Online Speech Detection and Dual-Gender Speech Recognition for Captioning Broadcast News

Toru Imai, Shoei Sato, Akio Kobayashi, Kazuo Onoe, and Shinichi Homma

NHK (Japan Broadcasting Corporation)
Science and Technical Research Laboratories, Tokyo, Japan
{imai.t-mq, satou.s-gu, kobayashi.a-fs, onoe.k-ec, homma.s-fc}@nhk.or.jp

ABSTRACT

This paper describes two new methods, online speech detection and dual-gender speech recognition, for captioning broadcast news. The proposed online speech detection performs dual-gender phoneme recognition and detects a start-point and an end-point based on the ratio between the cumulative phoneme likelihood and the cumulative non-speech likelihood with a very small delay from the audio input. As soon as the start-point is detected, the subsequent continuous speech recognizer with paralleled gender-dependent acoustic models starts a search using gender change information from the preceding phoneme recognizer to reduce computational cost. Speech recognition experiments on conversational commentaries and field reporting from Japanese broadcast news showed that the proposed speech detection method was effective in reducing false segmentations and also recognition errors in comparison with a conventional method using adaptive energy thresholds. The proposed dual-gender speech recognition with the new speech detection significantly reduced the word error rate by 11.2% relative to a conventional gender-independent system, while keeping the computational cost in real-time.

Index Terms: speech recognition, speech detection, low latency, broadcast captioning

1. INTRODUCTION

We are developing a speech recognition system for producing real-time closed captions for various kinds of live broadcasts. It has already been used in some news broadcasts for an anchorperson's read speech [1-2] and sports programs using a re-speak method [3]. One of the problems that should be solved to expand captioned TV programs by the online speech recognition system is a speech detection error due to a lower signal-to-noise ratio (SNR) caused by background noise and diverse speaking styles. So far, many kinds of speech detection methods have been proposed, and they can be classified into different types. A widely used conventional method is based on a short time log-energy level with two adaptive thresholds for speech and non-speech input [4]. It is simple and works well under many conditions, but it sometimes makes a deletion error at the beginning of an utterance that starts with an unstressed syllable, even without any background noise. Phoneme recognition [5] and Viterbi segmentation with speech and non-speech Gaussian mixture models (GMMs) [6] are two other ways to detect speech segments using more reliable frequency-domain characteristics, but they are designed for offline systems with inevitably long detection delays from the audio input. A top-down method based on pause detection [7] can be used in an online system, but it also requires a detection delay until a pause is detected. Our application of the online captioning needs to quickly output recognition results in order to display the captions with the smallest delay possible from a TV program. A local likelihood ratio between speech and non-speech [8] is another measure for classification. However, this is done frame by frame and the local decision produces rapidly changing results. Therefore, a final decision should be guided by several heuristics for smoothing, but its optimal setting is difficult and dependent on a task. Our purpose here is not to produce a precise frame-level classification but to detect a speech segment of moderate length without losing any speech data to reduce speech recognition errors.

Another problem in the online speech recognition system is the difficulty in recognizing spontaneous conversations between male and female speakers with diverse speaking styles. Gender-dependent acoustic models give a lower word error rate than global gender-independent acoustic models, but simply paralleled acoustic models of both genders require a higher computational cost and can not cope with a gender-mixed speech segment consisting of consecutive utterances by male and female speakers.

This paper proposes a new online speech detection method based on a ratio between the cumulative phoneme likelihood and the cumulative non-speech likelihood, and a new dual-gender continuous speech recognition method allowing search transitions between male and female acoustic models in a speech segment. The advantages with this speech detection method are a recognition-based measure without local decision, very early decision making suitable for an online system, and a by-product of gender change information useful in subsequent dual-gender continuous speech recognition. The proposed speech recognition system has the advantages of a flexible gender control for a lower word error rate and a lower computational cost.

Section 2 describes the proposed online speech detection method and section 3 presents the proposed dual-gender speech recognition method. Experimental results on broadcast news are given in section 4.

2. ONLINE SPEECH DETECTION

The proposed online speech detection method is based on a dual-gender phoneme recognizer, where male and female acoustic models run in parallel with a common beam threshold and accept transitions between both genders in a speech segment. Utilizing the cumulative phoneme likelihood compared with the cumulative non-speech likelihood, the dual-gender phoneme recognizer detects the start-point and the end-

point in the speech segment with only a small delay from the audio input.

The phoneme recognition is performed on a network as illustrated in Figure 1, where $g \in \{0,1\}$ denotes the gender (0 for male and 1 for female) of the speaker, sil_g is an acoustic model of a hidden Markov model (HMM) for non-speech, including silence, noise, and music, and $ph_{g,i}$ is the i-th context-independent phoneme (monophone) HMM for gender g. The male and the female acoustic models are set in parallel as a searched network. The acoustic models of each gender are trained with a lot of the corresponding gender's speech data under multiple acoustic conditions. The acoustic features are the same as the ones used in the subsequent continuous recognizer: the cepstrum of the Mel frequency cepstral coefficients (MFCC), the log-energy, and their first- and second-order regression coefficients. Suppose that the speech detection starts at frame τ, and both of the non-speech models sil_0 and sil_1 at the beginnings are activated to calculate the acoustic likelihood for the audio input. The decoder performs the Viterbi search while allowing a transition from sil_g to $ph_{g,i}$, and then, a transition from $ph_{g,i}$ to another $ph_{g,j}$ is repeated according to a phoneme bigram language model until it reaches the final non-speech acoustic models. Notable transitions in the proposed method are between both genders from $ph_{g,i}$ to $ph_{g',j}$ ($g \neq g'$) with a transition penalty score. It allows for consecutive speech utterances by male and female speakers without any pauses in between. The non-speech models of both genders are trained separately in this paper, but of course they can be identical.

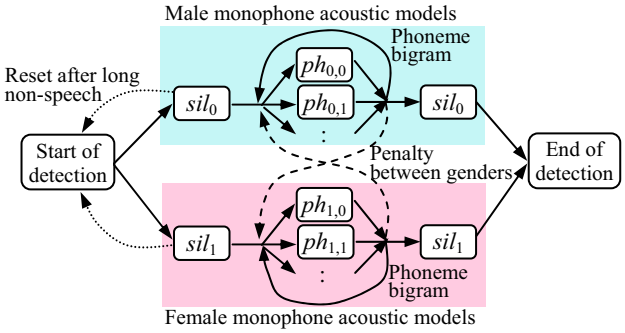

Figure 1: Proposed online speech detection.

2.1 Start-point detection

In order to quickly detect the start-point of speech, the method calculates the cumulative acoustic log-likelihood of the most likely phoneme sequence,

$$L_1 = \max_{h,g,i} L\left(x_\tau^t \middle| h, ph_{g,i}\right), \quad (1)$$

where x_τ^t denotes the audio input vectors of the acoustic features from the initial frame τ to the current frame t, and h is the history of the phoneme sequence preceding $ph_{g,i}$. Another needed value is the cumulative acoustic log-likelihood of the same gender's non-speech model at the beginning,

$$L_2 = L\left(x_\tau^t \middle| sil_g\right). \quad (2)$$

If their difference exceeds threshold θ_{start} at frame t,

$$L_1 - L_2 > \theta_{start}, \quad (3)$$

the method determines that the start-point of speech is T_{start} frames back from the beginning frame of the first phoneme in the most likely phoneme sequence in Equation (1). Then, the subsequent continuous speech recognizer starts decoding.

In order to absorb a very long non-speech segment before detecting a start-point, the initial frame τ is updated with the current frame t, if the condition in Equation (3) is not satisfied for the frame length of T_{idle}.

2.2 End-point detection

In order to detect an end-point of speech, two values are calculated: the best cumulative phoneme log-likelihood from the beginning non-speech model at the initial frame τ to the ending non-speech model at the current frame t,

$$L_3 = \max_{h,g} L\left(x_\tau^t \middle| h, sil_g\right), \quad (4)$$

and the best cumulative phoneme log-likelihood ending with any phoneme model of the same gender,

$$L_4 = \max_{h',i} L\left(x_\tau^t \middle| h', ph_{g,i}\right), \quad (5)$$

where h and h' are the histories of the phoneme sequences. If their difference exceeds threshold θ_{end} consecutively for the frame length of T_{end1} at frame t,

$$L_3 - L_4 > \theta_{end}, \quad (6)$$

and then the method determines that the end-point of speech is $t - T_{end2}$ ($T_{end2} < T_{end1}$). Then, the next speech detection is repeated again in the same way.

3. DUAL-GENDER SPEECH RECOGNITION

3.1 Parallel decoding

As soon as the start-point of an utterance is detected, the subsequent continuous speech recognizer starts decoding using the acoustic models for the male and the female in parallel in a single network (Fig. 2). The acoustic models of the gender-dependent and context-dependent phoneme (triphone) HMMs compose a phonetic lexicon tree for each gender with a non-speech model sil'_g at both ends. Since the male and female acoustic models search for the best word sequence in parallel, but with the common beam threshold, the word hypotheses for only the matched gender naturally remain alive and the other ones stop the search according to the score difference.

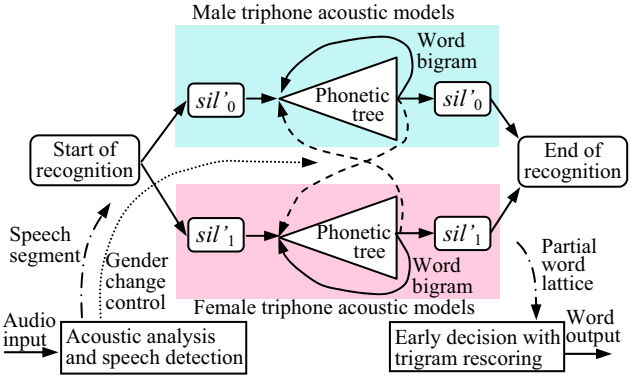

Figure 2: Proposed dual-gender speech recognition.

3.2 Controlled gender change

When the detected speech segment consists of a single gender

utterance, the simple dual-gender decoding has no problems. However, conversations between male and female speakers sometimes give a gender-mixed utterance due to a difficulty in chopping it up according to a gender change without any pauses in between. In order to restart the search for the deactivated gender's acoustic models besides the other remaining gender, the decoder utilizes the gender change information given by the simultaneously running speech detection. The speech detection based on the dual-gender phoneme recognition traces back on the searched space to get the most likely phoneme sequence at every audio input frame and tells their corresponding gender attributes to the continuous speech recognizer. As it takes time to obtain reliable gender attributes in speech detection, the continuous speech recognizer works with a delay of T_{delay} frames from the current frame t. The control method of the dual-gender search is expected to be effective at accepting a gender-mixed utterance and reducing the computational cost with a very small delay from the speech input.

The phonetic lexicon tree allows a loop transition with a word bigram language model and composes a word lattice along with the search as the first pass. In order to make a quick decision, the partial word lattice is rescored with a trigram language model as the second pass, without waiting for the end-point detection. The detailed decoding procedure is described in the next section.

4. EXPERIMENTS

4.1 Evaluation data

Experiments were performed to examine how effective the online speech detection and dual-gender speech recognition methods were. The evaluation speech data was 35 minutes of NHK's Japanese TV news consisting of conversational commentaries and field reporting spoken by 11 speakers, including announcers and reporters (9 males and 2 females). There were 409 utterances if segmented manually and 5,651 words in total. The SNR distributed from 0 to 56 dB and the average SNR was 26.7 dB.

4.2 Language model

The n-gram language models used in the system were phoneme bigrams for the speech detection, word bigrams for the first pass in the continuous speech recognition, and word trigrams for the second pass. The training texts were NHK's Japanese news manuscripts consisting of 127 M words extending back over 10 years. As more recent news had a higher possibility of frequently appearing, the n-gram language models were trained with a higher weighting factor to the latest news in an n-gram count level [1]. The n-gram language models were smoothed over by Good-Turing discounting, where cutoffs were set to two and three for bigrams and trigrams, respectively. There were 61 K vocabulary words and the trigram language model showed a perplexity of 26.3, with an out-of-vocabulary rate of 0.4% against the evaluation data.

4.3 Acoustic model

The gender-dependent acoustic models of the context-independent HMMs for the speech detection and the context-dependent HMMs for the continuous speech recognition were trained separately. There were 42 Japanese phonemes. The acoustic features of the audio input were 39 parameters (12 MFCCs with the log-energy and their first- and second-order regression coefficients) with RASTA processing [9] every 10 ms after digitization at 16 kHz and 16 bits with a Hamming window of 25 ms in width. The HMMs were trained with the maximum likelihood method from 340 hours of NHK's multi-conditioned news data for the male and 250 hours for the female. The context-independent models were 32-mixtured monophone HMMs. The context-dependent models were 16-mixtured and 4K-state-clustered triphone HMMs.

4.4 Decoder

In order to output recognition results for closed captions as quickly as possible, the decoder for the continuous speech recognition made a quick decision without waiting for the end of a speech segment [10]. The low latency decoder was based on 2-pass searches. In the first pass, the word-dependent N-best algorithm with a modified Viterbi beam search was carried out on the parallel and the static phonetic trees of the male and female acoustic models with a common beam threshold to compose a word lattice. During the first pass, the decoder periodically executed the second pass, which rescored the partial N-best word sequences given from the partial word lattice up to that time. If a rescored best word sequence had words in common with the previous one, that part was regarded as likely to be correct and was determined to be a part of the final result. This method is not theoretically optimal but makes a quick response with a negligible increase in word errors [10].

4.5 Results

The proposed speech detection method was applied to the evaluation data in comparison with the conventional energy-based method using two adaptive thresholds for speech and non-speech [4]. The parameters for the proposed speech detection and the dual-gender speech recognition were sub-optimally determined by using a different development test set: thresholds of $\theta_{start} = 20$ and $\theta_{end} = 5$, a gender change penalty of 50 in log-likelihood, and frames of $T_{start} = 20$, $T_{idle} = 100$, $T_{end1} = 33$, $T_{end2} = 18$, and $T_{delay} = 35$. As shown in Table 1, although the conventional energy-based method yielded relatively longer segments and wrongly dropped speech data at some start-points for the evaluation data, the proposed method based on the cumulative phoneme likelihood obtained moderate speech lengths and no missing start-points. The detection delays of the start- and end-points by the proposed method were 112 ms and 425 ms on average, respectively, which were almost the same as the energy-based method and small enough for the online captioning application.

While a continuous speech recognition experiment with the manual segmentation and the gender-independent acoustic models showed a word error rate (WER) of 11.6% for the evaluation data, a baseline system with the conventional energy-based speech detection gave the WER of 12.5% as

Table 1: Speech detection results

Speech detection	No. of segments	Avg. length	Standard deviation	Max. length	Missing points
Manual	409	4.8 s	2.9 s	21.7 s	0
Energy-based	336	5.4 s	6.4 s	84.8 s	4
Cumulative phoneme likelihood	343	5.4 s	4.2 s	26.0 s	0

Table 2: Speech recognition results (proposed methods identified with *)

Speech detection	Continuous speech recognition		Word error rate	Real time factor
Energy-based	Gender-independent acoustic models		12.5%	0.81
Cumulative phoneme likelihood*			12.2%	0.81
	Dual-gender parallel acoustic models*	No transition allowed	11.9%	0.93
		Transition allowed any time	11.3%	1.28
		Controlled transition*	11.1%	0.93

shown in Table 2. When the speech detection method was changed to the proposed method, the WER was reduced to 12.2% because of its better speech segmentation. In addition, when the gender-dependent acoustic models for the male and female were used in parallel, without allowing transitions between both genders in a speech segment, the WER was further reduced to 11.9%. If the transitions between both genders were allowed all the time, the WER was again reduced to 11.3%, but the real time factor became larger over the real-time. The real time factor of an average ratio between the recognition time and the speech segment length was computed on a Xeon 3.60-GHz machine. Finally, the gender transitions in a speech segment were controlled only at frames when a gender attribute changed in the proposed speech detection. Then, the real time factor was reduced to the same level as in the no transition case, and the WER was also reduced to 11.1%. The overall WER reduction from 12.5% for the conventional baseline system to 11.1% for the proposed system was a significant 11.2% relative improvement at a significance level of 0.05. In the final result, the rate of correctly identified genders was 99.7% of the words.

The speech recognition result was broken down according to some different SNR levels of the evaluation data as illustrated in Figure 3. It shows that a lower SNR of evaluation data yielded a higher relative WER reduction by the proposed methods of speech detection and the continuous speech recognition in comparison with the conventional energy-based and gender-independent methods.

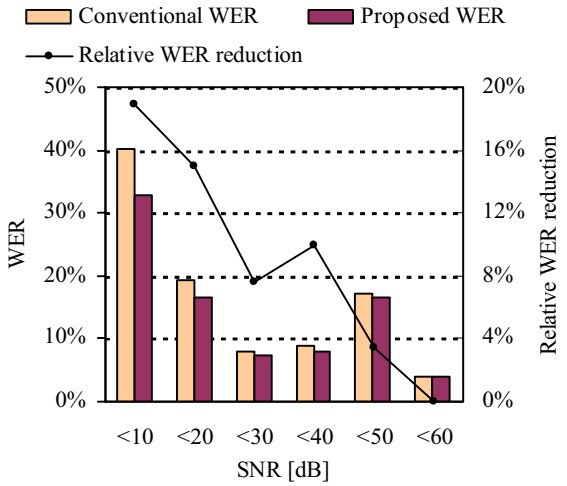

Figure 3: Speech recognition results with different SNR levels.

5. CONCLUSIONS

This paper describes the new methods of online speech detection and dual-gender speech recognition and showed their effectiveness in speech recognition experiments on broadcast news. The proposed online speech detection performs dual-gender phoneme recognition and detects a start-point and an end-point using cumulative phoneme likelihood. The subsequent continuous speech recognizer, with paralleled gender-dependent acoustic models, works with the gender change information from the preceding phoneme recognizer. The speech recognition experiments showed that the proposed speech detection produced moderate speech lengths without any missing start-points and reduced the WER. With the new speech detection, the proposed dual-gender continuous speech recognition significantly reduced the WER by 11.2% relative to the conventional energy-based and gender-independent system, while keeping the computational cost in real-time. In particular, the proposed method showed higher WER reduction against speech segments with a lower SNR.

6. REFERENCES

[1] A. Ando, T. Imai, A. Kobayashi, H. Isono, and K. Nakabayashi, "Real-Time Transcription System for Simultaneous Subtitling of Japanese Broadcast News Programs", *IEEE Trans. Broadcasting*, 46(3), pp.189-196, 2000.

[2] T. Imai, A. Kobayashi, S. Sato, S. Homma, K. Onoe, and T. Kobayakawa, "Speech Recognition for Subtitling Japanese Live Broadcasts, *The 18th International Congress on Acoustics (ICA)*, pp.I-165-168, 2004.

[3] T. Imai, A. Matsui, S. Homma, T. Kobayakawa, K. Onoe, S. Sato, and A. Ando, "Speech Recognition with a Re-Speak Method for Subtitling Live Broadcasts", *ICSLP*, pp.1757-1760, 2002.

[4] S. V. Gerven and F. Xie, "A Comparative Study of Speech Detection Methods", *Eurospeech*, pp.1095-1098, 1997.

[5] F. Kubala, H. Jin, S. Matsoukas, L. Nguyen, R. Schwartz, and J. Makhoul, "The 1996 BBN BYBLOS HUB-4 Transcription System", *DARPA Speech Recognition Workshop*, pp.90-93, 1997.

[6] J.-L. Gauvain, L. Lamel, and G. Adda, "The LIMSI Broadcast News Transcription System", *Speech Communication*, 37, pp.89-108, 2002.

[7] K. Takeda, S. Kuroiwa, M. Naito, and S. Yamamoto, "Top-Down Speech Detection and N-Best Meaning Search in a Voice Activated Telephone Extension System", *Eurospeech*, pp.1075-1078, 1995.

[8] R. Singh, M. L. Seltzer, B. Raj, and R. M. Stern, "Speech in Noisy Environments: Robust Automatic Segmentation, Feature Extraction, and Hypothesis Combination", *ICASSP*, pp.I-273-276, 2001.

[9] H. Hermansky and N. Morgan, "RASTA Processing of Speech", *IEEE Trans. Speech and Audio Processing*, 2(4), pp.578-589, 1994.

[10] T. Imai, A. Kobayashi, S. Sato, H. Tanaka, and A. Ando, "Progressive 2-Pass Decoder for Real-Time Broadcast News Captioning", *ICASSP*, pp.III-1559-1562, 2000.