DEVELOPMENT OF A KOREAN SPEECH RECOGNITION SYSTEM WITH LITTLE ANNOTATED DATA

Antoine Laurent, Lori Lamel

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

ABSTRACT

This paper investigates the development of a speech-to-text transcription system for the Korean language in the context of the DGA RAPID Rapmat project. Korean is an alpha-syllabary language spoken by about 78 million people worldwide. As only a small amount of manually transcribed audio data were available, the acoustic models were trained on audio data downloaded from several Korean websites in an unsupervised manner, and the language models were trained on web texts. The reported word and character error rates are estimates, as development corpus used in these experiments was also constructed from the untranscribed audio data, the web texts and automatic transcriptions. Several variants for unsupervised acoustic model training were compared to assess the influence of the vocabulary size (200k vs 2M), the type of language model (words vs characters), the acoustic unit (phonemes vs half-syllables), as well as incremental batch vs iterative decoding of the untranscribed audio corpus.

Index Terms— Speech recognition system, unsupervised acoustic training, korean, approximative transcripts

1. INTRODUCTION

Large Vocabulary Continuous Speech Recognition (LVCSR) systems are traditionally trained on large amounts of carefully transcribed speech data and huge quantities of written texts [1]. However, obtaining such training corpora is quite costly and requires expertise (generally via a native speaker) in the targeted language. One of the most frequently cited costs is that of obtaining this necessary transcribed acoustic data, which is an expensive process both in terms of manpower and time. Although there are ever-increasing amounts of audio data available from a variety of sources (radio, television, web, ...), for the vast majority there are no corresponding accurate word transcriptions [2]. Several research directions have addressed reducing the data production time and costs [3] and some audio training data, such as those produced within the DARPA EARS program, are associated with quick transcripts [4]. For some audio sources there are also associated texts, such as closed captions, summaries or other less closely related texts. A variety of methods have been investigated to use such resources for what is called lightly-supervised or unsupervised acoustic model training [5]. Most proposed methods rely on supervision from a language model. The different approaches vary on the use or not of confidence factors [6, 7, 8], on the use of iterative or doubling training [9] and on the supervision level [2]. [10, 11] present an analysis of training behavior for supervised and unsupervised approaches.

In this study, 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 text data (also from LDC or web downloads). We explored building several systems using different language models (LM) (in terms of vocabulary size and using chars instead of words) for the unsupervised training and for the decoding, using phonemes and "half-syllables" as acoustic units and using two different approaches for the unsupervised acoustic training. This speech recognition system will be used for the RAPMAT (Speech translation) project [1].

The next section overviews the characteristics of the Korean language, followed by a description of the approach and corpus used in this study. This is followed by a description of the language models and vocabulary selection, the phone set and acoustic models, after which experimental results are provided.

2. KOREAN LANGUAGE

For over a millennium, Korean was written with adapted Chinese characters called *Hanja*. In the 15th century a national writing system called *Hangeul* was proposed, and was completely adopted in the 20th century. As presented in [12], in

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

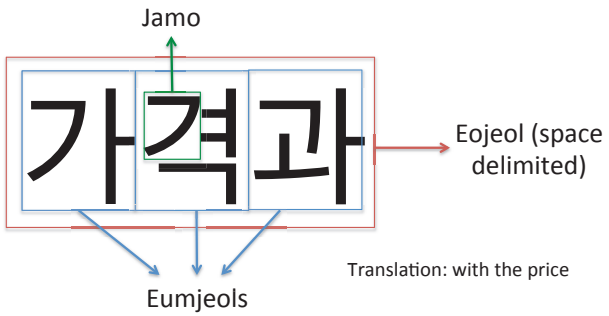

Fig. 1. Example of a Korean eojeol

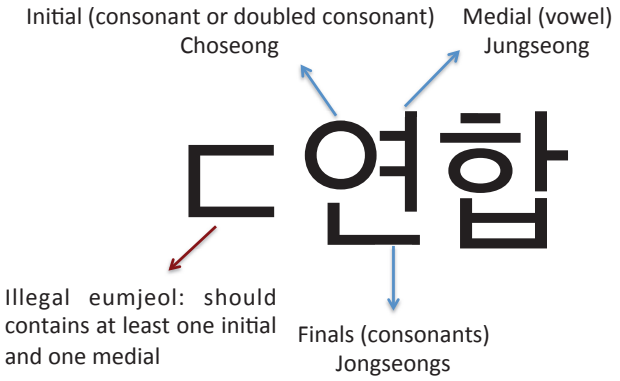

Fig. 2. Example of an illegal Korean eojeol

Korean writing, a space is placed between two adjacent word-phrases, each of which generally corresponds to two or three words in English in a semantic sense. It is an alpha-syllabary system [13]. As described in [14], sets of jamo (orthographic phoneme segments) are grouped into eumjeols (orthographic syllables), and sequences of eumjeols are grouped into eojeol (space-delimited orthographic words) (see figure 1).

Each eumjeols is composed of two or three elements: the choseong (the initial consonant), the jungseong (the vowel) and an optional jongseong (the final consonant), as represented in figure 2. (This illegal eojeol is present inside our corpus).

Most of the reported speech-to-text transcription results for the Korean language are substantially worse than reported performances on more resourced languages such as English or French. At least two factors contibute to this performance difference. Since eojeols generally represent more than one word, the vocabulary of speech recognition systems based on words (defined as space-separated elements) should contain a high number of entries for Korean [15]. 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. A proposed solution to this problem is the decomposition of words into morphs as proposed in [12, 16]. Another factor is the lack of suitable speech and text resources for model training. Only a few (relatively) small corpora are available for Korean via LDC or ELDA [17] and to the best of our knowledge, most of the previous speech recognition systems were built using undistributed internal data.

3. APPROACH AND CORPUS

The general approach taken in this work is similar to that of [1, 2, 9, 18] in that a speech recognizer is used to provide "approximate" transcripts for acoustic model training. The audio data is transcribed in incremental batches, and in successive iterations the models are trained on more data. We also propose another approach in which the whole available audio data set is iteratively transcribed and models improved. Our approaches were developed and tested on a corpus for which we do not have an exact transcription. Because nobody in our laboratory speaks or understands Korean, we do not know if this corpus, extracted from news websites, is close to the speech data extracted from the same websites.

In this study, different approaches for unsupervised training were explored in order to assess the overall impact on performance. Different sets of audio and text data were used for model training. First, we used an LDC corpus that contains Korean broadcast news transcripts and speech. This corpus only contains a small amount of data (9 hours – 70k words). 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. Additional data coming from three Korean news websites were also used: VOA[2], RFA[3] and NHK[4]. For NHK about 400 hours of data with approximate transcripts (5.5M words) were downloaded dating from November 2007. The approximate transcripts correspond to the HTML content for each news brief associated with an audio file. Data have been downloaded from the two other sources only since October 2013: we obtained 5 hours of audio data from the RFA website and 4 hours from the VOA website. Although the texts accompanying the audio data are considered as rough transcripts in general they only cover a small part of the audio and are not aligned with it. Therefore we could only use the audio data for unsupervised training of the acoustic models. A portion of the collected data from each source (RFA, VOA, and NHK) was reserved as an "approximate" development corpus. This corpus was automatically transcribed using our bootstrap system, and a DTW algorithm was used to align the automatic speech recognition outputs with the HTML page content, discarding parts in which no words were aligned. The selected development data contain about 10k words.

[2]http://www.voakorea.com/
[3]http://www.rfa.org/korean/
[4]http://www.nhk.or.jp/korean/

Table 1. Amount of training texts and interpolation weights for the component language models.

Data source	#words	4-gram
Newswire 2	55M	.197
NHK	5.5M	.750
RFA+VOA	70k	.042
LDC BN	70k	.010
LDC TEL	230k	.001

4. LANGUAGE MODELS

The texts from the LDC (Korean broadcast news transcripts and speech BN, telephone conversation transcripts TEL and newswire 2) and the texts from the NHK, VOA and RFA websites were used to build the language models. Component language models were estimated on each subset of training texts using a 2 million word vocabulary selected on the pooled data. The full 2, 3, and 4-gram language models were then obtained by interpolation of the back-off n-gram language models using the modified Kneser-Ney smoothing. Table 1 gives the interpolation weights for the LMs trained on the different data sources. The mixture weights were automatically determined using the EM algorithm to minimize the perplexity of a set of LM development data. This development corpus of 100k words is composed of portion of the available data from each sources NHK (87k words), VOA (2k words), RFA (2k words) and LDC broadcast news transcripts (8k words). Due to the high proportion of the NHK data in the text development corpus, this component has the highest contribution. Although speech transcripts generally have a very high weight, we believe their low contribution is due to their low representation (non for the LDC telephone speech corpus) in the development set.

An initial 200k word vocabulary was selected using the most probable words in the interpolated 1-gram word language model and a character LM was built. Table 2 provides some statistics for the 200k and for the 2M word vocabularies. For the 2M word dictionary, 46240 pronunciations are shared by more than one word, resulting in a total of 102085 (5.1%) homophones. The same tendency can be observed for the 200k word vocabulary. It is notable that for the 2M word vocabulary, up to 14 words can share the same pronunciation, which is double that reported for English (only 4 words) and French (7 words) [19, 20].

Normalization of the Korean texts was a large part of this work. We did not have any knowledge about this language, and used the available literature to help us clean the texts. We defined a list of obsolete and non EUC-KR characters (using http://en.wikipedia.org/wiki/Hangul), and removed the corresponding sentences from the training corpus. Sentences containing lines with illegal symbol sequence or containing symbols with only one jamo were also removed, as were sentences containing English words. After noticing that Korean texts contain many different separation characters, a set of 122 such characters were identified and then replaced by a space character.

Table 2. Vocabulary statistics. Average (avg) word length in phonemes and symbols, average number of pronunciations, homophone and maximum (max) homophone set size.

Vocabulary size	200k	2M
Avg. # phonemes/word	9.54	11.96
Avg. # symbols/word	3.87	4.81
Avg. # prons/word	1.04	1.06
# homophones	5.23%	5.10%
max homophone set size (#words)	9	14

Table 3. Korean 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	`i`, `e`, `a`, `o`, `u`
diphthongs	`E`, `O`, `A`, `U`

The LDC distributes a 25251-entry dictionary (LDC2003L02 Korean Telephone Conversations Lexicon) covering the words in the corpus of telephone conversations (LDC2003T08 Korean Telephone Conversations Transcripts). This dictionary is accompanied by a tool to automatically generate phonemic transcriptions for unseen words. We also used this tool to generate pronunciations for all lexical entries, and also to find illegal sequences of symbols.

5. ACOUSTIC MODELS

5.1. Phone set and acoustic units

Words of foreign origin excluded, Korean is written with 14 basic consonants and 5 double consonants formed from the basic consonants. There are 9 basic vowel sounds and 12 additional complex vowel sounds. These complex vowels are diphthongs and are comprised of two basic vowels or a sequence of a vowel and a semi-vowel offglide. Korean words are written from left to right and words are formed by writing each syllable in a block-like shape.

The phone set used in this work is composed of the 25 phonemes shown in Table 3. The Korean written language indicates strong (fortis) consonants by doubling them, however as there is no symbol in IPA to indicate this, we decided to not have a special phoneme for them. So each double (fortis) consonant is replaced by a single instance. In total there are 13 consonants, 9 vowels and 3 extra units for silence, breath and filler. For this preliminary work we did some simplifications but further experiments will be run with a bigger phone set to see the impact on the ASR performances.

Fig. 3. Example of phone and half-syllable units.

We also explored using half-syllable units (instead of phones) for acoustic modeling. This was inspired by the structure of Korean and published work using initial-final models for Chinese [21, 22]. We used the phonemic representation of each symbol and merged the component phonemes. Figure 3 illustrates the both the phone and half-syllable units. The first part of the syllable (the onset) contains all phonemes preceding the vowel. The second part is the remainder, that is the vowel and any ensuing consonants. Due to the structure of Korean, the half-syllable representation results in a set of 97 acoustic units. These models correspond to a single phone (21) or a sequence of two phones (74).

5.2. Acoustic modeling

The acoustic models were initialize via language transfer. Phones from English were associated with the Korean phones, and the corresponding context-independent models served as initial seed models. These models were used to segment the manually transcribed LDC BN data and a first set of Korean models were built. Several iterations of segmentation and model estimation were carried out, gradually increasing the model size. With under 9 hours of training data, only small models could be built. As a first check, these models were used to assess the word error rate (WER) on the 2 set-aside VOA files (1.5 hours) from LDC. An initial WER of 37.0% (18.5% of CER) was obtained, however this number is very optimistic as the training and dev data are extremely close.

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, along with the first and second derivatives. 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 tied states are obtained by means of a decision tree, where the 92 questions which concern the phone position, the distinctive features (vowel, consonant, nasal, stop, fricative, rounded, front, low, ...) and identities of the phone and the neighboring phones. The acoustic models are gender-independent. Silence is modeled by a single state with 1024 Gaussians.

In order to create the half-syllable units, the phone segmentations were merged, and new models trained.

As mentioned earlier several different configurations were explored for unsupervised training. In addition to comparing phone and half-syllable acoustic units, different size language models were used to decode the untranscribed data, and the incremental approach based on [1] has been compared with batch decoding of the full set of data. In the incremental approach the seed models are used to decode a small set of data, and the resulting transcripts are considered as ground truth references, and used instead of manual transcripts for segmentation. The amount of training data is roughly doubled at each iteration, allowing successively larger and more accurate models to be estimated. The new models are then used in the next decoding iteration. For the batch decoding, the seed models are used to decode all of the data, and a first set of models are estimated. All of the data are redecoded several times.

6. EXPERIMENTAL RESULTS

As mentioned earlier, one of the difficulties for Korean speech recognition is that the vocabulary size is very large. Table 4 shows the the Out-Of-Vocabulary (OOV) rates computed on the development texts with the 200k and 2 million word lists. With the 200k word list, the OOV rate is almost 10%, and is still 3% with 2 million words.

Table 4. PPL and OOV rate of the 200k and 2M word language models on the development text corpus (100k words)

Modele	PPL	OOV rate (%)
2M words 4g LM	732	3.0
200k words 4g LM	1596	9.7

Our test corpus is composed of 3.5 hours of data coming from RFA, VOA and NHK. Once discarded part that we were not able to align (using DTW and automatic transcripts, see section 3), it remains about 10k words (1.5 hour).

Table 5. Approximate WER and CER using 200k LM for decoding The column headings specify the LM used for unsupervised training.

Audio trn Sources	hours	200k LM WER	200k LM CER	2M LM WER	2M LM CER
LDC	9	50.6	32.1	50.6	32.1
Web	10	50.2	32.8	50.1	32.6
LDC+Web	19	49.1	31.5	48.8	31.0
LDC+Web	34	48.7	30.4	47.7	30.4
LDC+Web	79	48.0	29.9	47.6	29.7

Table 5 gives transcription results in terms of WER and CER (Character Error Rate) when decoding the test data with

the 200k LM. The column headers **200k LM** and **2M LM** correspond to the LM used for the unsupervised training. It can be seen that the WER is quite high. This can be explained in part by the fact that the references for the development corpus are only approximate, and also that word segmentation seems to be somewhat variable for the Korean language. As mentioned earlier, each Korean "word" is composed of on average about 4 symbols (see Table 2), an error in any symbol will engender a word error. [12]. The character error rate (CER) is also given. Comparing the first two lines, it can be seen that using about the same quantity of LDC (supervised) and Web (unsupervised) data gives almost the same WER and CER. This can be attributed to the fact that there is a temporal and source mismatch between the LDC data and the multi-source development corpus. As presented in section 5.2, first experiments using only LDC data as testing corpus, show a better WER and CER (37.0% and 18.5%), this can be explain by the fact that Web data included in the multi-source development corpus are noisy (approximative transcripts) and distant from the LDC data. Each iteration, roughly doubling the data, is seen to reduce both the WER and CER. Comparable performances are obtained for decoding the audio training data with the two language models.

Table 6. Approximate WER and CER using 2M LM for decoding. The column headings specify the LM used for unsupervised training.

Audio Sources	hours	200k LM WER	200k LM CER	2M LM WER	2M LM CER
LDC	9	47.5	29.4	47.5	29.4
Web	10	48.1	30.5	46.6	29.5
LDC+Web	19	45.2	28.2	42.3	27.0
LDC+Web	34	44.7	27.7	41.1	25.7
LDC+Web	79	44.0	27.0	40.2	25.4
LDC+Web	150	43.1	26.4	39.9	25.2

Table 6 shows the results obtained when using the 2M word LM to decoding the development set. The WER with the 2M word LM are about 10% lower than those with the 200k LM. As more data are added, the WER and CER decrease. It can also be seen that when using the 2M word LM to decode the development data, a larger difference is seen as a function of the LM used during unsupervised training with about a 4% absolute difference in the last two iterations.

Some experiments were also carried out using a character LM for the unsupervised training. The best acoustic models were used to decoded the dev data with the 2M word LM and the character LM. Decoding with the 2M word LM resulted in a WER of 44.5%. Decoding with the character LM and the 2M word LM give quite comparable CER as can be seen in Table 7.

Table 8 shows results obtained using the full set of untranscribed audio data during each iteration of unsupervised training. It can be seen that almost all the decrease in WER and CER is gained in the first iteration, after which there is little improvement. It is notable that we used much more data than for the other unsupervised training method.

Table 7. Approximate WER and CER using the character LM for the unsupervised training. The column headers specify the LM used to decode the development data.

Audio Sources	hours	2M LM WER	2M LM CER	Char LM CER
LDC+Web	79	44.5	26.7	27.0
LDC+Web	150	44.0	26.4	26.5

Table 8. Approximate WER and CER using the full training corpus during each unsupervised training iteration.

Audio Sources	Iteration	2M LM WER	2M LM CER
LDC 9h	0	47.5	29.4
LDC+Web 400h	1	39.7	25.3
LDC+Web 400h	2	39.6	25.3
LDC+Web 400h	3	39.5	25.1

Table 9 reports results using the half-syllable acoustic units. Both the WER and CER are significantly higher with these units than with the phone units. This is in contrast with reported results on Mandarin Chinese.

Table 9. WER and CER using 200k LM for decoding and unsupervised training with half-syllable acoustic units

Audio Sources	hours	200k LM WER	200k LM CER
LDC+Web	19	59.3	41.8
LDC+Web	79	53.8	34.5
LDC+Web	150	52.4	32.0

Experience with other (European) languages in the Quaero program showed larger WER decreases than we have observed here. We think that part of this lack of improvement for the Korean language is due to the use of approximate transcripts for the development corpus. We have recently been able to hire a native Korean speaker, and we will have a fully annotated development corpus within the coming months.

We have some preliminary results on a few minutes of speech that the Korean speaker has transcribed. It seems that the Korean language allows some flexibility in the location of word separators. So two transcribers will not necessarily segment the text in the same way. An example is shown in Figure 4. In this example, we can see that spaces are not placed at the same places in REF (original LDC transcript) and HYP (the references made by our transcriptor). There is a 50% WER and 8% CER difference between the two manual transcriptions.

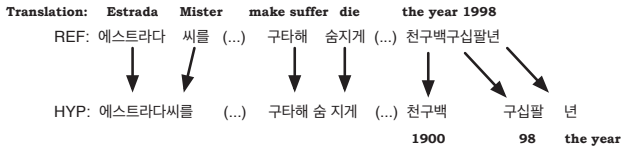

Fig. 4. Example of differences between LDC (REF) and Korean transcriber

The transcriber also corrected a few of the hardest files from the development set (those with the highest CER) according to the approximate transcripts. The CER of the best models was 28.8% with the approximate transcript, and is reduced to 18.7% after correction. We also scored some intermediary model sets and the same tendencies are observed as with the approximate ones. With this small sample, adding more data gives a larger relative CER reduction with the corrected transcripts (12%) than that measured with the approximate ones (7%).

7. CONCLUSION

This paper has described the development of a speech to text system for the Korean language for use in the RAPMAT project. As only very small amounts of annotated data were available, additional audio data were used in an unsupervised manner to improve the acoustic models. A subset of this data was also selected for use as development data for which only approximate transcripts were created by comparing associated web texts to the recognition hypotheses. The transcripts results show a decrease in terms of WER and CER when adding more audio data. The two unsupervised training strategies (incremental vs full batch processing) gave somewhat comparable results with the same quantity of data.

A native Korean speaker has just joined the team, and the preliminary results indicate that the proposed method based on approximate references seems to works well.

Since the normalization of the textual data was improved since the beginning of this work, we plan to repeat some of the unsupervised training with the new LMs. We also hope to have a better improvement using the unsupervised training method with an annotated development corpus. We will also train multilayer perceptron (MLP) acoustic models to improve the recognition accuracy as has been observed for other languages.

8. 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] 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.

[3] O. Kimball, C.L. 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.

[4] C. Cieri, D. Miller, and Walker 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.

[5] 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)*, 2001, vol. 1, pp. 477–480.

[6] 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)*, 2005, vol. 1, pp. 825–828.

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

[8] 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.

[9] 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.

[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] 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.

[13] I. Taylor, "The korean writing system: An alphabet? a syllabary? a logography?," in *Proceeding of Visible Language*, New York, USA, 1979, vol. 2, pp. 67–82.

[14] J. Lee and G. G. Lee, "A data-driven grapheme-to-phoneme conversion method using dynamic contextual converting rules for korean tts systems," *Computer Speech and Language*, vol. 23, pp. 423–434, 2009.

[15] Daniel 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.

[16] I-J. Choi, N.-H. Kim, and S. Y. Yoon, "Large vocabulary continuous speech recognition based on cross-morpheme phonetic information," in *Proceedings of International Conference on Spoken Language Processing (ISCA, Interspeech)*, 2004, pp. 453–456.

[17] 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.

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

[19] J.-L. Gauvain, L. Lamel, G. Adda, and M. Adda-Decker, "Large vocabulary speech recognition in english and french," in *Automatic Speech Recognition and Understanding (IEEE, ASRU)*, 1993.

[20] M. Adda-Decker and L. Lamel, "Dictionaries for multilingual speech processing," in *Multilingual Speech Processing*. 2006, pp. 123–168, Katrin Kirchhoff and Tanja Schultz, Elsevier.

[21] S. M. Chu, H.-K. Kuo, L. Mangu, Y. Liu, Y. Qin, Q. Shi, S. Zhang, and H. Aronowitz, "Recent advances in the ibm gale mandarin transcription system.," in *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP)*, 2008, pp. 4329–4332.

[22] L. Lamel, J.-L. Gauvain, V. Le, I. Oparin, and S. Meng, "Improved models for mandarin speech-to-text transcription," in *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP)*. IEEE, 2011, pp. 4660–4663.