INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Neural Whispered Speech Detection with Imbalanced Learning

Takanori Ashihara, Yusuke Shinohara, Hiroshi Sato, Takafumi Moriya, Kiyoaki Matsui, Takaaki Fukutomi, Yoshikazu Yamaguchi, Yushi Aono

NTT Media Intelligence Laboratories, NTT Corporation, Japan

takanori.ashihara.vk@hco.ntt.co.jp

Abstract

In this paper, we present a neural whispered-speech detection technique that offers utterance-level classification of whispered and non-whispered speech exhibiting imbalanced data distributions. Previous studies have shown that machine learning models trained on a large amount of whispered and non-whispered utterances perform remarkably well for whispered speech detection. However, it is often difficult to collect large numbers of whispered utterances. In this paper, we propose a method to train neural whispered speech detectors from a small amount of whispered utterances in combination with a large amount of non-whispered utterances. In doing so, special care is taken to ensure that severely imbalanced datasets can effectively train neural networks. Specifically, we use a class-aware sampling method for training neural networks. To evaluate the networks, we gather test samples recorded by both condenser and smartphone microphones at different distances from the speakers to simulate practical environments. Experiments show the importance of imbalanced learning in enhancing the performance of utterance level classifiers.

Index Terms: whispered speech, vocal effort, deep neural networks, imbalanced learning, class-aware sampling.

1. Introduction

Improvements in speech and language technologies have resulted in the deployment of speech-related systems such as automatic speech recognition, speaker recognition, and speech synthesis in real environments by using smartphones or smart speakers. The mobility of those devices encourages the adoption of speech applications in public/private settings where loud neutrally phonated speech is proscribed (e. g. in a library) or the speaker would like to protect some confidential information (e. g. during a meeting). In those situations/places, a speaker wishing to be quiet uses murmured communication like a whisper for application input. However, those applications basically assume normally/neutrally phonated speech input and the performance of speech processing systems degrade significantly when the data used for training the systems doesn't include sufficient whispered speech [1]. Of course, whispering is common speech mode in real environments as described above and should be included in large scale sample databases. Thus, for extracting whispered speech utterances from existing databases, a whispered speech detector is important to realize effective training. In addition, whispered speech detection itself is useful for enhancing the user experience such as the whispered mode in Alexa [2] which can respond to a whisper with a whisper.

Existing approaches to whispered speech detection have mainly focused on feature engineering to accentuate the differences between whispered and neutral speech. In [3], the signal processing-based approach is proposed; it computes the ratio of the spectral power in a low frequency-band to the power in a high frequency-band to ensure robustness to additive noise. Alternatively, Zhang and Hansen [4] investigated model-based classification of five speech modes, whisper, soft, normal, loud, and shouted speech, and their subsequent works centered on whispered speech detection [5, 6]. Their techniques are based on the Gaussian Mixture Model (GMM) trained by engineered features, specifically spectral information entropy (SIE) from divided sub-frames and sub-bands and the SIE ratio between the high band and the low band. More recently, [7] introduced deep neural networks (DNNs) to carry out the frame-level classification of whispered/non-whispered speech. This work evaluated long short-term memory (LSTM) in a comparison with the simple multilayer perceptron (MLP) as a baseline; the former achieved a higher frame accuracy. Furthermore, the models trained by using only log mel-filterbank coefficients performed remarkably well even without the additional engineering features because DNNs are able to automatically acquire internal representations for classification from a large number of features. For the task of predicting utterance-level classification, inference modules, such as the mean of posterior score over the frames, can be introduced to frame-level DNN classifiers.

However, it is difficult in many cases to accumulate a sufficient amount of whispered utterances for training DNNs. Therefore, in this study, while we follow the DNN based detector approach, our focus is on the design of low-resource learning techniques that permit small amounts of whispered speech data to yield effective training.

To alleviate the low-resource problem, this paper assumes that a large amount of real non-whispered speech data is available and we examine the imbalanced learning that results when the classifier is trained using a small amount of whispered utterances in combination non-whispered acoustic variability. Note that we use only log mel-filterbank coefficients because this feature is shared by other speech processing systems (e. g. speech recognition, voice activity detection) and reduces the computational demands. It is well known that extremely imbalanced data distributions have a detrimental effect on model performance. The recent systematic study of [8] showed that the oversampling strategy with convolutional neural networks (CNNs) offers higher accuracy than the undersampling strategy and also does not lead to overfitting unlike some classical machine learning models. In this article, we apply one of the oversampling methods called class-aware sampling [9] as it is well suits the training of neural networks. We also compare CNNs that can directly optimize the utterance-level target with frame-level DNNs [7].

The remainder of this article is organized as follows. Section 2 shows related work. Section 3 overviews imbalanced learning and class-aware sampling. Experimental results are presented in Section 4. Finally, the conclusion is drawn in Section 5.

http://dx.doi.org/10.21437/Interspeech.2019-2161

2. Related work

The acoustic characteristics of whispered speech have been analyzed in several languages such as English [4, 10], Japanese [11], Mandarin [12] and Czech [13]. In comparison with neutral speech, whispered speech has three key characteristics. *i*) spectral power is lowered. *ii*) vowel formant frequencies at lower frequencies such as F1/F2 are shifted to higher frequencies. *iii*) spectral tilt is much flatter. These studies indicate that the acoustic differences between whispered/neutral speech is language independent.

Imbalanced data is a common problem where the number of examples of class A is significantly higher than that of class B in the binary classification task. To deal with imbalance, there are two main strategies: data-level approach and classifier-level approach [8, 14].

The data-level approach such as oversampling and undersampling is dominant for addressing imbalance; it harmonizes the number of examples for both classes. Several oversampling variants exist, but the most basic one is random oversampling from the minority class [15]. A more advanced technique called SMOTE [16] employs synthetic examples for the minority class. Class-aware sampling method [9] is included in this category as a neural networks specific approach. Undersampling, on the other hand, removes majority class examples [17, 18].

The classifier-level approach is also used for addressing imbalance. Cost-sensitive methods use specialized costs to address the problem of misclassification of examples [19]. The approach of one-class classification uses only one class data during training, and calculates a distance or error between the input data during testing. This approach is often used in novelty detection or anomaly detection task [20].

3. Method

3.1. Training with Imbalanced data

Imbalanced data is a combination of minority and majority (in terms of amount) class data. In the case of whispered speech utterance detection, for example, minority class data corresponds to whispered speech data and majority class data corresponds to non-whispered speech data.

If a model is trained with a severely imbalanced data distribution in the same way as the assumed balanced data, the resulting model would misclassify most minority class instances as majority class (the strong bias treats minority class as noise) which degrades performance. This negative effect of class imbalance was seen in our preliminary experiment on deep neural network models. According to recent evidence of imbalanced learning with CNNs for image recognition [8], oversampling outperforms undersampling in most cases. This paper examines the effect of oversampling with CNNs on whispered utterance detection performance.

3.2. Class-aware sampling

Class-aware sampling, one of the oversampling methods for neural networks, was proposed in [9]. This algorithm, shown as Algorithm 1, addresses the issue of data imbalance by sampling the same numbers of data in each class while making mini-batches for a stochastic gradient descent (SGD) algorithm.

A data list of both classes is prepared and training batches are created by sampling the list as uniformly as possible with respect to classes. Of particular note, the number of each class examples is one half the minibatch size in our study. If the sampling reaches the end of data list, this class list is shuffled randomly to reorder the speech data. In doing so, neural network learning is based on well-balanced class data with additional majority class acoustic variability. This algorithm can be interpreted as a form of oversampling because the minor class data appears many times in the training phase.

Algorithm 1 Generate mini-batch with class-aware sampling

Input:
l_{c1} : Minor amount class data list (e. g. whispered speech);
l_{c2} : Major amount class data list (e. g. neutral speech);
n : Batch size;
Output: A mini-batch
while End of training **do**;
  Get minor class data x_{c1} (size of $n/2$) from l_{c1};
  if End of l_{c1} **then**;
    SHUFFLE(l_{c1});
  Get major class data x_{c2} (size of $n/2$) from l_{c2};
  if End of l_{c2} **then**;
    SHUFFLE(l_{c2});
  return CONCATENATE(x_{c1}, x_{c2});

3.3. Utterance-level DNNs

The goal of this paper is to classify whispered speech on a per-utterance rather than per-frame basis. Therefore, we follow frame-level DNNs, but with structural extensions to allow utterance-level classification. A schematic diagram of the utterance-level model structure is given in Figure 1.

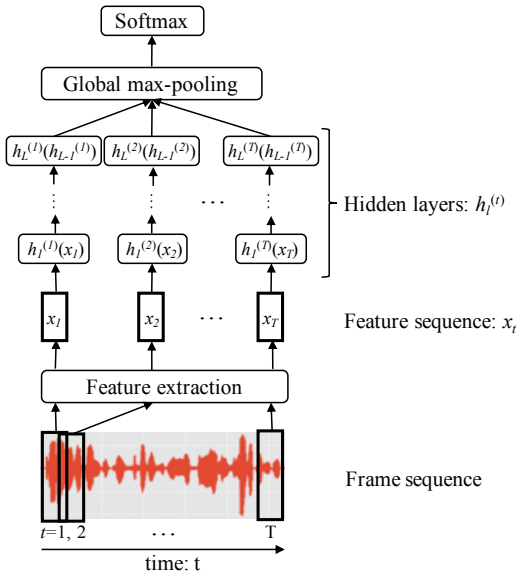

Figure 1: *The schematic architecture of utterance-level model that can jointly learn utterance-level representations with utterance-level inference module*

Specifically, we add a global max-pooling layer after the frame-level convolutional layers $h_l^{(t)}$ from sequence of feature x_t and hence these models can optimize the inference module of utterances simultaneously. Global max-pooling for joining temporal features has been used in the background of audio event

classification [21] or text classification [22].

4. Experiment and result

We begin with an explanation of the corpus used in our experiment in Section 4.1. Second, the metrics used to evaluate the proposed models are described in Section 4.2 and the details of the experimental setup used to train models are given in Section 4.3. Finally, the evaluation results are shown in Section 4.4.

4.1. Data set

The experiments utilized three databases: CHAracterizing INdividual Speakers (CHAINS) [23], Corpus of Spontaneous Japanese (CSJ) [24], and newly created test set. An overview of each data set is shown in Table 1.

Table 1: *Overview of each corpus. Note that the value of created corpus is based on one microphone.*

Corpus	# of speaker	# of hour	# of utterance
CHAINS [23]	36	4.7	4896
CSJ [24]	1388	645.0	417384
Created	10	4.4	2000

The CHAINS corpus consists of several speaking styles of each speaker and is mainly used in speaker identification research. Since the purpose of this article is to construct a whispered speech detection model, we extract whispered and non-whispered related speech from the corpus, specifically WHISPER reading and SOLO reading. WHISPER reading occupies about 2.4 hours of the corpus. Several utterances of fable reading were recorded successively, and thus we divide the speech into utterance units in advance. We split samples into those for 5 speakers and the others to yield a development set and training set, respectably.

The CSJ corpus mostly consists of academic presentation speech and simulated public speech. Because the data was recorded in a presentation situation, we treat all data as non-whispered speech data. To examine the effect of class-aware sampling, we treat CHAINS corpus as the balanced class distribution data and CHAINS+CSJ as the imbalanced class distribution data in the experiment. Data preparation is performed by using CSJ recipe in Kaldi [25].

In addition to open-access corpuses, we recorded speech data anew to evaluate our approach. Our ultimate goal for whisper detection is to classify input speech as either whisper speech or not robustly in even practical environments such as smartphone applications, smart speakers and so on. In real use cases, for example, when someone speaks in a whisper, they speak close to the microphone for preserve privacy. On the other hand, people using smart speakers tend to speak at various distances from the devices. To capture those realistic conditions, the capture system (16kHz sampling) used two types of microphone, condenser microphone and smartphone microphones (Apple iPhone 8, SONY Xperia XZ1, SAMSUNG Galaxy 8+ and HUAWEI P20 pro), for each subject. Three microphone standoff distances were used: 3 cm (i. e., very close condition), 15 cm (i. e., neutral condition) and 50 cm (i. e., far-field condition). Overall, 6 microphones captured samples utterance in two speech modes (neutral and whisper) simultaneously. The subjects were 5 male and 5 female native Japanese speaking subjects. As each participant spoke 100 utterances per speech mode, each microphone captured a total of 2000 utterances. Note that the test database was balanced unlike the training database.

4.2. Evaluation metrics

The binary classification task is often evaluated by accuracy, precision, recall, and F_1 measure. Because our approach uses data with imbalanced class distribution, the probability threshold for best accuracy tends to be biased [14]. Therefore, we use the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) from binary label and prediction probability as evaluation metrics. ROC point can be obtained by True Positive Rate (TRP) and False Positive Rate (FPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \quad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{1}$$

where TP, FN, FP and TN represent True Positive, False Negative, False Positive, True Negative, respectively. ROC curve can also be obtained by ROC points for all possible prediction thresholds and ROC-AUC is calculated as the area under the ROC curve. The ideal value of ROC-AUC is 1.0, that is all positive samples are classified as positive (i. e., TPR is 1.0) and no negative samples are misclassified as positive (i. e., FPR is 0). A random prediction, on the other hand, will represent an ROC-AUC score less than 0.5.

4.3. Experimental setup

To train the whispered speech utterance detection model, we transform the raw audio into 64 dimensional log mel-filterbank coefficients in every 25 ms window with 10 ms fixed frame rate, which follows [7]. We note that this paper does not use the engineering feature as described in Section 1 to ease the computation burden. As the normalizing feature, we subject the data set to pre-estimated global cepstral mean and variance normalization (CMVN). Signal preprocessing is performed by using the Kaldi toolkit [26]. Whispered/non-whispered labels are propagated to all frames for the frame-level model which is in marked contrast to utterances in the utterance-level model.

The experiment examined three model architectures: MLP, LSTM and CNN. In [7], MLP and LSTM models were trained frame-by-frame and so, the output is the probability of each frame being whispered speech or not. Therefore, when judging whether an utterance as either whispered speech utterance or not, an inference module is needed. According to their findings, since the mean of all frame scores yields the best inference overall, we follow the module for utterance-level inference. The MLP model architecture consists of 3 hidden fully-connected layers with 40 nodes with ReLU activation function and the LSTM model consists of 2 hidden layers with 64 cells.

In addition to the frame level models, we train the CNN model which outputs utterance-level scores directly from frame-level features. The CNN model architecture consists of 4 convolutional layers with $(7\times3\times64)$ filters with max-pooling (1×2), $(1 \times 7 \times 128)$ filters with max-pooling (1×4), $(1 \times 8 \times 256)$ filters, and $(7 \times 1 \times 512)$ filters with global max-pooling with dropout. Finally, the output vector of this convolutional layers is sent to 256 nodes fully-connected layer with dropout followed by softmax classifier (Figure 2). Each convolutional layer has ReLU activation function. The padding method is basically same except for the 3rd layer which uses valid padding. The dropout rate is 0.2.

All models are optimized by the SGD algorithm with learning rate of 0.01 and momentum of 0.9. The validation by using

development set is conducted at the end of each training epoch. Note that each epoch is total duration of the entire training set in balanced training and total duration of only whispered speech data in imbalanced training.

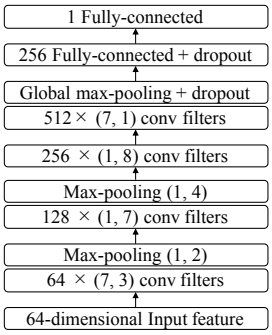

Figure 2: *CNN model structure that can perform utterance-level classification.*

4.4. Results

The experiment's results are shown in Table 2 and Table 3. Note that the ROC-AUC values were calculated from TP and FP percentage and hence the values move between 0 and 100. Table 2 show condenser microphone results with 3 distances for each model, and Table 3 shows the smartphone microphone results with 3 distances for each model. The first two rows of the tables show the baseline models: frame-level MLP and LSTM. The last two rows show the proposed models: utterance-level CNN and utterance-level CNN with imbalanced data. Each value is the average of 3 times of trials.

Table 2: *Experimental result of the condenser microphone in each distance compared in terms of AUC score.*

Model	3 cm	15 cm	50 cm
MLP	93.53	96.93	83.29
LSTM	95.47	91.12	76.28
CNN	99.23	99.93	87.23
CNN (imbalance)	**99.63**	**99.96**	**98.82**

Table 3: *Experimental result of the smartphone microphone in each distance compared in terms of AUC score.*

Model	3 cm	15 cm	50 cm
MLP	95.34	89.98	88.33
LSTM	93.42	80.42	66.97
CNN	98.84	97.63	73.09
CNN (imbalance)	**99.38**	**97.36**	**91.68**

Compared with the condenser microphone in Table 2, the smartphone microphone in Table 3 yielded degraded overall performance. Since training used only the speech data captured by the condenser microphone, the acoustic differences between the two recording devices impact the performance. In particular, the far-field condition shown as 50 cm in both tables reduces the performance of all models due to the acoustic differences between training and test set.

Among the baseline models, MLP offers better performance than LSTM in low-resource data condition and LSTM model degrade the performance severely with the smartphone microphone. This indicates that the LSTM model is more sensitive to the amount of the training data and the acoustic characteristics than MLP.

CNN models show better performance relative to the baseline models in even close/neutral distance conditions. This result means that direct utterance-level optimization is effective for whispered speech utterance detection. In addition, CNN models also show robustness against microphone variability. This indicates that the max-pooling layer along the frequency axis offsets spectral variation [27]. Furthermore, imbalanced learning outperforms CNN with balanced data distribution, notably in the far-field (50 cm) condition. The performance improvement is induced from the reduction of FPR and this can be confirmed by ROC curves in Figure 3 that is shown high FPR of CNN with balanced data distribution in low TPR. Therefore, the imbalanced learning helps to acquire the neural representation of non-whispered speech effectively.

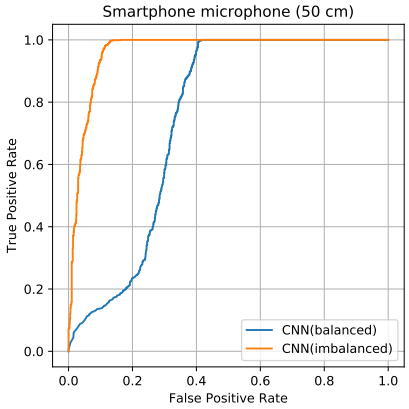

Figure 3: *ROC curves of CNN model with balanced and imbalanced data distribution on the smartphone microphone (50 cm) condition.*

5. Conclusions

In this paper, we investigated the impact of imbalanced learning on the performance of whispered speech detection, which attempts to classify an utterance as either whispered or non-whispered speech. In particular, a class-aware sampling method is used in the training phase to alleviate the effect of imbalance in the class distribution. An experiment showed that our utterance-level CNN with imbalanced learning achieved the best ROC-AUC score of almost 1.0 in close/neutral conditions and almost 0.9 in far-field condition, even though the whispered speech data totaled approximately 2.4 hours.

In future work, we would like apply other training methods to imbalanced class data, for example cost-sensitive training, one-class classification and so on. Furthermore, we will confirm the performance with data augmentation approaches such as speed perturbation [28] and pseudo-whisper [1] in combination with our approach to compensate the variability of other acoustic factors.

6. References

[1] S. Ghaffarzadegan, H. Bořil, and J. H. L. Hansen, "Generative modeling of pseudo-whisper for robust whispered speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1705–1720, 2016.

[2] Alexa is now even smarter—new features help make everyday life more convenient, safe, and entertaining. [Online]. Available: https://press.aboutamazon.com/news-releases/news-release-details/alexa-now-even-smarter-new-features-help-make-everyday-life-more

[3] S. J. Wenndt, E. J. Cupples, and R. M. Floyd, "A study on the classification of whispered and normally phonated speech," *In Proc. International Conference on Spoken Language Processing (ICSLP)*, pp. 649–652, 2002.

[4] V. C. Zhang and J. H. L. Hansen, "Analysis and classification of speech mode: whispered through shouted," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2289–2292, 2007.

[5] V. C. Zhang and J. H. L. Hansen, "An entropy based feature for whisper-island detection within audio streams," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2510–2513, 2008.

[6] V. C. Zhang and J. H. L. Hansen, "Advancements in whisper-island detection within normally phonated audio streams," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 860–863, 2009.

[7] Z. Raeesy, K. Gillespie, C. Ma, T. Drugman, J. Gu, R. Maas, A. Rastrow, and B. Hoffmeister, "LSTM-based whisper detection," *In Proc. Spoken Language Technology Workshop (SLT)*, pp. 139–144, 2018.

[8] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249 – 259, 2018.

[9] L. Shen, Z. Lin, and Q. Huang, "Relay backpropagation for effective learning of deep convolutional neural networks," *In Proc. European Conference on Computer Vision (ECCV)*, pp. 467–482, 2016.

[10] S. Ghaffarzadegan, H. Bořil, and J. H. L. Hansen, "UT-Vocal Effort II: Analysis and constrained-lexicon recognition of whispered speech," *In proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2544–2548, 2014.

[11] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Communication*, vol. 45, no. 2, pp. 139 – 152, 2005.

[12] C. Zhang, B. Li, S. Chen, and Y. Yang, "Acoustic analysis of whispery voice disguise in mandarin chinese," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1413–1416, 2018.

[13] P. Zelinka, M. Sigmund, and J. Schimmel, "Impact of vocal effort variability on automatic speech recognition," *Speech Communication*, vol. 54, no. 6, pp. 732 – 742, 2012.

[14] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Systems with Applications*, vol. 73, pp. 220 – 239, 2017.

[15] C. X. Ling and C. Li, "Data mining for direct marketing: Problems and solutions," *In Proc. International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 73–79, 1998.

[16] N. Chawla, K. Bowyer, L. O. Hall, and W. Philip Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of artificial intelligence research (JAIR)*, vol. 16, pp. 321–357, 2002.

[17] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," *In Proc. International Conference on Machine Learning (ICML)*, pp. 179–186, 1997.

[18] X. Liu, J. Wu, and Z. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 2, pp. 539–550, 2009.

[19] C. Elkan, "The foundations of cost-sensitive learning," *In Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 973–978, 2001.

[20] D. M. Tax and R. P. Duin, "Support vector data description," *Machine Learning*, vol. 54, no. 1, pp. 45–66, 2004.

[21] H. Phan, L. Hertel, M. Maaß, and A. Mertins, "Robust audio event recognition with 1-max pooling convolutional neural networks," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2016.

[22] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," *In Proc. International Conference on Machine Learning (ICML)*, pp. 160–167, 2008.

[23] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, "The CHAINS corpus: CHAracterizing INdividual Speakers," *In Proc. International Conference on Speech and Computer (SPECOM)*, pp. 431–435, 2006.

[24] S. Furui, K. Maekawa, and H. Isahara, "A japanese national project on spontaneous speech corpus and processing technology," *In Proc. ASR2000 - Automatic Speech Recognition: Challenges for the new Millenium*, pp. 244–248, 2000.

[25] T. Moriya, T. Shinozaki, and S. Watanabe, "Kaldi recipe for japanese spontaneous speech recognition and its evaluation," *Fall Meeting of Acoustic Society of Japan (ASJ)*, pp. 155–156, 2015.

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Vesel, "The Kaldi speech recognition toolkit," *In Proc. Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.

[27] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C. Laurent, Y. Bengio, and A. Courville, "Towards end-to-end speech recognition with deep convolutional neural networks," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 410–414, 2016.

[28] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 3586–3589, 2015.