INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Use of Global and Acoustic Features Associated with Contextual Factors to Adapt Language Models for Spontaneous Speech Recognition

Shohei Toyama, Daisuke Saito, Nobuaki Minematsu
Graduate School of Engineering, The University of Tokyo
toyama@gavo.t.u-tokyo.ac.jp, dsk_saito@gavo.t.u-tokyo.ac.jp, mine@gavo.t.u-tokyo.ac.jp

Abstract

In this study, we propose a new method of adapting language models for speech recognition using para-linguistic and extra-linguistic features in speech. When we talk with others, we often change the way of lexical choice and speaking style according to various contextual factors. This fact indicates that the performance of automatic speech recognition can be improved by taking the contextual factors into account, which can be estimated from speech acoustics. In this study, we attempt to find global and acoustic features that are associated with those contextual factors, then integrate those features into Recurrent Neural Network (RNN) language models for speech recognition. In experiments, using Japanese spontaneous speech corpora, we examine how i-vector and openSMILE are associated with contextual factors. Then, we use those features in the reranking process of RNN-based language models. Results show that perplexity is reduced by 16% relative and word error rate is reduced by 2.1% relative for highly emotional speech.
Index Terms: contextual factors, global features, spontaneous speech, language models, adaptation, reranking

1. Introduction

Recently, we can find many automatic speech recognition (ASR) systems embedded into various electronic devices, but the input to these systems is often voice commands. As these devices become more prevalent, it should be more necessary for them to accept spontaneous speech. Here, we can point out various differences of speakers' behaviors found between voice commands and spontaneous speech. In the latter, one often communicates with others by controlling not only linguistic information but also para-linguistic and even non-verbal information such as speaking styles and gestures [1]. To understand him/her, listeners identify the spoken words while interpreting para-linguistic and non-verbal information transmitted via speech and motions, which are related to age, gender, emotion, regional accent, attitude, and so on. It is certainly possible to adapt language models to those factors by treating them as discrete labels and using class-based language models [2]. With RNN language models, however, to adapt these models, we can use raw and continuous features related to these labels and also combine different types of features very flexibly [3, 4, 5].

What kind of acoustic features are associated with contextual factors? As far as the authors know, language model adaptation to contextual factors were examined only by using acoustic features related to small linguistic units such as syllable and word [5, 6], but we can claim that long-span features are highly correlated with some contextual factors when they are acoustically realized by static bias of speech features. Then in this paper, we focus on global and acoustic features associated with contextual factors and examine how they can be used for RNN language model adaptation. In experiments, we investigate i-vector and openSMILE features extracted from individual utterances and use them for language model adaptation. The resulting models are tested using highly emotional speech corpora.

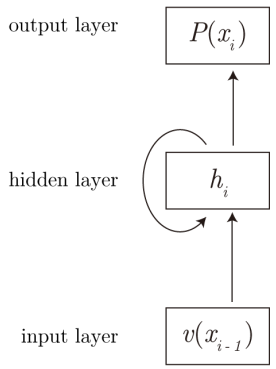


Figure 1: *Recurrent Neural Network Language Model*

2. Related works

The basic RNN language model [7] is schematically shown in Figure 1. Word x_{i-1} is converted to fixed length feature vector $v(x_{i-1}) \in \mathbb{R}^n$, and is combined with previous hidden layer $h_{i-1} \in \mathbb{R}^n$. The current hidden layer h_i is calculated as follows:

$$h_i = f(W_{hh}h_{i-1} + W_{xh}v(x_{i-1}) + b_h), \quad (1)$$

where W_{hh} and $W_{xh} \in \mathbb{R}^{n\times n}$ are weight matrices, $b_h \in \mathbb{R}^n$ is a bias vector, and $f(\cdot)$ is called activation function like hyperbolic tangent. From h_i, the following word is predicted:

$$P(x_i) = \mathrm{softmax}(W_x h_i + b_x), \quad (2)$$

where $W_x \in \mathbb{R}^{n\times v}$ is a weight matrix and $b_x \in \mathbb{R}^v$ is a bias vector. $P(x_i) \in \mathbb{R}^v$ is an output vector whose dimension v is equal to the vocabulary size. Each dimension represents the probability that its corresponding item in the vocabulary is observed after the given history. To avoid the well-known gradient vanishing problem, hidden layer prediction, denoted as Equation 1, can be replaced with Long Short-Term Memory (LSTM) [8].

As for language model adaptation based on contextual factors, we can say that there are two types of approaches of using additional features for adaptation: linguistic features and acoustic features. In the former, both local and global features were examined in [9, 10]. Here, the local features are related to morpheme [10, 11, 12] or word [9, 13] and the global features are related to sentence [14] or document [3, 10, 15]. Further, socio-situational settings were also examined for adaptation in [9].

http://dx.doi.org/10.21437/Interspeech.2017-717

On the other hand in the latter, only the prosodic features extracted from segments of a word unit or a syllable unit were used in previous studies [5, 6]. In these works, the durations of word, pause, syllable and the F_0 statistics of syllable were used to analyze the context and they were inserted as additional features to the hidden and output layers of RNN language models.

3. Language model adaptation with global and acoustic features

3.1. Global and acoustic features

Contextual factors such as speaker identity and emotion can modify an utterance globally. For example, while the variances of F_0 and power become higher when a speaker speaks intensely, they become lower when he/she does indifferently [16]. In this paper, i-vector [17] and openSMILE [18] features are adopted because they are extracted globally from individual utterances and are often used to estimate speaker identity and emotion, respectively.

3.2. Integrating the features into RNN language models

As Fu [5] integrated prosodic features into RNN language models, we input i-vector and openSMILE features to both the hidden layer and the output layer of RNN language models (Figure 2). They are fed to a simple feed forward neural network:

$$d = \tanh(W_d a(S) + b_d), \tag{3}$$

where $a(S) \in \mathbb{R}^j$ indicates global and acoustic features extracted from raw speech signals S, $W_d \in \mathbb{R}^{j \times k}$ is a weight matrix, and $b_A \in \mathbb{R}^k$ is a bias vector.

As we use the LSTM structure in the recurrent part of Figure 2, each word is predicted by the following equations:

$$P(x_{i+1}|x_1, \cdots, x_i) = \mathrm{softmax}(W_{xh} h_i + W_{xd} d + b_x), \tag{4}$$

$$i_i = \sigma(W_{ix} x_i + W_{ih} h_{i-1} + W_{id} d + b_i), \tag{5}$$
$$f_i = \sigma(W_{fx} x_i + W_{fh} h_{i-1} + W_{fd} d + b_f), \tag{6}$$
$$o_i = \sigma(W_{ox} x_i + W_{oh} h_{i-1} + W_{od} d + b_o), \tag{7}$$
$$g_i = \tanh(W_{gx} x_i + W_{gh} h_{i-1} + W_{gd} d + b_g), \tag{8}$$
$$c_i = f_i \otimes c_{i-1} + i_i \otimes g_i, \tag{9}$$
$$h_i = o_i \otimes \tanh(c_i), \tag{10}$$

where W_{*x} and $W_{*h} \in \mathbb{R}^{n \times n}$, and $W_{*d} \in \mathbb{R}^{k \times n}$ are weight matrices. $b_x \in \mathbb{R}^v$ and $b_i, b_f, b_o, b_g, g_i \in \mathbb{R}^n$ are bias vectors. $\sigma(\cdot)$ and $\tanh(\cdot)$ are the element-wise sigmoid and hyperbolic tangent functions. $\otimes$ multiplies arguments elementwisely. The entire network of Figure 2 is trained by backpropagation through time [19].

In the following section, we examine analytically and experimentally how i-vector and openSMILE features are associated with contextual factors. After that, in Section 5, we carry out speech recognition experiments to verify the effectiveness of our proposed method.

4. Experiments on feature mapping

In this section, analytical experiments are done to examine how i-vector and openSMILE features are associated with some predefined contextual labels assigned to the utterances in the corpora that we use. For easy inspection, those features will be mapped and visualized by t-SNE [20].

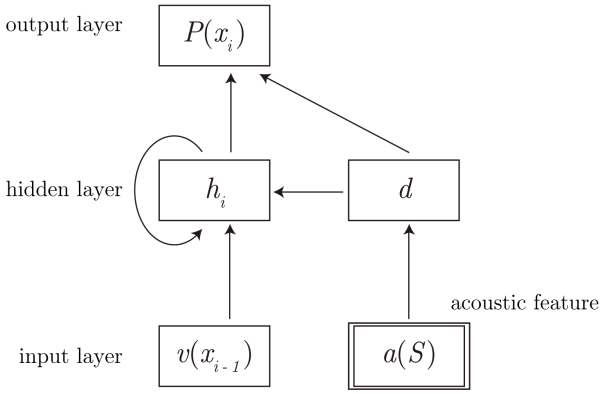


Figure 2: *Recurrent Neural Network Language Model with Prosodic Features*

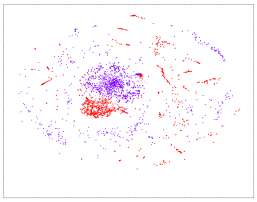
Table 1: *Data sets used in the mapping experiments*

type	#speeches	#words	length (hours)
CSJ1	987	3,411,409	274.4
CSJ2	1,715	3,859,488	329.9
CSJ3	58	156,323	12.2
CSJ4	523	221,948	21.0
CSJ5	19	303,795	24.1
UUDB	3,426	19,983	2.1
OGVC	7,621	35,931	13.2

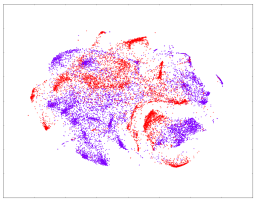
4.1. Experimental settings

Three Japanese speech corpora were used in this analysis, which are listed in Table 1. Corpus of Spontaneous Japanese (CSJ) [21] is the biggest corpus as spoken Japanese corpus. It consists of 5 sub-corpora: 1) academic presentations, 2) simulated public speeches, 3) reading, 4) dialogue and 5) others. Since the first and second sub-corpora occupy over 90% of CSJ, we can say that CSJ is basically a corpus of formal speech. Utsunomiya University spoken Dialogue Database for paralingistic information studies (UUDB) [22] and Online Gaming Voice Chat corpus with emotion label (OGVC) [23] are also spontaneous corpora which are designed to analyze natural conversation containing various kinds of para-linguistic information. Considering the above-mentioned difference between CSJ and the other two, we call UUDB and OGVC as emotional corpora henceforth. In UUDB, seven pairs of university students joined recording, where each pair were talking to sort randomly presented cartoon pictures. In OGVC, some pairs of young players are playing massively multiplayer online role-playing games with voice chat, so emotional expressions are easily detected. It is also found in UUDB and OGVC that each utterance has a smaller number of words compared to CSJ. Further in OGVC, a larger number of proper nouns such as character names and city names are used.

As for global and acoustic feature extraction, a 100-dimensional i-vector was extracted from each utterance in the three corpora by KALDI toolkit [24]. 13-dimensional MFCCs were extracted using a 25-ms Hamming window and a 10-ms window shift. After estimating a 256-mixture Gaussian model,

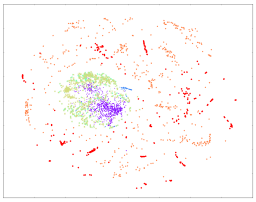


i-vectors

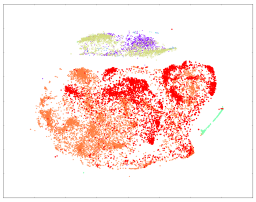


openSMILE features

Figure 3: *Gender-based visualization of global features: Two colors represent two genders.*



i-vectors



openSMILE features

Figure 4: *Speech-type-based visualization of global features: Seven colors represent seven speech types.*

the UBM was formed from the training part of CSJ. In addition, the "emobase" feature set were used as openSMILE features. This feature set is specially designed to recognize emotion, which is a set of 988 features obtained by taking 19 statistics for 26 low-level descriptors and their delta.

4.2. Results and discussion

The i-vectors and openSMILE features are plotted on a two-dimensional plane through t-SNE, where colors are used to indicate differences of gender and speech type. Figure 3 is gender-based visualization and Figure 4 is speech-type-based visualization, where seven different types of speeches (five types in CSJ, UUDB, and OJVC) are represented by seven colors.

In both plots of Figure 3, although the entire space is not clearly divided into two regions, male and female, if we are allowed to divide the entire space into small regions, each region seems to have the major gender, male or female. In Figure 4, two speech types of UUDB and OGVC, which are emotional corpora, are drawn in orange and red, respectively. Five types of CSJ speeches are plotted in other five colors. It is clearly shown that CSJ speeches, which occupy more than 95% of the three corpora, are clustered into small regions. In other words, less than 5% of speeches are occupying an extremely large portion of the entire space, showing an extremely large acoustic diversity of emotional speeches in UUDB and OGVC.

As for differences between i-vectors and openSMILE features, while feature distribution is very continuous in openSMILE features both in Figure 3 and Figure 4, a certain amount of data are plotted in a scattered way in i-vectors. It is found that most of the scattered plots correspond to short utterances in UUDB and OGVC, which contain a few words. In this case, an utterance contains only a small number of phonemes and, since insufficient data are available for MAP adaptation, the resulting GMM supervector will have different properties from one another.

5. Experiments on speech recognition

To evaluate the effect of integrating global and acoustic features into RNN language models, we carried out an ASR experiment using Japanese spontaneous corpora. The performances of RNN language models were evaluated by both adjusted perplexity (APP) [25] and word error rate (WER). APP is designed to reduce probabilities of unknown words by using a penalty term. APP for word sequence $X = \{x_1, \cdots, x_N\}$ is given by

$$APP(X) = \left(\prod_{i=1}^{N} \frac{1}{p(x_i|x_1, \cdots, x_{i-1})} {m_{unk}}^{\frac{1}{n_u}} \right)^{\frac{1}{N}}, \quad (11)$$

Table 2: *Data sets used in the ASR experiments*

set	type	#utterances	#words
training	CSJ1	148,268	3,411,409
	CSJ2	238,055	3,859,488
	CSJ3	23,263	156,323
	CSJ4	16,909	221,948
	CSJ5	13,991	303,795
	UUDB	3,426	19,983
	OGVC	7,621	35,931
develop	CSJ1	4,000	100,385
testing	eval1	1,272	28,923
	eval2	1,292	29,716
	eval3	1,385	19,668
	UUDB	1,414	7,673
	OGVC	1,493	8,755

where m_{unk} is the number of kinds of unknown words, and n_{unk} is the number of observed unknown words.

5.1. Experimental settings

The same Japanese corpora that were used in Section 4 were used again in this experiment. The baseline ASR system was constructed by following the KALDI CSJ recipe. It can generate N-best hypotheses for an input utterance by using the KALDI CSJ default language models, which are trigram models. Our proposed language model was evaluated in reranking the hypotheses as postprocessing. Here, we used three CSJ standard evaluation sets as testing utterances, where each set consisted of ten speeches. In addition, after dividing each of UUDB and OGVC into training and testing parts with no speaker overlap, their testing parts were also used to evaluate our model. The speech corpora used here are summerized in Table 2 and the testing sets of CSJ are denoted as eval1, eval2, and eval3.

Before experiments, we modified the transcriptions provided by UUDB and OGVC. In these transcriptions, vocal behaviors of breath, cough, laugh, and sigh were annotated using special symbols. We converted them into tokens by using a text analyzer Mecab [26] and the new tokens were also registered into the pronunciation lexicon used in our ASR system. The size of the resulting word vocabulary (v) became 65,751, and out-of-vocabulary words were replaced by <unk>.

In Equations (4) to (10) of our RNN language model, the dimensionality of word embedding and hidden layer and that of global and acoustic features (i-vector or openSMILE) are denoted as n and k, respectively. Here, k is the dimensionality after network-based dimension reduction. In the experiments,

Table 3: *Adjusted perplexities in test data sets*

model	eval1	eval2	eval3	UUDB	OGVC
3-gram	78.8	83.1	90.8	289.9	647.5
LSTM-RNN	57.5	62.6	58.7	177.8	345.5
+i-vector	**57.2**	**61.2**	**57.4**	165.6	333.1
+openSMILE	57.3	63.2	62.1	**149.6**	**289.8**

Table 4: *WERs (%) in test data sets*

model	eval1	eval2	eval3	UUDB	OGVC
(3-gram)	10.80	8.64	9.03	41.29	46.92
LSTM-RNN	**10.28**	8.27	**8.73**	42.08	46.87
+i-vector	10.29	8.27	8.74	41.62	**46.68**
+openSMILE	10.32	**8.25**	8.77	**41.18**	47.00

n and k were 200 and 10.

In RNN training, the cross-entropy error was backpropagated using stochastic gradient descent, where the length of word history was 35. The learning rate was scheduled by ADAM [27]. Parameters of the RNN were randomly initialized over a uniform distribution of $[-0.1, 0.1]$. For regularization, we used dropout [28] with probability 0.5 and mini-batch training with batch size 64. The norm of the gradients was constrained to be less than or equal to 5, so that if the L_2 norm of the gradient exceeds 5 then it will be set to 5 before updating.

Our RNN language models were incorporated into the ASR process by rescoring the 100-best hypotheses, which were generated from the lattices of the baseline ASR system. To compute the final score for each hypothesis, two scores, a trigram score and a RNN score, are interpolated. Optimization of the interpolation rate was done by selecting the best rate from 0.25, 0.5, and 0.75, which can maximize the recognition accuracy of a development set (See Table 2).

5.2. Results and discussion

The adjusted perplexity scores of four models of KALDI CSJ trigram, LSTM-RNN, LSTM-RNN with i-vector, and LSTM-RNN with openSMILE are shown in Table 3. By comparing the APP of LSTM-RNN and that of its enhanced version with i-vectors, we can observe 1-2% improvements for CSJ evaluation sets and 4-7% improvements for emotional test data.

On the other hand, adaptation with openSMILE features lead to a slight negative effect on CSJ evaluation set, nevertheless, they improve APP for emotional test data remarkably by 16%. Differences of improvement between formal speech and emotional speech are attributed to domain mismatch between training and testing. As shown in 2, less than 5% of the training data are emotional and effectiveness of adaptation is easily found in UUDB and OGVC.

We can show an example of perplexity reduction. The APP of a test utterance in OGVC: "<laugh> これ は やばい って (This must kill you.) <laugh>" decreases from 37.9 to 30.5 with openSMILE features integrated into LSTM-RNN. In this utterance, two verbal expressions are extremely characteristic in terms of their acoustic and para-linguistic salience. They are <laugh> and やばい. Especially the latter one is found often in utterances of the younger generation with much excitement.

The WERs for four language models tested are shown in Table 4. The four models are the baseline trigram, LSTM-RNN, and their adapted models with global and acoustic features. We can observe effectiveness of RNN rescoring for all the CSJ evaluation data sets, where almost no domain mismatch exists between training and testing. With i-vectors or openSMILE features, however, additional improvements are difficult to find.

As for emotional speech, there are 2.1% improvement for UUDB test data with openSMILE features and 0.4% improvement for OGVC test data with i-vector features. The magnitude of WER reduction is smaller than that of APP reduction, which is often reported in [5, 6]. Compared to those previous studies, our testing utterances were by far more emotional and the recognition accuracy of the baseline ASR system is much lower. This means that the quality of N-best hypotheses has to be degraded in our experiments, which reduces usability of rescoring.

6. Summary

In this paper, we proposed to integrate global and acoustic features into RNN language models for spontaneous speech recognition. From the results in Section 4, we can claim that both i-vector and openSMILE features include information on speaker gender and speech type. Further, openSMILE features are distributed in a continuous way, but i-vectors are distributed in a scattered way when their length is short. The ASR experiments in Section 5 reveal that i-vector and openSMILE features are very effective to reduce APP, but their effectiveness on ASR is limited in the current experimental setting. Here, the baseline acoustic models were built using the same training data of CSJ, UUDB, and OGVC that were used for the baseline language models. Since emotional utterances are very rare, the quality of the N-best hypotheses was unsatisfactory for reranking.

For future work, we will re-examine our proposed method of language model adaptation with adapted acoustic models, which are highly expected to give us more accurate N-best hypotheses. Also we are interested in using other emotional corpora which give us lower WER and conducting analytical investigations on what kind of different contributions are found for WER reduction between i-vector and openSMILE features. Further, we will optimize the network structure for language models and combine both local and global features extracted acoustically and linguistically for adaptation.

7. References

[1] D. McNeill, *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992.

[2] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based n-gram models of natural language," *Computational linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

[3] W. Jin, T. He, Y. Qian, and K. Yu, "Paragraph vector based topic model for language model adaptation," in *INTERSPEECH*, 2015, pp. 3516–3520.

[4] X. Chen, T. Tan, X. Liu, P. Lanchantin, M. Wan, M. J. Gales, and P. C. Woodland, "Recurrent neural network language model adaptation for multi-genre broadcast speech recognition." in *INTERSPEECH*, 2015, pp. 3511–3515.

[5] T. Fu, Y. Han, X. Li, Y. Liu, and X. Wu, "Integrating prosodic information into recurrent neural network language model for speech recognition," in *APSIPA*, 2015, pp. 1194–1197.

[6] S. R. Gangireddy, S. Renals, Y. Nankaku, and A. Lee, "Prosodically-enhanced recurrent neural network language models," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015, pp. 2390–2394.

[7] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *INTERSPEECH*, 2010, pp. 1045–1048.

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, pp. 1735–1780, 1997.

[9] Y. Shi, P. Wiggers, and C. M. Jonker, "Towards recurrent neural networks language models with linguistic and contextual features." in *INTERSPEECH*, 2012, pp. 1664–1667.

[10] A. Mansikkaniemi and M. Kurimo, "Unsupervised topic adaptation for morph-based speech recognition." in *INTERSPEECH*, 2013, pp. 2693–2697.

[11] D. Renshaw and K. B. Hall, "Long short-term memory language models with additive morphological features for automatic speech recognition," in *ICASSP*, 2015, pp. 5246–5250.

[12] E. Arisoy and M. Saraçlar, "Multi-stream long short-term memory neural network language model," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[13] R. Masumura, H. Masataki, T. Oba, O. Yoshioka, and S. Takahashi, "Use of latent words language models in asr: a sampling-based implementation," in *Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 8445–8449.

[14] Y. Ji, G. Haffari, and J. Eisenstein, "A latent variable recurrent neural network for discourse-driven language models," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 332–342.

[15] Z. Tüske, K. Irie, R. Schlüter, and H. Ney, "Investigation on log-linear interpolation of multi-domain neural network language model," in *ICASSP*, 2016, pp. 6005–6009.

[16] H. Mori, K. Maekawa, and H. Kasuya, *What Does Speech Convey?: Speech Science of Emotion, Paralinguistic Information, and Speaker Individuality*, ser. Acoustic science series / Phonetic Society of Japan. Corona Publishing Co., Ltd., 2014, no. 12.

[17] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[18] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.

[19] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[20] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[21] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of japanese," in *Second International Conference on Language Resources and Evaluation*, 2000, pp. 947–952.

[22] H. Mori, T. Satake, M. Nakamura, and H. Kasuya, "Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics," *Speech Communication*, pp. 36–50, 2011.

[23] Y. Arimoto, H. Kawatsu, S. Ohno, and H. Iida, "Emotion recognition in spontaneous emotional speech for anonymity-protected voice chat systems," in *INTERSPEECH*, 2008, pp. 322–325.

[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[25] J. Ueberla, "Analysing a simple language model· some general conclusions for language models for speech recognition," *Computer Speech & Language*, vol. 8, no. 2, pp. 153–176, 1994.

[26] T. Kudo, "Mecab: Yet another part-of-speech and morphological analyzer," *http://mecab.sourceforge.net/*, 2005.

[27] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *The International Conference on Learning Representations*, 2015.

[28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.