in *Proceedings of International Conference on Spoken Language Processing (ISCA, Interspeech)*, 2008, pp. 2374–2377.

[9] L. Lamel, J.-L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model trainings," *Computer Speech and Language*, vol. 16, pp. 115–129, 2002.

[10] S. Novotney and R. Schwartz, "Analysis of low-resource acoustic model selg-training," in *Proceedings of International Conference*

on Spoken Language Processing (ISCA, Interspeech), 2009, pp. 244–247.

[11] S. Novotney, R. Schwartz, and J. Ma, "Unsupervised acoustic and language model training with small amounts of labelled data," in *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP)*, 2009, pp. 4297–4300.

[12] A. Laurent and L. Lamel, "Development of a korean speech recognition system with little annotated data," in *International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU'14)*, 2014.

[13] S. P. Rath, D. Povey, K. Vesel`y, and J. H. Cernock`y, "Improved feature processing for deep neural networks," *Proceedings of International Conference on Speech Communication and Technology (ISCA, Interspeech 2013)*, 2013.

[14] G. Hinton, L. Deng, D. Yu, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, 2012.

[15] A. L. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust asr," *Proceedings of International Conference on Speech Communication and Technology (ISCA, Interspeech 2012)*, pp. 20–25, 2012.

[16] F. Weninger, M. Wollmer, J. Geiger, B. Schuller, J. Gemmeke, A. Hurmalainen, T. Virtanen, and G. Rigoll, "Non-negative matric factorization for highly noise-robust asr : To enhance of to recognize ?" *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 2012)*, pp. 4681–4684, 2012.

[17] O.-W. Kwon and J. Park, "Korean large vocabulary continous speech recognition with morpheme-based recognition units," *Speech Communication*, vol. 39, pp. 287–300, 2003.

[18] D. K., T. Schultz, and A. Waibel, "Data-Driven Determination of Appropriate Dictionary Units for Korean LVCSR," in *Proceedings of International Conference on Speech Processing (ICSP'99)*, Seoul, Korea, 1999, pp. 323–327.

[19] S. Strassel, N. Martey, and D. Graff, "Korean broadcast news speech & korean broadcast news transcripts, ldc2006s42, ldc2006t14, 2006," Linguistic Data Consortium, 2006.

[20] T. Schultz, "Globalphone: a multilingual speech and text database developed at karlsruhe university." in *Proceedings of International Conference on Spoken Language Processing (ISCA, Interspeech)*, 2002.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *Proceedings of IEEE ASRU*, 2011.

[22] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *Proceedings of IEEE ICASSP*, 2014.

[23] L. Lamel, J.-L. Gauvain, and G. Adda, "Lightly supervised acoustic model training," *ITRW ASR*, vol. 1, pp. 150–154, 2000.