Automatic Acoustic Segmentation for Speech Recognition on Broadcast Recordings

Gang Peng, Mei-Yuh, Hwang, Mari Ostendorf

Dept. of Electrical Engineering, Univ. of Washington, Seattle, WA 98195, USA

{gpeng,mhwang,mo}@ee.washington.edu

10.21437/Interspeech.2007-742

Abstract

This paper investigates the issue of automatic segmentation of speech recordings for broadcast news (BN) and broadcast conversation (BC) speech recognition. Our previous segmentation algorithm often exhibited high deletion errors, where some speech segments were misclassified as non-speech and thus were never passed on to the recognizer. In contrast with our previous segmentation models, which only differentiated between speech and non-speech segments, phonetic knowledge is applied to represent speech by using multiple models for different types of speech segments. Moreover, the "pronunciation" of the speech segment has been modified to loosen the minimum duration constraint. This method makes use of language specific knowledge, while keeping the number of models low to achieve fast segmentation. Experimental results show that the new segmenter outperforms our previous segmenter significantly, particularly in reducing deletion errors.

Index Terms: speech recognition, automatic acoustic segmentation, machine translation, broadcast news, broadcast conversation

1. Introduction

In the area of automatic speech recognition (ASR), both training and decoding are usually done at the utterance level for efficiency reasons, both in speed and memory. In BN/BC speech recognition, the test data are usually given as entire BN/BC recordings. Therefore, an ASR system needs to first segment the recorded show into utterances, which are typically about a few seconds long surrounded by pauses. In this paper, we refer to such automatically-segmented utterances as "auto utterances".

Table 1. CERs of our 2006 Mandarin recognizer.

	Sub	Del	Ins	Overall
eval04	4.3%	7.6%	0.3%	12.2%
eval06	8.6%	7.6%	1.6%	17.8%

We have previously built a state-of-the-art Mandarin BN/BC speech recognizer [1]. In our error analysis, we found the deletion errors were much more frequent than insertion errors, as indicated by the character error rates (CERs) in Table 1, where *eval04* is the DARPA Rich Transcription 2004 evaluation set, which includes one hour of BN shows, and *eval06* is the DARPA GALE 2006 evaluation set, which includes about one hour of BN shows and one hour of BC shows.

The high deletion error rate was caused partially by the imperfect acoustic models and language models, their interaction with each other and with the decoding structure, but more commonly caused by the suboptimal automatic acoustic segmentation. For example, the following two auto utterances in one CCTV show in *eval04* both had the utterance endings cut too early, indicated by brackets below:

…今年一季度,中国遭遇反倾销同比[增长百分之八十]

"The first quarter of this year, China has experienced [an 80% increase] in anti-dumping"

如果说在调查过程中相关的问题得不到很好的解决, 根据对外贸易壁垒调查规则暂行规定的[这个规定]…

"If the related problems have not been properly solved in the course of investigations, according to [the rules] of the Interim Regulations on survey of Foreign Trade Barriers…"

High deletion errors not only degrade ASR performance, but are also propagated to any downstream processing such as machine translation (MT), leading to increased error rates for MT.

In this paper, we try to attack the most common source of deletion errors: improving automatic acoustic segmentation such that fewer speech segments will be dropped while also minimizing new insertion errors (where noises become recognized as word sequences).

The remainder of this paper is organized as follows. Section 2 explains both the previous and the new acoustic segmentation algorithms. Section 3 introduces the corpora used in this paper. Section 4 describes in detail the experimental results and observations, followed by a summary and further work in Section 5.

2. Segmentation Algorithms

2.1 Previous Segmenter

Our previous segmenter operated without any knowledge of the underlying phoneme sequence contained in the speech waveform. As Figure 1(a)

illustrates, the segmentation problem was treated as a speech recognition problem with a finite-state grammar on a vocabulary of three words: speech, noise, and silence. Segments composed of noise and silence exclusively were discarded. The "pronunciation" of speech is 18 or more consecutive *fg* (foreground) models; that of silence, exactly 2 *bg* (background) models; that of noise exactly 2 *rej* (reject) models, as indicated in Table 2. Each of the *fg*, *bg*, and *rej* models is a 3-state left-to-right hidden Markov model (HMM). Each state was modeled by 300 Gaussians of 39-dimension Mel-frequency cepstral coefficients (MFCCs) and deltas.

After a show is recognized as a sequence of the three words, initial utterance boundaries are defined by locations where there is a long pause. Speaker boundaries are then determined by Gaussian mixture models of MFCC in the manner of speaker clustering. Short consecutive utterances from the same speaker are further merged into a longer utterance. The final utterances average 6 secs long each.

Because of the pronunciation structure of speech, each speech segment had a minimum duration of 54 (18×3) frames, or 540ms approximately (each frame is 10ms long in our system.). This constraint might be too long, especially for BC recordings, where there is often short spontaneous speech, as multiple speakers interrupt each other during the conversation.

2.2 New Segmenter

Our new segmenter makes use of broad phonetic knowledge of the target language, which is Mandarin in our experimental setup. While keeping silence and noise modeling intact, we divide speech into three categories: a Mandarin syllable with a voiceless initial (*I1*), a Mandarin syllable with a voiced initial (*I2*), and a non-Mandarin word (referred as to a foreign word in this context). As shown in Table 2, Mandarin syllables are initials followed by finals (*F*). Mandarin syllables without initials are also modeled by *I2 F*. A foreign word is modeled by two *forgn* models. A BN/BC show is then defined as a sequence of these five words characterized by the grammar in Figure 1(b), and the pronunciation of the vocabulary is defined in the second half of Table 2.

Table 2. Decoding vocabulary for automatic acoustic segmentation.

	Vocabulary	Pronunciation
Previous segmenter	speech	18+ fg
	noise	rej rej
	silence	bg bg
New segmenter	Mandarin1	I1 F
	Mandarin2	I2 F
	Foreign	forgn forgn
	noise	rej rej
	silence	bg bg

To sum up, there are two main improvements in the new segmenter. First, 6 phonetic HMMs are used to model five possible words instead of 3 HMMs for 3 words. This gives more accurate modeling, while the auto-segmentation system remains small and fast. Second, a speech segment can now be as short as one syllable whose minimum duration is only 60ms.

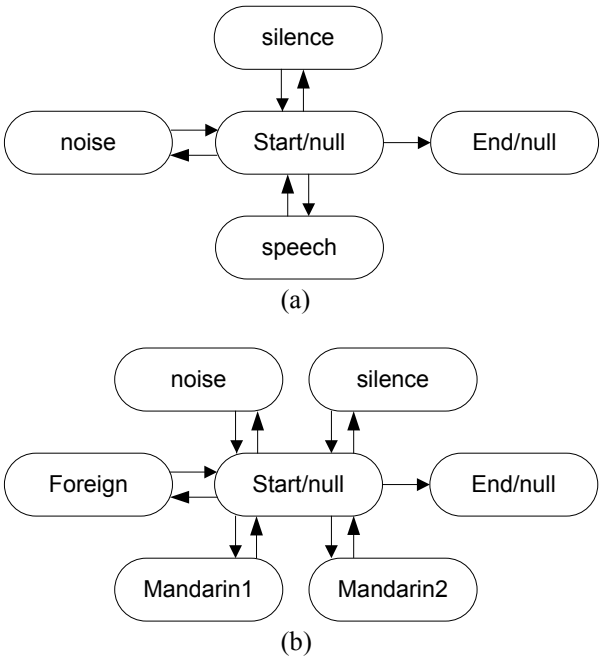


Figure 1. Finite state grammars for our previous segmenter (a) and new segmenter (b).

3. Data Corpora

In this paper, we evaluate our acoustic segmentation algorithms on our state-of-the-art Mandarin large-vocabulary continuous speech recongizer (LVCSR). Our acoustic data for training the final ASR models are all from various LDC Mandarin corpora as listed in Tables 3. There are 465 hours of speech data for training the 2006's acoustic models [1], and 871 hours of speech data for the up-to-date system. However, we use only the *Hub4* data for training our acoustic segmentation models described in Section 2.

Table 3. Acoustic training data, in hours.

Data	Year	BN	BC	Overall
Hub4	1997	29.8		
TDT4	2000-2001	89.1		
GALE Year 1	2004-2006	194.6	152.3	
Sub-total		313.3	152.3	465
GALE Year 2	2006	162.7	242.6	
Total		476	395	871

Table 4 shows the amount of test data used in this paper. The *eval04* and *eval06* are both evaluation data released from LDC directly, while *uwbc06* is defined by randomly selecting 3 BC recordings from GALE Year 2 data (Global Autonomous Language Exploitation), which is a new DARPA-sponsored project, aiming at translating either Mandarin or Arabic

speech into English text. These three recordings were excluded from our training collections. As for the data used for language models (LMs), please refer to [1].

Table 4. Acoustic test data, in hours.

Data	Year	BN	BC	Overall
eval04	2004	1.0		1.0
eval06	2006	1.16	1.05	2.21
uwbc06	2006		1.56	1.56

4. Experimental results and discussion

4.1. Experimental results

The segmenters are first tested in the mode of single-pass speaker-independent (SI) decoding. The segmented auto utterances are decoded by using within-word triphone models with 42-dimension (39-dimension of MFCC and deltas plus 3 pitch features), trained with a minimum phone error (MPE) criterion [2, 3], and a tightly pruned bigram.

Table 5. Single-pass SI decoding results in terms of CER.

	eval04			
	Sub	Del	Ins	Overall
Baseline	9.9%	6.1%	0.9%	16.9%
New segmenter	10.2%	5.5%	1.0%	16.7%
Oracle	10.3%	5.3%	1.0%	16.6%
	eval06			
	Sub	Del	Ins	Overall
Baseline	15.0%	7.2%	2.5%	24.7%
New segmenter	15.2%	5.7%	2.7%	23.6%
Oracle	15.0%	6.0%	2.3%	23.3%
	uwbc06			
	Sub	Del	Ins	Overall
Baseline	19.3%	10.9%	1.4%	31.5%
New segmenter	19.5%	8.9%	1.6%	30.0%
Oracle	19.6%	9.0%	1.5%	30.1%

The SI decoding results are shown in Table 5. The "Baseline" indicates the segmentation was done by our previous segmenter; while "New segmenter" indicates the segmentation was done by our currently developed segmenter. For all 3 test sets, the new segmenter achieves the lowest CERs. The error distributions show clearly that the main benefit of the new segmenter is in the lower deletion errors, indicating that the new segmenter may have extracted more speech segments from the recordings. Those speech segments, discarded by our previous segmenter, usually have lower speech quality in term of signal-to-noise ratio (SNR) or relatively low energy. The new segmenter recovers most of them. These additional segments definitely challenged the recognition system more, leading to the increases in both substitution and insertion errors as shown in Table 5.

Since some of our test sets include some BC recordings, we also experimented with adding 10 hours of BC data to retrain the acoustic segmentation models. Unfortunately, we found the BC data degraded the performance as shown in Table 6. This might result from two possible reasons: the quality of BC data is usually lower than that of BN data due to lower SNR and overlapped speech; and the accuracy of the transcriptions of GALE data is lower than that of Hub4 data.

Table 6. Single-pass SI decoding results with acoustic segmentation models trained with additional BC data.

	Hub4 only	Hub4 + 10 hrs of BC data
eval04	16.7%	17.0%
eval06	23.6%	23.7%
uwbc06	30.0%	31.0%

Moreover, we also tried different model topologies. As is generally known, the duration of Mandarin finals is longer than that of initials. We extended the number of states for finals from 3 to 6, but this also hurt our system similar to the effect of adding BC data.

Finally, the new segmenter is integrated into our recognition system [1], which consists of two sets of acoustic models. One set of the models is trained with MFCC plus pitch features (referred to as the MFCC model). The other set of the models is trained by attaching phone posterior features generated by a multi-layer perceptron (referred to as the MLP model) [4]. The hypotheses produced by those two sets of models are used to cross adapt each other, followed by a second pass decoding to generate two sets of N-best lists, which are then rescored by two 5-gram LMs. The two 5-gram scores are next combined with acoustic scores for system combination via a confusion network [5] at the character level, to output the final recognized character sequence.

In Table 7, the "Baseline" stands for our 2006's recognition system, which uses the previous acoustic segmenter. All acoustic models for that system are trained with 465 hours of speech data. Compared with the insertion error rates, the deletion error rates are much higher. Then we retrain the MLP model with 871 hours of speech data. The corresponding system using the previous acoustic segmenter is denoted as "Baseline*". Although the deletion error rate for *eval04* decreases from 7.6% to 6.1%, an equal increase in substitution and insertion errors occurs, making no contribution to the overall error rate. For *eval06*, the "Baseline*" decreases the overall error rate from 17.8% to 17.2%, and for *uwbc06* from 22.3% to 21.2%.

Finally the segmenter in "Baseline*" is replaced by our new segmenter, and the corresponding system is "New segmenter*" because of the added training data in the acoustic models. As shown in Table 7, the overall error rates for all 3 test sets are further lowered, verifying that the new segmenter indeed contributes an improvement in our best system.

Table 7. Multi-pass recognition results of eval04, eval06, and uwbc06 for 3 different systems.

	Eval04			
	Sub	Del	Ins	Overall
Baseline	4.3%	7.6%	0.3%	12.2%
Baseline*	5.5%	6.1%	0.6%	12.2%
New segmenter*	5.5%	5.3%	0.6%	11.4%
Oracle	5.5%	4.8%	0.6%	11.0%
	eval06			
	Sub	Del	Ins	Overall
Baseline	8.6%	7.6%	1.6%	17.8%
Baseline*	8.8%	6.7%	1.7%	17.2%
New segmenter*	8.9%	6.1%	1.8%	16.8%
Oracle	8.5%	6.2%	1.7%	16.4%
	uwbc06			
	Sub	Del	Ins	Overall
Baseline	10.9%	10.6%	0.8%	22.3%
Baseline*	10.9%	9.4%	0.9%	21.2%
New segmenter*	11.0%	7.5%	1.1%	19.6%
Oracle	10.7%	7.6%	0.9%	19.3%

4.2. Discussion

To understand the segmentation behavior, we examine the differences from different segmentation algorithms. In Table 8, "Oracle" is the total duration of the test set according to the manual transcription; "Baseline" represents the total extracted duration using our previous segmenter, and "New S" similarly by the new segmenter. In addition, the amount of speech extracted by our previous segmenter but discarded by our new segmenter is indicated by "B Only", and that kept by the new segmenter but not by the previous segmenter is indicated in the "N Only" column. It is obvious that the new segmenter is more effective in extracting speech from the corresponding recordings for all 3 testing sets.

In examining those "B Only" segments, we find that they are mostly non-speech, or very short speech but heavily corrupted by background noise. Therefore, there is little benefit in recovering those "B Only" segments.

Table 8. The amount of speech (in seconds) extracted from the acoustic recordings.

	Oracle	Baseline	New S	B Only	N Only
eval04	3587.4	3449.9	3489.2	7.8	47.1
eval06	7927.2	7601.9	7754.5	19.4	172.0
uwbc06	5626.6	5333.3	5459.4	44.1	170.2

We also tried to understand the benefit of the language specific phonetic knowledge in the segmentation. In one setup, we simply merged *I1*, *I2*, *F* and *forgn* with one single foreground model *fg*, rendering the only difference between our two segmenters in the pronunciation of "speech", either with a minimum duration of 540ms or 60ms. This setup degraded the performance by absolutely 0.1% on all three sets. It implies that the duration constraint is more important than the more detailed speech modeling, which makes sense given that the new speech modeling is not much more detailed than the previous one. It is possible that bigger gains could be obtained using a language-dependent phone loop with a phone sequence n-gram for the loop to model "speech".

5. Summary and Future Work

In this paper, we present an improved acoustic segmenter to recover some deletion errors for ASR. The basic algorithm can be applied to any language. More detailed language specific knowledge can be incorporated into the speech detection process (the modeling of speech) if the segmentation is still not satisfactory. However, we find it is more important to allow an auto utterance to include short pauses between short speech segments, rather than to constrain a speech segment to be at least half of a second long.

In order to verify whether more detailed decoding is helpful for segmentation, phone loop decoding for speech detection needs to be investigated. Also, separate recognition can be done first for different segmentations, and then hypotheses from different segmentations may be combined by ROVER (recognizer output voting error reduction) via a confusion network [5].

6. Acknowledgements

This research is based on work supported by the Defense Advanced Research Project Agency (DARPA) under Contract No. HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations in this paper are those of the authors and do not necessarily reflect the views of DARPA.

7. References

[1] Hwang, M.-Y., Lei, X., Zheng, J., Cetin, O., Wang, W., Peng , G., and Stolcke, A., "Advances in Mandarin broadcast speech recognition", Interspeech 2007.

[2] Povey, D., and Woodland, P.C., "Minimum phone error and I-smoothing for improved discriminative training", ICASSP 2002.

[3] Zheng, J., Cetin, O., Hwang, M.-Y., Lei, X., Stolcke, A., and Norgan, N., "Combining discriminative feature, transform, and model training fro large vocabulary speech recognition", ICASSP 2007.

[4] Chen, B., Zhu, Q., and Morgan, N., "Learning long-term temporal features in LVCSR using neural networks", ICSLP 2004.

[5] Mangu, L., Brill, E., and Stolcke, A., "Finding consensus in speech recognition: Word error minimization and other applications of confusion networks", Computer Speech and Language, 373-400, 2000.