Acoustic Model Training with Detecting Transcription Errors in the Training Data

Gakuto KURATA, Nobuyasu ITOH, and Masafumi NISHIMURA

IBM Research - Tokyo, Kanagawa, Japan

{gakuto, iton, nisimura}@jp.ibm.com

Abstract

As the target of Automatic Speech Recognition (ASR) has moved from clean read speech to spontaneous conversational speech, we need to prepare orthographic transcripts of spontaneous conversational speech to train acoustic models (AMs). However, it is expensive and slow to manually transcribe such speech word by word. We propose a framework to train an AM based on easy-to-make rough transcripts in which fillers and small word fragments are not precisely transcribed and some transcription errors are included. By focusing on the phone duration in the result of forced alignment between the rough transcripts and the utterances, we can automatically detect the erroneous parts in the rough transcripts. A preliminary experiment showed that we can detect the erroneous parts with moderately high recall and precision. Through ASR experiments with conversational telephone speech, we confirmed that automatic detection helped improve the performance of the AM trained with both conventional ML criteria and state-of-the-art boosted MMI criteria.

Index Terms: Acoustic model, Rough transcripts, Lightly supervised training

10.21437/Interspeech.2011-183

1. Introduction

As the purpose of Automatic Speech Recognition (ASR) has moved from the dictation of clean read speech to the transcription of spontaneous conversational speech, we need to build acoustic models (AMs) for conversational speech. In order to build an AM, utterances and corresponding transcripts are needed. However, since conversational speech contains unclear utterances, word fragments, laughter, breathing noises, and the like, manually preparing orthographic transcripts (***precise transcripts***) in which every word in the utterances is precisely transcribed is costly [1, 2, 3].

To tackle the situation when precise transcripts are not available, we prepared ***rough transcripts*** instead of precise transcripts. For the rough transcripts, the requirement of the transcription quality is relaxed and therefore the costs of manual transcription are reduced. However, the rough transcripts are not suitable for AM training. For example, transcription errors can have negative effects.

In this paper, we propose a framework to train an AM with excluding erroneous parts from the rough transcripts. The key idea is to focus on the phone duration in the result of forced alignment between the rough transcripts and the corresponding utterances and then to eliminate the parts that include the phones with anomalous durations. We call the transcripts from which the erroneous parts have been excluded the ***purified transcripts***. With the conventional ML criteria and state-of-the-art boosted MMI (bMMI) [4] criteria, we trained the AM using the rough transcripts or the purified transcripts and then compared their performance in ASR experiments.

2. Transcription specification

In this section, we explain the procedure to prepare the transcripts of the audio data. Then we compare the precise and the rough transcripts we use.

2.1. Procedure

After we collect the audio data, we need to manually transcribe it. To reduce the work for the human transcribers and to improve the efficiency of AM training, the audio data is segmented into utterances of up to 30 seconds by using a voice activity detection system [5]. Then the human transcribers transcribe these segmented utterances.

2.2. Comparison between precise and rough transcripts

Considering that the transcripts are used to train an AM, we want orthographic verbatim transcripts that also include special indicators for laughter, breathing noises, background noises, and the like. First, we prepared such transcripts (precise transcripts). It takes about 20 hours of work for each hour of utterances to prepare these precise transcripts in our environment, which is similar to the workload for the careful transcriptions used in the Fisher corpus [1].

To reduce the workload, we relaxed the transcription quality requirements and reduced the number of rechecks. As a result, though the quality of the transcripts declined, the workload for the manual transcription was cut to one-third. We call these rough transcripts.

To estimate the quality of the rough transcripts, we selected some samples of the rough transcripts and the corresponding utterances. Then we transcribed these utterances under the specifications of the precise transcripts. By using the precise transcripts as a reference, we found out that the Character Error Ratio (CER) of the rough transcripts was 8.1%. In addition, we found that 9.0% of the utterances contained unmarked laughter, breathing noises, and background noises.

The erroneous parts in the rough transcripts where transcription errors occur or where noises are left unannotated can have negative effects in AM training. We are going to propose a method to automatically detect such parts.

3. Proposed method

In this section, we first describe our assumptions. Then we explain how to detect the erroneous parts in the rough transcripts.

3.1. Assumptions

The situation we are assuming is that a certain, but not sufficient amount of audio data and corresponding precise transcripts were available. In order to boost the accuracy of the ASR, we additionally collected new audio data. To reduce the workload, we prepared only rough transcripts for the newly recorded data instead of precise transcripts.

Copyright © 2011 ISCA

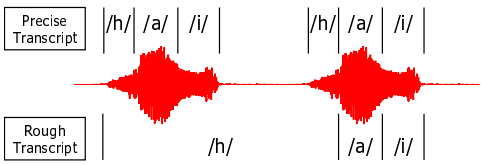

Figure 1: *Example of forced alignment with precise and rough transcript.*

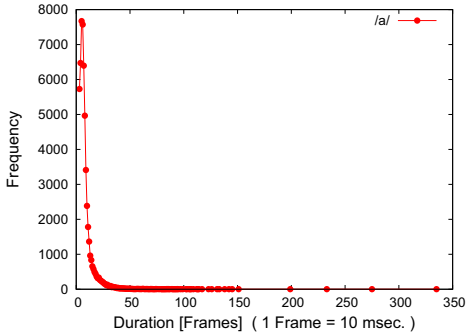

Figure 2: *Duration distribution of Japanese vowel* /a/.

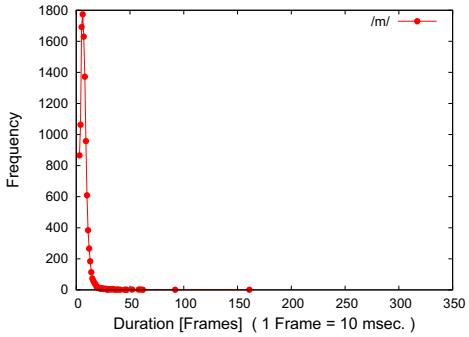

Figure 3: *Duration distribution of Japanese consonant* /m/.

3.2. Detection of erroneous parts in rough transcripts

We propose a framework to detect the erroneous parts in the rough transcripts. The key idea is to focus on the phone duration in the forced alignment.

3.2.1. Intuitive idea

When training an AM, we use forced alignment that aligns the transcripts and the utterances at the phone level. To build a good AM, precise alignment is important. **Figure 1** shows a result of forced alignment for the utterance "Hai, Hai" ("Yes, yes" in Japanese) with the precise transcript "Hai, Hai" and the rough transcript "Hai". Note that the pronunciation of "Hai" is "/h/ /a/ /i/" and the characters between the slashes (/) represent the individual phones. When the rough transcript is used, the durations of some phones, in this case /h/, become unnaturally long to compensate for the error in the transcript. We call this a ***phone duration anomaly***. A phone duration anomaly is caused by a mismatch between the utterances and the rough transcripts. Therefore, by detecting phone duration anomalies, we can find the erroneous parts in the rough transcripts.

In order to automatically detect erroneous parts in the rough transcripts, we first conduct forced alignment on the rough transcripts with the utterances. Because the normal duration varies between phones, we accumulate statistics for the durations of each phone. **Figure 2** and **Figure 3** show the distributions of

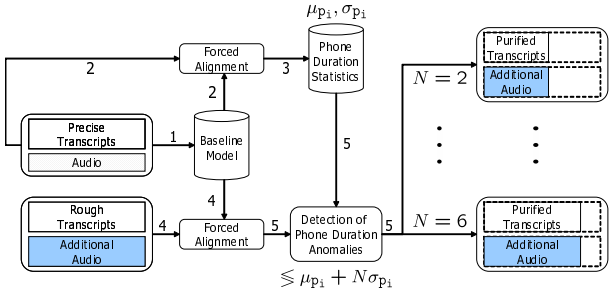

Figure 4: *Detection of phone duration anomaly.*

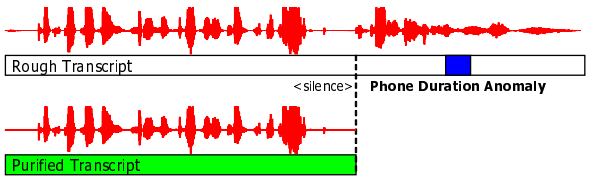

Figure 5: *Elimination after phone duration anomaly.*

the durations of the Japanese vowel /a/ and consonant /m/. The horizontal axis is the duration and the vertical axis is the frequencies of the phone with each duration. The shape of the distributions seems like the combination of the Gaussian-like distribution on the left and the outliers on the right. The distributions of the other phones also have similar shapes. We believe that the Gaussian-like distribution is the natural fluctuation of the duration and the outliers are the phone duration anomalies. Therefore, by detecting these outliers and eliminating the utterances containing the outliers, we can purify the rough transcripts.

3.2.2. Procedure

Figure 4 shows the procedure to detect phone duration anomalies and to purify the rough transcripts. Note that the audio data corresponding to the precise transcripts and the rough transcripts are not the same. The numbers along with the arrows in **Figure 4** correspond to the steps in parentheses in the following text.

To detect the outliers, we need to model the Gaussian-like distribution on the left. As stated in the assumption in Section 3.1, we have the precise transcripts and estimate the baseline AM based on them (**Step 1**). We conduct forced alignment over the precise transcripts with the corresponding utterances using the baseline AM (**Step 2**). By accumulating statistics from this alignment, we can calculate the mean μ_{p_i} and the standard deviation σ_{p_i} of the duration for each phone p_i (**Step 3**). Using the baseline AM, we conduct forced alignment with the rough transcripts and the newly recorded utterances (**Step 4**). Then by using the integer parameter $N = 2, \cdots, 6$, we assume the alignment of the phone p_i whose duration is more than $\mu_{p_i} + N\sigma_{p_i}$ as a phone duration anomaly. After we detect phone duration anomalies, we cut out the utterances and the transcripts that include the phone duration anomalies. One possible option is to discard the whole utterance and the transcript including a phone duration anomaly, but this approach loses too much data. We retain the part from the beginning to the silence just before the phone duration anomaly and discard the rest of the utterance. Because we use the Viterbi algorithm in our forced alignment, the part following the phone duration anomaly tends to be unreliable. **Figure 5** shows an intuitive example. By elim-

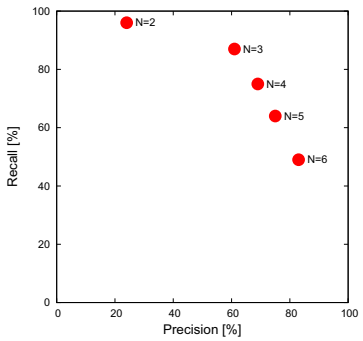

Figure 6: *Precision and recall of erroneous part detection.*

inating the parts after the phone duration anomalies, we get the purified transcripts for each N from the rough transcripts (**Step 5**). As N becomes larger, the size of the parts being discarded becomes smaller and the size of the purified transcripts becomes larger in consequence.

4. Experiment

We conducted a preliminary experiment to confirm whether or not our approach proposed in Section 3.2 can detect phone duration anomalies. Then we compared the performance of the AM using all of the rough transcripts and the AMs using the purified transcripts, which are the subset of the rough transcripts, in ASR experiments.

4.1. Preliminary experiment

Our preliminary experiment estimated the precision and the recall in the detection of phone duration anomalies. First we trained the baseline AM from 50 hours of conversational telephone speech and the corresponding precise transcripts. Then we tried to detect erroneous parts in the rough transcripts for the additional speech data while changing the parameter N from 2 to 6. For the precision, we randomly selected 100 detected anomalies for each N and manually checked whether each of the detected anomalies is actually incorrect because of the incompleteness of the rough transcripts. For the recall, we manually collected 100 transcription errors from the rough transcripts in advance and checked whether these 100 errors were appropriately detected as phone duration anomalies for each N. **Figure 6** shows the precision and the recall for each N. When $N = 2$, though the recall was high, the precision was low, because many alignments were detected as phone duration anomalies. When $N = 6$, the result was reversed, because only a few alignments were detected. Considering the training of the AM, a value which achieves both high recall and high precision is preferred

There was one additional finding when we manually checked the results. We had expected that phone duration anomalies were mainly caused by transcription errors. In addition to this source, we also found some phone duration anomalies that were triggered by laughter, breathing noises, or background noises. It's natural that the conversational telephone speech is recorded in real environments, not in a recording studio. These unexpected noises interfere with AM training and automatically detecting them is beneficial.

4.2. AM training with purified transcripts

Figure 7 shows the flow of the AM training. First we built the baseline model using the precise transcripts and the corresponding utterances. The upper flow in **Figure 7** is a normal proce-

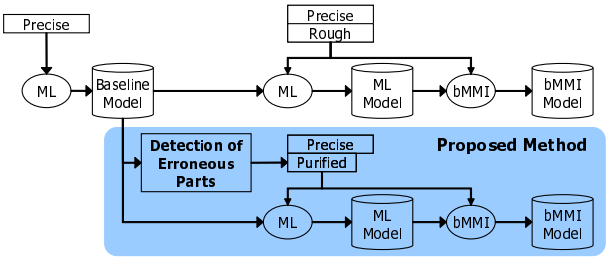

Figure 7: *Training flow.*

Table 1: *Size of precise and rough transcripts.*

Transcripts	Size [Hours]
Precise	50
Rough	12

dure in which we trained the AMs using the precise transcripts and all of the rough transcripts with ML criteria and bMMI criteria afterwards. The lower flow is our proposed method in which we used the precise transcripts and the purified transcripts. In this case, we trained several AMs with ML criteria while changing the parameter N. Then we trained the AM with bMMI criteria using the N that achieved the best performance in ML training. We now describe the experimental setup and discuss the results.

4.2.1. Experimental setup

Table 1 shows the size of utterances and the transcripts used to train the AMs. In the AM, the phones were represented as context-dependent, 3-state, left-to-right HMMs. The HMM states were clustered by using a phonetic decision tree. The number of states and Gaussians were kept at 5k and 200k in all of the experiments.

We conducted an ASR experiment using real-life data from a Japanese call center. We randomly selected 8 calls and used the utterances of the agents as the test data. This test data included 3 male and 5 female agents. The number of utterances was 886 and their total duration was 44.9 minutes. The number of characters in the transcribed text was 15,041.

We estimated the 3-gram language model (LM) with modified Kneser-Ney smoothing [6] from a corpus that was collected for the target call center and prepared a lexicon of 20,860 words with 24,095 pronunciations. We used the same LM and the same lexicon in all of the experiments.

To measure the ASR accuracy, we used the CER. This is because of the ambiguity in word segmentation in Japanese. For example, "東京都知事" (Governor of Tokyo) can be segmented into words in four ways: (1) "東京都知事", (2) "東京都" / "知事", (3) "東京" / "都知事", and (4) "東京" / "都" / "知事". In all of these cases, the same characters are used and the number of the characters is 5. However, the number of words seems to change from 1 to 3 because of the ambiguity and the Word Error Rate (WER) fluctuates accordingly. Therefore, the CER is a suitable criterion in Japanese.

4.2.2. Training while changing size of purified transcripts

We trained 5 AMs with ML criteria while changing the parameter N from 2 to 6 and measured the performance by conducting ASR on the test data. **Table 2** shows the results. The leftmost column shows the parameter N. The middle column shows the ratio of the purified transcripts to the rough transcripts. The rightmost column shows the CER over the test data. The sec-

Table 2: *CER when changing parameter N.*

N	Ratio	CER
$-\infty$ (Used Precise Transcripts Only.)	0	23.9
2	0.30	23.4
3	0.53	23.6
4	**0.69**	**22.4**
5	0.80	22.7
6	0.86	23.2
∞ (Added All of Rough Transcripts.)	1.0	22.9

Table 3: *Result by ML training and bMMI training.*

Size of Rough Transcripts	Training Criteria	CER
-	ML	23.9
All	ML	22.9
Purified ($N = 4$)	ML	**22.4**
All	bMMI	21.5
Purified ($N = 4$)	bMMI	**20.7**

ond row from the top shows the result when only the precise transcripts were used. The bottom row shows the result when all of the rough transcripts were used.

Comparing the CERs, with $N = 4$, the performance was the best, the CER of 22.4%, with improving by 0.5% from the result, the CER of 22.9%, when all of the rough transcripts were used. This result shows that excluding the detected erroneous parts from the rough transcripts improved the performance of the AM even though the size of the training data was decreased[1]. Referring again to **Figure 6** in Section 4.1, both the precision and the recall were moderately high when $N = 4$ and this resulted in good performance in the AM training.

4.2.3. Training with ML and bMMI criteria

With bMMI criteria, we trained one AM using all of the rough transcripts and the other AM using the purified transcripts ($N = 4$) and then compared their performance. **Table 3** shows the results. The columns from the left shows the size of the rough transcripts used in the training, the training criteria, and the CER respectively. The second row from the top shows the results when only the precise transcripts were used. The third and the fourth rows show the results when training with ML criteria using all of the rough transcripts and the purified transcripts. These are duplicated from **Table 2**.

The fifth and sixth rows show the results when training with bMMI criteria. The improvement from all of the rough transcripts to the purified transcripts was 0.8%, from 21.5% to 20.7%. This improvement was larger than the improvement of 0.5% when trained with ML criteria. Discriminative training such as bMMI seems to be more affected by the erroneous transcripts.

5. Related Work

In the area of lightly supervised training when precise transcripts are not available, many researchers have studied techniques in which the results of ASR [7], the minutes of the meetings [8], and the closed captions [9, 10] are leveraged. When using the results of ASR, confidence measure can be exploited to remove erroneous portions [11, 12, 13]. This scheme can be applied when using the rough transcripts by exploiting the likelihood in forced alignment. Leveraging other utterance verification research seems promising [14, 15]. Research on OOV (Out-Of-Vocabulary) detection may also help [16]. We would like to contrast and combine our proposed method with such related work in the future.

[1]This is even more beneficial because it reduces the computational resources for training.

6. Conclusion

Preparing the precise transcripts for spontaneous conversational speech is time-consuming, costly, and labor-intensive. Though the rough transcripts contain some errors, the workload for preparation is much smaller than that for the precise transcripts. We proposed a method to train an AM by purifying the rough transcripts. Our method focuses on the phone duration in forced alignment to remove the parts that contain phone duration anomalies. In the preliminary experiment, we confirmed that we could detect the erroneous parts with moderately high precision and recall. We compared the performance of the AM trained with all of the rough transcripts and the AM trained with the purified transcripts. We confirmed that the purified transcripts automatically generated from the rough transcripts were superior to the rough transcripts in both ML and bMMI training. This suggests that we can leverage the rough transcripts more effectively by automatically excluding the erroneous parts.

7. References

[1] C. C. David *et al.*, "The Fisher corpus: a resource for the next generations of speech-to-text," in *Proc. LREC*, 2004, pp. 69–71.

[2] N. Morgan *et al.*, "The meeting project at ICSI," in *Proc. HLT*, 2001, pp. 246–252.

[3] B.-J. P. Hsu *et al.*, "Language model parameter estimation using user transcriptions," in *Proc. ICASSP*, 2009, pp. 4805–4808.

[4] D. Povey *et al.*, "Boosted MMI for model and feature-space discriminative training," in *Proc. ICASSP*, 2008, pp. 4057–4060.

[5] E. Marcheret *et al.*, "The IBM RT06s evaluation system for speech activity detection in CHIL seminars," in *Proc. MLMI*, 2006, pp. 323–335.

[6] S. F. Chen *et al.*, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language*, vol. 13, no. 4, pp. 359–393, 1999.

[7] T. Kawahara *et al.*, "Language model transformation applied to lightly supervised training of acoustic model for congress meetings," in *Proc. ICASSP*, 2009, pp. 3853–3856.

[8] M. Paulik *et al.*, "Lightly supervised acoustic model training on EPPS recordings," in *Proc. INTERSPEECH*, 2009, pp. 224–227.

[9] L. Lamel *et al.*, "Lightly supervised and unsupervised acoustic model training," *Computer Speech & Language*, vol. 16, pp. 115–129, 2002.

[10] L. Chen *et al.*, "Lightly supervised acoustic model training using consensus networks," in *Proc. ICASSP*, vol. 1, 2004, pp. 189–192.

[11] T. Kemp *et al.*, "Unsupervised training of a speech recognizer: Recent experiments," in *Proc. EUROSPEECH*, 1999, pp. 2725–2728.

[12] F. Wessel *et al.*, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 23–31, 2004.

[13] H. Jiang, "Confidence measures for speech recognition: A survey," *Speech Communication*, vol. 45, no. 4, pp. 355–470, 2005.

[14] M. Pitz *et al.*, "Automatic transcription verification of broadcast news and similar speech corpora," in *Proc. DARPA Broadcast News Workshop*, 1999, pp. 157–159.

[15] M. Gubian *et al.*, "Novelty detection as a tool for automatic detection of orthographic transcription errors," in *Proc. SPECOM*, 2009, pp. 509–514.

[16] A. Rastrow *et al.*, "A new method for OOV detection using hybrid word/fragment system," in *Proc. ICASSP*, 2009, pp. 3953–3956.