SPEECH RECOGNITION ROBUST AGAINST SPEECH OVERLAPPING IN MONAURAL RECORDINGS OF TELEPHONE CONVERSATIONS

Masayuki Suzuki, Gakuto Kurata, Tohru Nagano, Ryuki Tachibana

Watson multimodal, IBM

{szuk,gakuto,tohru3,ryuki}@jp.ibm.com

ABSTRACT

Monaural (single-channel) recording is sometimes used for telephone conversations in call centers. Generally speaking, the accuracy of automatic speech recognition of a monaural recording is worse than that of the multi-channel recording of the same conversation where each speaker's voice is separately recorded. The major reason is that the recognition system fails not only at the overlapping segments where the voices of the multiple speakers overlap, but also at the neighboring segments surrounding the overlapping segments. In this paper, we tackle this problem by using a combination of garbage modeling and noise-robust monaural acoustic modeling. Our proposed method trains the models by making use of multi-channel recordings and transcripts, which are relatively easy to prepare than monaural recordings and transcripts. We present experimental results where the proposed methods reduced the error rates by approximately 3% relative to the baseline methods for both of GMM-HMM and CNN-HMM cases. Because the proposed method is quite simple, the proposed method is easy to deploy to wide range of ASR systems for monaural speech transcription.

Index Terms— Overlap, Monaural speech, Garbage model, Noise robust, Telephone conversation

1. INTRODUCTION

Recent study in the field of automatic speech recognition (ASR) has greatly improved the recognition accuracy for transcribing naturally spoken utterances. However, because voices of the multiple speakers participating in the conversation often overlap and because there are application scenarios where only a single microphone can be used, it is often the case that overlapping segments degrade the recognition accuracy and spoil the value of the application [1, 2]. A common application area where monaural recordings are sometimes used is recording of telephone conversation in call centers to reduce the size of the stored data. As another example, recordings of conversations in meetings or at sales counters made by using a single microphone inevitably include overlapping segments.

The ASR accuracies of these monaural recordings with overlapping segments are generally worse than with multi-channel recordings where each speaker's voice is recorded separately. The overlapping segments cause burst errors not only in the overlapping segments but also in their neighboring segments because the decoding process of ASR propagates errors to surrounding segments.

In this paper, we address this problem by combining garbage modeling [3, 4, 5, 6] and noise-robust acoustic modeling [7, 8]. Overlapping occurs when two or more speakers speak simultaneously. If the power of the voices is about the same, it is very difficult to recognize any one of the voices. In this case, our garbage modeling handles the overlapping segments. This may result in some deletion errors in the overlapping segments, but successful removal of the burst errors of insertions and substitutions can lead to overall error rate reduction. Contrarily, if the power of the voice of one of the speakers is dominant, the dominant voice can be recognized, ignoring the other voices as overlapping background noise. To build a noise-robust acoustic model, we use training data including overlapping speech noise in which one of the voices has a dominant power.

Training of a monaural acoustic model usually requires monaural recordings and transcripts with labels of overlapping speech segments. However, it is quite costly to prepare transcripts that contain overlapping segments, which make labeling work even more difficult. To avoid this problem, we generate simulated monaural recordings and transcripts from multi-channel recordings and transcripts, which are relatively easy to obtain.

Our contributions in this study are three-fold. First, we tackled accuracy degradation problems caused by overlapping speech segments in monaural recordings of telephone conversations. As far as we know, this is the first work to tackle this problem. Second, we combined garbage modeling and noise-robust acoustic modeling to handle two types of speech overlapping segments: those where the power is about the same and those where it is not. Third, we proposed a method for making monaural recordings and transcripts from multi-channel recordings and transcripts. Thanks to this, we do not need to manually prepare transcripts of monaural recordings.

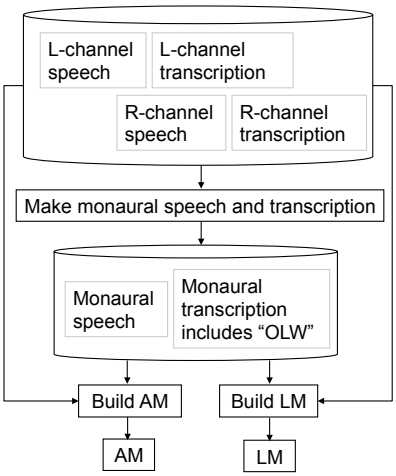

Fig. 1. Flow chart of the model training process of the proposed method

2. PROPOSED METHOD

2.1. Overall flow of model training

Fig. 1 shows an overall flow chart of the model training process of the proposed method. Although we can handle three or more speakers, we consider only two speakers here, which is usually true for telephone conversations.

By simply adding L- and R-channel recordings, we can make simulated monaural recordings. With the single channel recording and transcript of each of the L and R channels, we can do forced alignment to get transcripts with time stamps. Then, we generate transcripts of the simulated monaural recordings by concatenating the transcripts of the two channels along with the time axis and by replacing the words in the overlapping segments with labels of speech overlapping. Finally, we build an acoustic model (AM) and language model (LM) by using the monaural data.

Our dictionary includes a special "overlap-word" (OLW). Our phoneme set includes a special "overlap-phone" (OLP) and an entry in the pronunciation dictionary for OLW is set to OLP. The following sections describe the details of the building process of the proposed AM and LM including OLP and OLW, respectively.

2.2. AM

"Transcripts for AM" in Fig. 2 shows an example of generated transcripts for AM training. An overlapping segment of "right"/"uh-huh" is replaced with "right", which has a dominant power, and an overlapping segments of "wait"/"order" is replaced with an OLW, where the powers of two voices are about the same. Concretely speaking, we calculated the SNR for each overlapping segment by regarding the speech having the larger power as the signal, and if and only if the SNR is higher than a threshold, we regard the segment as dominant speech with background noise.

An AM of an OLP trained by using OLW segments can handles overlapping speech, where the powers of the voices of the speakers are about the same. However, AMs of the other phonemes trained by using this data are robust against overlapping speech background noise in which one of the voices has a dominant power because the training data include the same type of noise. The garbage modeling and the noise-robust AM are complementary, and a combined model can handle the two types of speech overlapping.

2.3. LM

"Transcripts for LM" in Fig. 2 shows an example of generated transcripts for LM training. Both overlapping segments of "right"/"uh-huh" and "wait"/"order" are replaced with OLW in this example. By replacing all overlapping segments, we can gather as many word contexts for overlapping segments as possible.

Although word context information for OLW is useful for garbage modeling, it is effective to assign a smaller probability to the OLW in the LM to reduce false alarm errors because false rejection of OLW does not affect ASR results but false alarm of OLW increases deletion errors. To adjust the probability of the OLW without loss of context information, we interpolate the LM and a normal LM that does not include OLW. The normal LM is trained by using transcripts of multi-channel recordings. The interpolation weight can be used as a tuning parameter to balance false alarm and false reject errors.

2.4. Decoding

We decode the test data by using the proposed noise-robust AM and LM with garbage model. We discard any OLW decoded by the system. After identifying the overlapping segments by using the proposed method, instead of removing the OLW, it would also be possible to introduce more advanced methods to recognize the segments, but this is future work and we only removed overlapping segments in this work.

3. EXPERIMENTS

3.1. Experimental conditions

We conducted Japanese monaural telephone conversation ASR by using in-house data. These conversations were between agents and customers. All of the speech was recorded at 8 kHz with 16 bit sampling. We used a voice activity detection (VAD) system to extract speech segments of the data in advance. The test data was 1 hour of manually transcribed monaural recordings. The training data was 150 hours of two-channel data and transcripts. We call this "data A". Some parts of data A are two-channel data, but only the

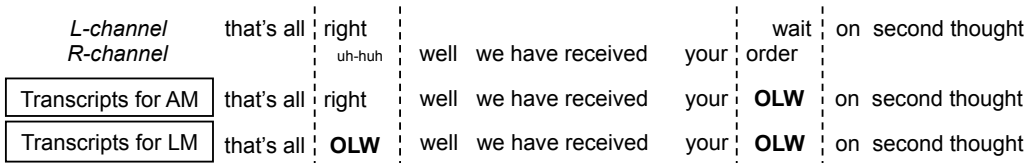

Fig. 2. Examples of generation process of monaural scripts with OLW from two-channel data

agent-side data exist. We were able to artificially generate 30 hours of simulated monaural recordings with artificial transcripts with OLW from data A where both the agent-side and caller-side data exist. We call this "data B". 12% of the data B are overlap segments according to the VAD results on two-channel data.

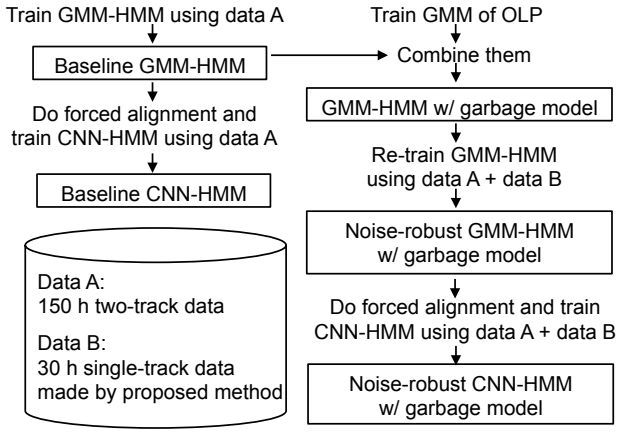

Fig. 3. Training recipe of AMs

We used the AM recipe of Fig. 3. Data A was used for baseline AMs. Both data A and data B were used for proposed AMs. The threshold to regard one speaker as dominant in data B was set to 10 dB (See Section 2.2). The GMM-HMM with garbage model was made by combining the baseline GMM-HMM and the GMM of OLP. This GMM for OLP was trained with speech segments aligned to OLW in data B. We re-trained the GMM-HMM by using data A and data B to create the noise-robust GMM-HMM with garbage model. Then, using the GMM-HMM, we did forced alignment for data A and data B. We used this data to train the noise-robust convolutional neural network (CNN)-HMM [9, 10, 11] with garbage model.

For the GMM-HMM, we used 4,500 HMM states and 150,000 Gaussians. For the OLP, 1 HMM state and 100 Gaussians were added. The HMM states were clustered by using a phonetic decision tree of quinphones while the OLP was regarded as monophone. The features were derived from 13-dimension perceptual linear prediction (PLP) features. The acoustic context was taken into account by splicing 9 adjacent frames of mean- and variance-normalized PLP features and then projecting into a 40-dimension feature space by using linear discriminant analysis (LDA), followed by global semi-tied covariance (STC). Feature space and model space boosted maximum mutual information (bMMI) training was used to train the GMM-HMM. At decoding time, maximum likelihood linear regression (MLLR) model adaptation was used.

For the CNN-HMM, we used two convolutional layers and six hidden layers. The input features were in a 24-dimension log MEL filter bank and its delta, and delta-delta, and their 11 adjacent frames ($3 \times 24 \times 11$). The means and variances were globally normalized. The weight share window size for the first convolutional layer was 9×9 (time $\times$ frequency). We used only frequency axis pooling with a window size of 3. The weight share window size for the second convolutional layers was 3×4 (time $\times$ frequency). No pooling was used in the second layer. The numbers of nodes in the hidden layers were 1024, 1024, 1024, 1024, 1024, and 512. All of the activation functions were sigmoid functions. In the final layer, a softmax activation function was used to calculate the posterior probabilities of the HMM states. The prior probabilities of the HMM states were estimated by using the training data in an ML manner. The pre-training was done with layer-wise cross entropy training, followed by cross entropy based fine-tuning.

The LM was a word 3-gram estimated by using modified Kneser-Ney smoothing [12]. The baseline LM was trained by using transcripts of data A. We also trained an LM by using data B. The proposed LM was made by interpolating the baseline LM and the LM from data B. The interpolation weight was set to 8:2 (See Section 2.3).

The decoder was our WFST decoder [13]. The LM weight was tuned manually to achieve the lowest character error rate (CER) for each type of AM.

3.2. Results

Table 1 shows the results of the experiments. Note that the CER is measured over all data, both overlapped and non-overlapped speech. We can see the sole effect of garbage modeling and noise-robust acoustic modeling in the second line and the third line, respectively. Both techniques had gain from the baseline GMM-HMM. The fourth line shows that the combination of the two techniques had further gain. The proposed model increased the relatively small number of deletion errors, but greatly decreased the larger number of substitution and insertion errors caused by the overlapping segments, so that overall CERs were reduced. The same trend

Table 1. CER, number of substitution, insertion, and deletion errors of various AM types

AM type	Garbage model	Noise-robust AM	CER	#sub.	#ins.	#del.
GMM-HMM			31.8%	2591	573	1517
GMM-HMM	✓		31.2%	2501	551	1531
GMM-HMM		✓	31.0%	2503	545	1509
GMM-HMM	✓	✓	30.9%	2406	526	1612
CNN-HMM			29.5%	2490	654	1189
CNN-HMM	✓	✓	28.5%	2176	550	1451

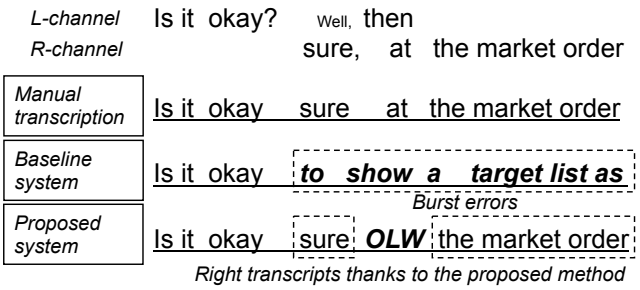

Fig. 4. An artificial example of manual and system transcripts for a monaural recording

can be found for CNN-HMM. The proposed CNN-HMM reduced the CER approximately 1 point (3% relative) relative to the baseline CNN-HMM.

Fig. 4 shows typical examples of transcripts of the manual and proposed systems. We can see that the proposed garbage model of overlap absorbed the overlapping segments that caused burst errors of substitutions or insertions.

4. RELATION TO PRIOR WORKS

4.1. Garbage modeling

Widely used garbage modeling in ASR can be roughly categorized into two groups according to the types of garbage modeling. The first type is called a "tee model", which introduces garbage words that do not affect the context but are inserted at an arbitrary word boundary [14]. This type of garbage modeling is often used for the modeling of short pauses.

The other type of garbage modeling considers the garbage words as regular words in decoding time, and removes the garbage words after the decoding. If the garbage words tend to appear in certain linguistic contexts, then it works well. This type of garbage modeling is often used for the modeling of filler words [15, 16]. In addition, [3] found this type of garbage modeling is effective for speaker noises such as the noise of breathing, which has linguistic context dependency.

[2] analyzed contexts around overlapping segments and found linguistic tendencies. Considering their result, using the latter type of garbage modeling for speech overlapping is reasonable.

4.2. Noise-robust acoustic modeling

Multi-condition training is a well-known and effective approach to training a noise-robust AM [7, 8, 17, 18, 19]. Multi-condition training is a technique for using speech data of various types and signal-to-noise ratio (SNR) levels of various noise as training data. We used multi-condition training for our noise-robust AM, where the training data includes clean speech and speech with overlapping background noise where the SNR is larger than a threshold.

4.3. Overlap modeling in speaker diarization tasks

Aside from ASR studies, there are many research projects on tackling the overlap problem in meeting diarization tasks [20, 21]. [21] proposed a GMM-HMM-based segmenter of non-speech, speech, and overlapping segments for post-processing of a baseline meeting diarization system. They pointed out that false rejections of overlapping segments did not affect the baseline system, but false alarms increased errors of diarization. Because there is a trade-off relationship between false alarms and false rejection, they tuned their overlap detector for low false alarms (and possibly high false rejection) operating points. Following this, we tuned our proposed model for low false alarms by using the interpolation weight of the LM.

5. CONCLUDING REMARKS

We proposed a combination of garbage modeling and noise-robust acoustic modeling robust against speech overlapping in monaural recordings of telephone conversations. If the powers of the voices of the speakers are about the same, the garbage model absorbs the overlapping segments, and if the power of the voice of one of the speakers is dominant, the noise-robust AM can recognize the dominant voice. Our proposed method does not need any monaural recordings and manual transcripts as input data because we generate them from multi-channel recordings and transcripts. We confirmed that the proposed method reduced the CER by approximately an absolute value of 1 point, which is approximately 3% relative to the baseline CNN-HMM system.

Our future works includes applying an advanced technique [22, 23, 24, 25] for segments recognized as OLW.

6. REFERENCES

[1] Elizabeth Shriberg, Andreas Stolcke, and Don Baron, "Observations on overlap: findings and implications for automatic processing of multi-party conversation.," in *INTERSPEECH*, 2001, pp. 1359–1362.

[2] Ozgür Cetin and Elizabeth Shriberg, "Speaker overlaps and asr errors in meetings: Effects before, during, and after the overlap," in *Acoustics, Speech and Signal Processing (ICASSP), 2006 IEEE International Conference on*. IEEE, 2006.

[3] G. Sarosi, B. Tarjan, A Balog, T. Mozsolics, P. Mihajlik, and T. Fegyo, "On modeling non-word events in large vocabulary continuous speech recognition," in *Cognitive Infocommunications (CogInfoCom), 2012 IEEE 3rd International Conference on*, Dec 2012, pp. 649–653.

[4] V.R. Raman and G.J. Vysotsky, "Methods and apparatus for generating and using garbage models for speaker dependent speech recognition purposes," Nov. 24 1998, US Patent 5,842,165.

[5] Chafic Mokbel, Laurent Mauuary, Lamia Karray, Denis Jouvet, Jean Monné, Jacques Simonin, and Katarina Bartkova, "Towards improving asr robustness for psn and gsm telephone applications," *Speech Communication*, vol. 23, no. 1, pp. 141–159, 1997.

[6] J Caminero, Daniela de la Torre, Luis Villarrubia, C Martin, and Lis Hernandez, "On-line garbage modeling with discriminant analysis for utterance verification," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*. IEEE, 1996, vol. 4, pp. 2111–2114.

[7] Michael L Seltzer, Dong Yu, and Yongqiang Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7398–7402.

[8] Richard P Lippman, Edward Martin, Douglas B Paul, et al., "Multi-style training for robust isolated-word speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), '87 IEEE International Conference on*. IEEE, 1987, vol. 12, pp. 705–708.

[9] Tara N Sainath, Brian Kingsbury, George Saon, Hagen Soltau, Abdel-rahman Mohamed, George Dahl, and Bhuvana Ramabhadran, "Deep convolutional neural networks for large-scale speech tasks," *Neural Networks*, 2014.

[10] Jui-Ting Huang, Jinyu Li, and Yifan Gong, "An analysis of convolutional neural networks for speech recognition," 2015.

[11] Takuya Yoshioka, Shigeki Karita, and Tomohiro Nakatani, "Far-field speech recognition using cnn-dnn-hmm with convolution in time," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4360–4364.

[12] Stanley F Chen and Joshua Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language*, vol. 13, no. 4, pp. 359–393, 1999.

[13] H. Soltau, G. Saon, and B. Kingsbury, "The IBM Attila speech recognition toolkit," in *Spoken Language Technology Workshop (SLT)*, Dec. 2010, pp. 97–102.

[14] S Young, G Evermann, D Kershaw, G Moore, J Odell, D Ollason, V Valtchev, and P Woodland, "The htk book (revised for htk version 3.4. 1)," *Cambridge University*, 2009.

[15] Yuya Akita and Tatsuya Kawahara, "Efficient estimation of language model statistics of spontaneous speech via statistical transformation model," in *Acoustics, Speech and Signal Processing (ICASSP), 2006 IEEE International Conference on*. IEEE, 2006.

[16] Kengo Ohta, Masatoshi Tsuchiya, and Seiichi Nakagawa, "Evaluating spoken language model based on filler prediction model in speech recognition.," in *INTERSPEECH*, 2008, pp. 1558–1561.

[17] Masayuki Suzuki, Takuya Yoshioka, Shinji Watanabe, Nobuaki Minematsu, and Keikichi Hirose, "Feature enhancement with joint use of consecutive corrupted and noise feature vectors with discriminative region weighting," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 10, pp. 2172–2181, 2013.

[18] Osamu Ichikawa, Steven J Rennie, Takashi Fukuda, and Masafumi Nishimura, "Channel-mapping for speech corpus recycling," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7160–7164.

[19] D. Kolossa and R. Haeb-Umbach, Eds., *Robust Speech Recognition of Uncertain or Missing Data*, Springer, 2011.

[20] Xavier Anguera Miro, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals, "Speaker diarization: A review of recent research," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 2, pp. 356–370, 2012.

[21] Kofi Boakye, Beatriz Trueba-Hornero, Oriol Vinyals, and Gerald Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings," in *Acoustics, Speech and Signal Processing (ICASSP), 2008 IEEE International Conference on*. IEEE, 2008, pp. 4353–4356.

[22] Mikkel N Schmidt and Rasmus Kongsgaard Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *INTERSPEECH*, 2006.

[23] Bin Gao, WL Woo, and SS Dlay, "Single-channel source separation using emd-subband variable regularized sparse features," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 961–976, 2011.

[24] Dennis L Sun and Gautham J Mysore, "Universal speech models for speaker independent single channel source separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 141–145.

[25] Emad M Grais, Mehmet Umut Sen, and Hakan Erdogan, "Deep neural networks for single channel source separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 3734–3738.