American Journal of Signal Processing 2016, 6(1): 19-23
DOI: 10.5923/j.ajsp.20160601.03

Phone Labeling Based on the Probabilistic Representation for Dysarthric Speech Recognition

Yuki Takashima[1,*], Toru Nakashika[2], Tetsuya Takiguchi[1], Yasuo Ariki[1]

[1]Graduate School of System Informatics, Kobe University, Kobe, Japan
[2]Graduate School of Information Systems, The University of Electro-Communications, Chofu, Japan

Abstract In this paper, we discuss speech recognition for persons with articulation disorders resulting from athetoid cerebral palsy. Because the speech style for a person with this type of articulation disorder is quite different from a physically unimpaired person, a conventional speaker-independent acoustic model for unimpaired persons is hardly useful to recognize it. Therefore, a speaker-dependent model for a person with an articulation disorder is necessary. In our previous work, a feature extraction method using a convolutional neural network was proposed for dealing with small local fluctuation of dysarthric speech, and its effectiveness was shown in a word recognition task. The neural network needs a training label (teaching signal) to train the network using back-propagation, and the previous method used results from forced alignment using HMMs as the training label. However, as the phoneme boundary of an utterance by a dysarthric speaker is ambiguous, it is difficult to obtain the correct alignment. If a wrong alignment is used, the network may be inadequately trained. Therefore, we propose a probabilistic phoneme labeling method using the Gaussian distribution. In contrast to the general approach, we deal with the phoneme label as the soft label, that is, our proposed label takes the continuous value. This approach is effective for the dysarthric speech which is the ambiguity of the phoneme boundary. The effectiveness of this method has been confirmed by comparing its effectiveness with that of forced alignment.

Keywords Articulation disorders, Feature extraction, Convolutional neural network, Bottleneck feature, Phoneme labelling

1. Introduction

Recently, the importance of information technology in the welfare-related fields has increased. For example, sign language recognition using image recognition technology [1], text reading systems from natural scene images [2], and the design of wearable speech synthesizers for voice disorders [3] have been studied. However, there has been very little research on orally-challenged people, such as those with speech impediments. It is hoped that speech recognition systems will one day be able to recognize their voices.autism spectrum disorders and typically developing children.

One of the causes of speech impediments is cerebral palsy. Cerebral palsy results from damage to the central nervous sys-tem, and the damage causes movement disorders. There are various types of cerebral palsy. In this paper, we focused on persons with articulation disorders resulting from the athetoid type as in [4]. Athetoid symptoms develop in about 10-15% of cerebral palsy sufferers. Some people who have difficulty in speaking can communicate with others using sign language recognition or a speech synthesis system. However, many of those who have articulation disorders resulting from athetoid cerebral palsy are physically impaired, making the use of sign language difficult or impossible and speaking is the only communication method they have available. Several works for dysarthric speech recognition have been proposed. In [4], Mastumasa et al. investigated a Metamodel [5] and an acoustic model approach to increase recognition accuracy. In [6], Christensen et al. modified the pronunciation so they represent the specific speech impairments of the speaker. In [7], Christensen et al. investigated adaptation from out-of-domain (normal speech) models into the target domain (disordered speech) focusing on the feature extraction stage. Our previous work [8] also employed a convolutional neural network (CNN [9, 10, 11]) to deal with small local fluctuations of dysarthric speech.

Because the speaking style of persons with articulation disorders is quite different from physically unimpaired persons due to the involuntary movement of their muscles, the conventional speaker-independent acoustic model is not very useful, and the recognition accuracy is considerably low. For dysarthric speech recognition, we previously proposed the robust feature extraction method using a convolutive bottleneck network (CBN [17]) that consists of a CNN and a bottleneck layer. A CNN is regarded as a

* Corresponding author:
y.takasima@me.cs.scitec.kobe-u.ac.jp (Yuki Takashima)
Published online at http://journal.sapub.org/ajsp

successful tool and has been widely used in recent years for various tasks, such as image analysis [12, 13, 14], spoken language [15], and music recognition [16]. A CNN consists of a pipeline of convolution and pooling operations followed by a multi-layer perceptron. In dysarthric speech, the key points in time-spectral local areas of an input feature map are often shifted slightly due to the fluctuation of the speech uttered by persons with articulation disorders. Thanks to the convolution and pooling operations, we can train the CNN robustly to deal with small local fluctuations. In [8], the networks were trained by back-propagation using training labels obtained from the forced alignment using HMMs (hidden Markov models); however, it is difficult to obtain the correct alignment because of the unclear spectra. The use of wrong training labels may result in the incorrect training of the networks.

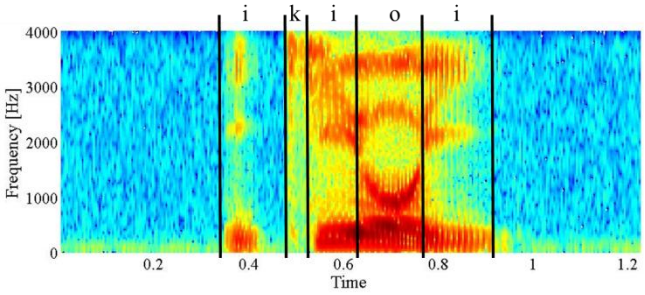

Figure 1. Example of a spectrogram for /ikioi/ spoken by a physically unimpaired person

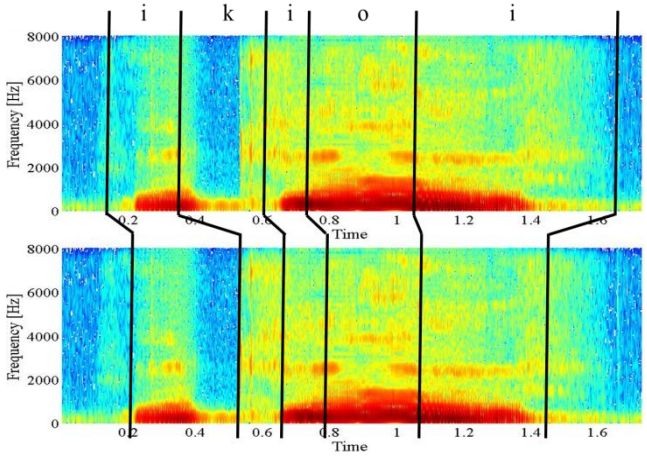

Figure 2. Example of a spectrogram for /ikioi/ spoken by a person with an articulation disorder using forced alignment (top) and manual alignment (bottom)

Figs. 1 and 2 show the spectrograms for an utterance in Japanese (/ikioi/) spoken by a physically unimpaired person and a person with an articulation disorder. Fig. 2 shows a comparison of the forced and manual alignment. The dysarthric speech signal is not obviously clearer than the signal uttered by a physically unimpaired person. As shown in Fig. 2, the phoneme boundaries obtained by forced alignment are not in the proper location (but it is difficult to obtain the correct alignment even if we do alignment manually). In this paper, we propose a soft phoneme labeling method with a probability expression, which takes the ambiguity of the phoneme boundary into account. In [8], the binary value was used as the training (phoneme) label, but the hard decision may be unfavorable for the training label because the boundary between the adjacent phonemes is very unclear. In our approach, the phoneme labeling is represented using a Gaussian distribution configured with the mean which represents the center of each phoneme period. The phoneme label is given from the posterior probability of a GMM. The rest of this paper is organized as follows: in Section 2, feature extraction using CNN is described. In Section 3, our phoneme labeling method is described. In Section 4, the experimental data are evaluated, and the final section is devoted to our conclusions.

2. Feature Extraction Using CBN

2.1. Flow of the Feature Extraction

First, we prepare the input feature for training a CBN from a speech signal. After calculating the short-term mel spectrum from the signal, a mel-map is obtained by merging the mel spectra into a 2D feature with several frames, allowing overlaps. For the output units of the CBN, phoneme labels that correspond to the input mel-map are used. The parameters of the CBN are trained by back-propagation with stochastic gradient descent, starting from random values. The input mel-map is converted to the bottleneck feature by using the CBN. Extracted features are used as the input feature of hidden Markov models (HMMs).

2.2. CBN

A CBN [17] consists of an input layer, a layer of convolution layer and pooling layer, fully-connected Multi-Layer Perceptrons (MLPs) with a bottleneck structure, and an output layer. The MLP stacks some layers, and the number of units in the middle layer is reduced as "bottleneck features". The number of units in each layer is discussed in the experimental section. Since the bottleneck layer has reduced the number of units for the adjacent layers, we can expect that each unit in the bottleneck layer aggregates information and behaves as a compact feature descriptor that represents an input with linear discriminant analysis (LDA) or PCA. In this paper, an audio feature is input to a CBN, and the extracted bottleneck feature is used for speech recognition.

3. Phoneme Labeling Based on Probabilistic Representation

For an arbitrary utterance included K phonemes, we note $X_t \in \{1,2,\dots,K\}$ the random variable that indicates a phoneme at time t. For example, $X_k = k$ indicates that a phone label at time t is the k-th phoneme in an utterance. In

this paper, the probability $p(X_k = k)$ is defined as follows:

$$p(X_t = k) = \frac{N(\mu_k, \sigma_k)}{\sum_{k'=1}^{K} N(\mu_{k'}, \sigma_{k'})} \quad (1)$$

where $N(\mu, \sigma^2)$ is the Gaussian probability density function with mean μ and variance σ^2. K and k are the number of phonemes included in the utterance and its index, respectively. In this paper, μ and σ^2 are the center of the k-th phoneme duration and its variance, and are defined as follows, respectively:

$$\mu_k = \frac{b_{k-1} + b_k}{2} \quad (2)$$

$$\sigma_k = \alpha(|\mu_k - b_{k-1}| + |\mu_k - b_k|) \quad (3)$$

where b_k is the boundary time between the k-th phoneme and $(k+1)$-th phoneme, and α is the non-negative hyperparameter that controls the variance ($b_k = 0$ and b_K are the start time and the end time of the utterance, respectively.) In our approach, we first give the phoneme boundary for each utterance. However, it is not the correct boundary but only an approximate boundary. Next, the phoneme duration and its variance are set up using the phoneme boundary based on (2) and (3). Finally, the existence probabilities for all frames are calculated by using (1), and the obtained probabilities are regarded as the phoneme labels. From the characteristics of the Gaussian distribution, the probability becomes high around the mean, but lower in the distance. Consequently, we expect that the soft phoneme labeling gives a good representation of not only the steady state around the center of the phoneme duration but also the unclear phoneme boundary.

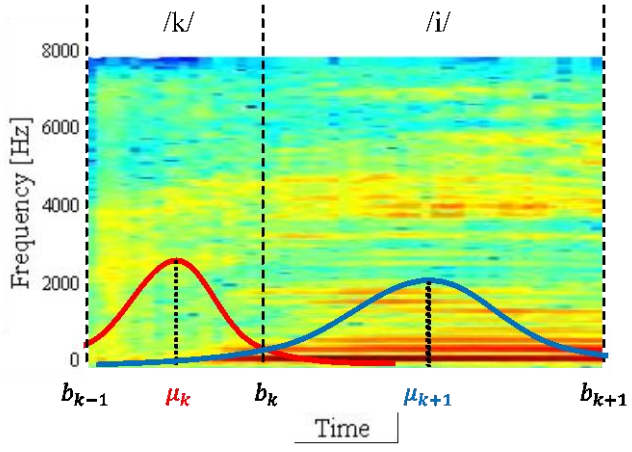

Figure 3. Illustration of the proposed Gaussian labeling for an utterance /ki/ in Japanese

4. Experimental Evaluation

4.1. Recognition Results Using a Speaker-independent Acoustic Model

At the beginning, we attempted to recognize utterances using a speaker-independent acoustic model for unimpaired people (This model is included in Julius 1). The acoustic model consists of a triphone HMM set with 25-dimensional MFCC features (12-order MFCCs, their delta and energy) and 16 mixture components for each state. Each HMM has three states and three self-loops. For a person with an articulation disorder, a recognition rate of only 24.07% was obtained, but for a physically-unimpaired person, a recognition rate of 99.54% was obtained for the same task. It is clear that the speaking style of a person with an articulation disorder differs considerably from that of a physically-unimpaired person. Therefore, it is considered that a speaker-dependent acoustic model is necessary for recognizing speech from a person with an articulation disorder.

4.2. Word Recognition Experiments (Speaker A)

4.2.1. Experimental Conditions

In this section, our method was evaluated on a word recognition task for one male person (referred to as "speaker A") with an articulation disorder. We recorded 216 words included in the ATR Japanese speech database A-set [18], repeating each word five times. The utterance signal was sampled at 16 kHz and windowed with a 25-msec Hamming window every 10 msec. Then we clipped each utterance manually. In our experiments, the first utterances of each word were used for evaluation, and the other utterances (the 2nd through 5th utterances) were used for the training of the CBN and acoustic models. A mel-map feature was constructed by merging mel spectra into a 2D feature with 13 frames. We used HMMs (54 context-independent phonemes) with 3 states and 8 Gaussian mixtures for the acoustic model. We trained and evaluated a CBN that has 30 units in the bottleneck (BN) layer.

4.2.2. Evaluation Results and Discussion

First, we investigated the effectiveness of a hyperparameter α in (3) which controls the variance. The best performance over the test data was obtained at $\alpha = 0.4$.

Fig. 4 depicts the difference between the hard labeling and soft labeling, and shows an example of the training label, the manual alignment, and the proposed alignment. In the proposed alignment, a phone label transitions to another one gradually. Fig. 5 shows the experimental results using the probabilistic (soft) phoneme labeling method ("Gaussian"), comparing with the conventional MFCC features and the forced alignment ("forced"). Fig. 5 also shows the results in the case of the manual alignment ("manual"). In this experiments, α is set to 0:4 in (3). The alignment obtained from our method provided a better recognition accuracy than both forced alignment and manual alignment. We consider that the network is trained flexibly by using the soft-labels.

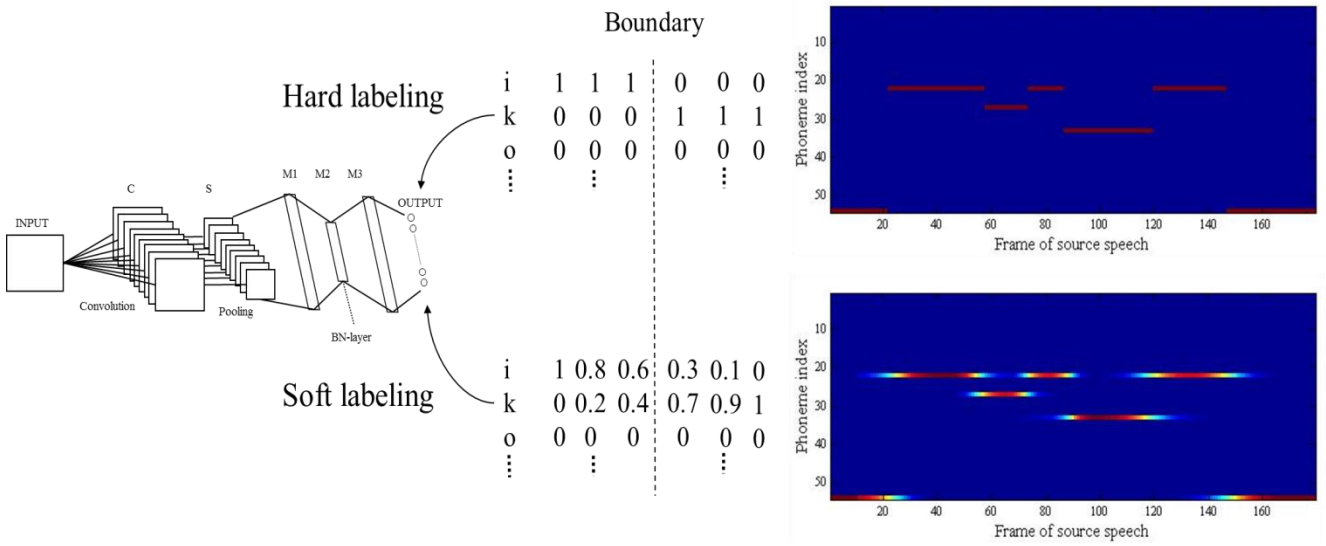

Figure 4. Analysis results of the label for an utterance /ikioi/ of "speaker A". (top) manual alignment, (bottom) proposed alignment α=0:4 (vertical and horizontal axes indicate the phoneme index and the frame of speech, respectively). The color lines indicate the values of teaching signal for CBN. Note: the 54-th phoneme index is a short pause

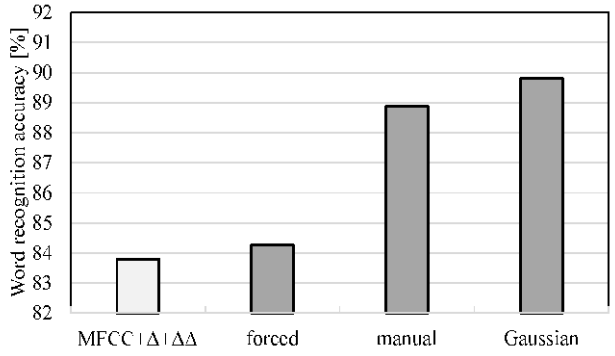

Figure 5. Word recognition accuracy for utterances of "speaker A" using each phoneme labeling method and conventional MFCC features

4.3. Word Recognition Experiments (speaker B)

We confirmed that the recognition accuracies improved by using a probabilistic phoneme labeling method in previous experiments. However, the symptoms of articulation disorders and the tendency of fluctuations in dysarthric speech vary from speaker to speaker. In this section, we show experimental results using speech uttered by another person (female; "speaker B") with an articulation disorder.

4.3.1. Experimental Conditions

We conducted the same word recognition experiments as in the previous experiments using speech uttered by "speaker B". The speech data consist of 200 words, each of which was repeated three times (600 words in total). In the experiments, the first utterances of each word (200 words) were used for the test, and the other utterances (400 words) were used for the training of a CBN and the acoustic models. The other configurations were set to be the same as the experiments with the "speaker A".

4.3.2. Experimental Conditions

Again we investigate the effectiveness of a hyperparameter α, and the best performance over the test data was obtained at $\alpha = 0.5$.

Fig. 6 shows experimental results for "speaker B". When using manual alignment, the recognition accuracies were not improved as compared with forced alignment. This is because the forced alignment was obtained accurately as compared with the manual alignment in the case of "speaker B". It might be too difficult to carry out alignment manually due to the unclear phoneme boundaries. Also, a few mistakes will have a bad influence on the performance because of the limited (small amount of training) data. Nevertheless, as shown in Fig. 6, the probabilistic labeling method could achieve the best accuracy for each condition.

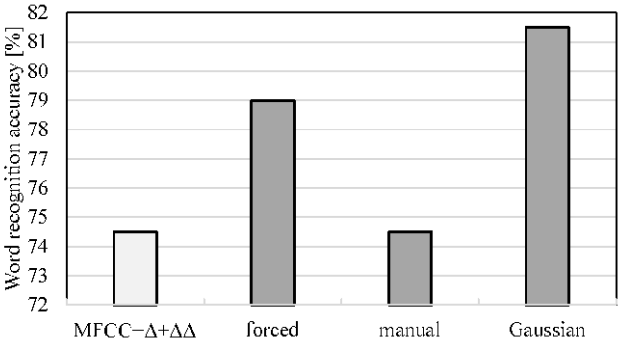

Figure 6. Word recognition accuracy for utterances of "speaker B" using each phoneme labeling method and conventional MFCC features

5. Conclusions

In this paper, we proposed a probabilistic (soft) phoneme labeling method, for persons with articulation disorders,

based on a Gaussian distribution for the training label that is used to train a CBN. In our recognition experiments, a CBN trained using probabilistic labels demonstrated better performance compared with the forced alignment using HMMs. In the future, we will study a labeling method that uses the forward backward probabilities from the HMMs.

REFERENCES

[1] Stephen Cox and Srinandan Dasmahapatra, "High-level approaches to confidence estimation in speech recognition," IEEE Transactions on Speech and Audio Processing, vol. 10, no. 7, pp. 460–471, 2002.

[2] Y. Takeuchi H. Kudo M. K. Bashar, T. Matsumoto and N. Ohnishi, "Unsupervised texture segmentation via wavelet-based locally orderless images (WLOIS) and SOM," CGIM, 2003.

[3] Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano, "Speaking aid system for total laryngectomees using voice conversion of body transmitted artificial speech," in INTERSPEECH, 2006.

[4] Hironori Matsumasa, Tetsuya Takiguchi, Yasuo Ariki, Ichao Li, and Toshitaka Nakabayashi, "Integration of metamodel and acoustic model for speech recognition," in INTERSPEECH, 2008, pp. 2234–2237.

[5] Omar Caballero Morales and Stephen J. Cox, "Modelling confusion matrices to improve speech recognition accuracy, with an application to dysarthric speech," in INTERSPEECH, 2007, pp. 1565–1568.

[6] Heidi Christensen, Phil D. Green, and Thomas Hain, "Learning speaker-specific pronunciations of disordered speech," in INTERSPEECH, 2013, pp. 1159–1163.

[7] Heidi Christensen, M. B. Aniol, Peter Bell, Phil D. Green, Thomas Hain, Simon King, and Pawel Swietojanski, "Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech," in INTERSPEECH, 2013, pp. 3642–3645.

[8] Toru Nakashika, Tetsuya Takiguchi, Yasuo Ariki, S. Duffner, and C. Garcia, "Dysarthric speech recognition using a convolutive bottleneck network," in ICSP, 2014, pp. 505–509.

[9] Yann LeCun and Yoshua Bengio, "The handbook of brain theory and neural networks," chapter Convolutional Networks for Images, Speech, and Time Series, pp. 255–258. MIT Press, 1998.

[10] Yann LeCun, L ́eon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[11] Honglak Lee, Peter Pham, Yan Largman, and Andrew Y Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in Advances in neural information processing systems, 2009, pp. 1096–1104.

[12] Christophe Garcia and Manolis Delakis, "Convolutional face finder: A neural architecture for fast and robust face detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 11, pp. 1408–1423, 2004.

[13] Manolis Delakis and Christophe Garcia, "Text detection with convolutional neural networks," in VISAPP (2), 2008, pp. 290–294.

[14] Raia Hadsell, Pierre Sermanet, Jan Ben, Ayse Erkan, Marco Scoffier, Koray Kavukcuoglu, Urs Muller, and Yann LeCun, "Learning long-range vision for autonomous off-road driving," Journal of Field Robotics, vol. 26, no. 2, pp. 120–144, 2009.

[15] Gr ́egoire Montavon, "Deep learning for spoken language identification," in NIPS Workshop on deep learning for speech recognition and related applications, 2009.

[16] Toru Nakashika, Christophe Garcia, and Tetsuya Takiguchi, "Local-feature-map integration using convolutional neural networks for music genre classification," in INTERSPEECH, 2012.

[17] Karel Vesel`y, Martin Karafi ́at, and Franti ̌sek Gr ́ezl, "Convolutive bottleneck network features for LVCSR," in ASRU, 2011, pp. 42–47.

[18] Akira Kurematsu, Kazuya Takeda, Yoshinori Sagisaka, Shigeru Katagiri, Hisao Kuwabara, and Kiyohiro Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," Speech Communication, vol. 9, no. 4, pp. 357–363, 1990.