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

Constructing Acoustic Distances between Subwords and States Obtained from a Deep Neural Network for Spoken Term Detection

Daisuke Kaneko [1], *Ryota Konno* [1], *Kazunori Kojima* [1], *Kazuyo Tanaka* [2], *Shi-wook Lee* [3], *Yoshiaki Itoh* [1]

[1] Iwate Prefectural University, Japan
[2] University of Tsukuba, Japan
[3] National Institute of Advanced Industrial Science and Technology, Japan
y-itoh@iwate-pu.ac.jp

Abstract

The detection of out-of-vocabulary (OOV) query terms is a crucial problem in spoken term detection (STD), because OOV query terms are likely. To enable search of OOV query terms in STD systems, a query subword sequence is compared with subword sequences generated using an automatic speech recognizer against spoken documents. When comparing two subword sequences, the edit distance is a typical distance between any two subwords. We previously proposed an acoustic distance defined from statistics between states of the hidden Markov model (HMM) and showed its effectiveness in STD [4]. This paper proposes an acoustic distance between subwords and HMM states where the posterior probabilities output by a deep neural network are used to improve the STD accuracy for OOV query terms. Experiments are conducted to evaluate the performance of the proposed method, using the open test collections for the "Spoken&Doc" tasks of the NTCIR-9 [13] and NTCIR-10 [14] workshops. The proposed method shows improvements in mean average precision.

Index Terms: spoken-term detection, deep neural network, out-of-vocabulary query term, acoustic distance

1. Introduction

Research on spoken-document retrieval (SDR) and spoken-term detection (STD) have been actively conducted to realize efficient searching of vast quantities of audiovisual data [1,2,3]. STD is the task of finding matched sections in spoken documents with a query consisting of one or more words. Query terms are often out-of-vocabulary (OOV) words not contained in an ASR dictionary, such as technical terms, geographical names, personal names, and neologisms. Therefore, OOV query terms must be retrievable through STD systems. To enable that, subword recognition using monophones, triphones, and so on is performed in advance for all spoken documents, and subword sequences for spoken documents are prepared beforehand. When query terms are given to the system, it converts the query terms to a sequence of subwords, and then searches for that sequence among the documents subword sequences. Matching between a query sequence and those of spoken documents is conducted by a continuous dynamic programming (CDP) algorithm that continuously applies the dynamic time warping (DTW) algorithm between a query term and spoken documents at the subword level [4]. The acoustic distances are used as the local distance between any two subwords during CDP to enable approximate matching.

In recent years, many researchers have reported significant improvements in the accuracy of speech recognition by using a deep neural network (DNN) [6,7]. We previously proposed acoustic distances defined from the statistics between states of the hidden Markov model (HMM) and showed their effectiveness in STD [4]. This paper proposes an introduction of posterior probabilities output by DNN in constructing acoustic distances.

The proposed acoustic distances are a kind of confusion matrix. Two types of acoustic distance are proposed, the acoustic distance between subwords and that between states composing a subword HMM. Training data are first segmented and corresponded to subwords and states by forced alignment. When applying the training data of a state A to DNN, the output probability of a state B for data A corresponds to the possibility of a misrecognition from A to B. In this way, a confusion matrix for states is constructed. A confusion matrix for subwords is constructed by averaging the three distances of the three states that compose each subword HMM. Although spoken documents are converted to word transcriptions or subword transcriptions by ASR using DNN [8, 9], there are few reports of using DNN to construct acoustic distances between subwords or states.

Section 2 describes the conventional STD system and acoustic distance. Section 3 describes the proposed acoustic distances in detail. The "Spoken&Doc" open test collections in the NTCIR-9 and NTCIR-10 workshops are used to evaluate the retrieval accuracy of the proposed method in Section 4, and our conclusions are presented in Section 5.

2. Conventional STD System and Acoustic Distances for OOV Terms

For OOV query terms, spoken documents are transformed to subword sequences by subword recognition using ASR beforehand in our conventional STD system. Given a query, a query subword sequence is automatically obtained according to Japanese conversion rules. A CDP algorithm continuously applies DTW to search for a query subword sequence among subword sequences of spoken documents. CDP performs matching between a query subword sequence and subword sequences of spoken documents. Acoustic

http://dx.doi.org/10.21437/Interspeech.2017-634

distances between subwords are used as a local distance in CDP. These are obtained from the statistics of HMMs composed of subwords, as described in detail in the next section. We have previously demonstrated the effectiveness of these acoustic distances [4].

2.1. Acoustic Distance between Subwords/States

This section explains in detail the method for constructing acoustic distances between subwords and states [4]. Let a subword HMM be composed of three states, here. First, the distance between the same *i*-th state $s(1 \le i \le 3)$ of two subwords is defined. The likelihood of each state is computed using a Gaussian mixture model (GMM) in the state. A state distribution is extracted from the two states. The nearest two distributions correspond to the distance between the two states, and the nearest distributions are obtained by computing all the distances between any two distributions derived from each pair of states. The distance of the nearest distributions is regarded as the distance between states. The formulation of the sate distance is described below.

u and *v* are distributions of two states, μ_{ud} and σ_{ud}^2 are the mean and variance of dimension *d* of distribution *u*, respectively. The distance between two distributions is obtained by Eq. (1), which denotes the Bhattacharyya distance $BD(u,v)$ between *u* and *v*. Here, *Dim* is the dimensions of a feature vector.

$$BD(u,v) = \frac{1}{4}\sum_{d=1}^{Dim}\left\{\frac{(\mu_{ud}-\mu_{vd})^2}{\sigma_{ud}^2+\sigma_{vd}^2} + \log\frac{(\sigma_{ud}^2+\sigma_{vd}^2)^2}{4\sigma_{ud}^2\sigma_{vd}^2}\right\} \quad (1)$$

Let $s_p(i)$ and $m_p(i,j)$ be the *i*-th state of subword *p* and the *j*-th distribution in the *i*-th state of subword *p*. As shown in Eq. (2), the minimum Bhattacharyya distance is obtained by computing all distances between any distributions (*j, k*) in the *i*-th states. The minimum distribution distance is regarded as the *i*-th state distance $SD(s_p(i), s_q(i))$.

$$SD(s_p(i), s_q(i)) = \min_{j,k} BD(m_p(i,j), m_q(i,k)) \quad (2)$$

All state distances are computed for all state pairs, and an acoustic-distance matrix between states is constructed.

The acoustic distance *AD* (*p, q*) between subwords *p* and *q* is defined as the average of the *M* state distances, as shown in Eq. (3).

$$AD(p,q) = \frac{1}{M}\sum_{i=1}^{M} SD(s_p(i), s_q(j)) \quad (3)$$

All acoustic distances are computed for all subword pairs, and an acoustic-distance matrix between subwords is constructed.

2.2. Confusion Matrix

A confusion matrix is often used to represent the acoustic distance between two subwords. The confusion matrix corresponds to the error tendency from a given subword or state to another subword or state. A large amount of a speech corpus different from the evaluation data is used to construct the confusion matrix. First, an ASR transforms the speech corpus into a subword sequence. From the results, the number of occurrences of each subword and the number of mistakes too different subwords are counted in the other subword, and the probability that each subword is mistakenly replaced with another subword is computed. For example, N_A is the number of occurrences of subword *A*, and N_A^B is the number of mistaken replacements of *A* with *B*. The probability of replacing *A* with *B* is obtained by Eq. (4).

$$P_A(B) = \frac{N_A^B}{N_A} \quad (4)$$

Generally speaking, many training data are required to calculate the exact distance to obtain the correct confusion matrix. By using the confusion matrix, it is possible to calculate the likelihood of error between each sound, and the distance can be derived from the acoustic point of view from the value.

3. Proposed Method

3.1. DNNs

A DNN is a multilayer neural network. A typical speech recognizer is a combination of an HMM and a DNN (DNN–HMM). DNN input is a feature vector of spoken data such as mel-frequency cepstral coefficients (MFCCs) and log filter-bank parameter vectors. DNN outputs are posterior probabilities of each HMM state. A feature vector includes feature vectors of several frames before and after the current frame, making it a high-dimensional feature vector. Each output node in an output layer is associated with a HMM state. When a feature vector of one frame is given to the DNN, each output node generates a probability, which becomes the posterior probability of the associated HMM state.

3.2. Constructing Acoustic Distance and Confusion Matrix Using DNN Probabilities

Each DNN output node outputs a posterior probability for each state of a subword. The acoustic distance between states is defined using this posterior probability, and a confusion matrix between states is constructed. We explain the process along with the example in Fig. 1.

Note the section of phone "k" in phone sequence "i k a" in the training data. Triphone "i-k+a" is composed of three states and the corresponding section for each state is obtained by forced alignment, as shown in Fig. 1. For example, a feature vector sequence (shaded areas) corresponding to the first state S_a of triphone "i-k+a" is input to DNN. If the posterior probability of the first state S_d of triphone "i-s+a" is as large as that of state S_a, state S_d can be regarded as similar to state S_a. An acoustic distance between states is accordingly defined as described below.

3.2.1. Defining Acoustic Distance between States

All sections corresponding to the *i*-th state S_{pi} of subword *p* in the training data are extracted by forced alignment, and feature vectors of the corresponding sections are input to

DNN. The output probabilities for other states are obtained for each frame, because an average output probability for state S_{qj} against all the frames corresponding to state S_{pi} represents the similarity between S_{qj} and S_{pi}.

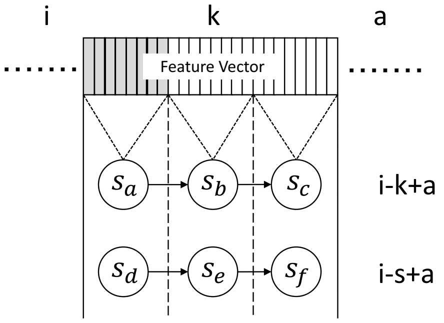

Figure 1: *Forced alignment of training data*

The corresponding sections in the training data for state are extracted by forced alignment. Let the *n*-th frame among all corresponding sections be $F_s(n)$, N_s be the total number of frames corresponding to state $s(1 \leq i \leq N_s)$, and $P_t(F_s(n))$ be the posterior probability of state *t* at the frame $F_s(n)$.

The average of the posterior probability $P_t(F_s(n))$, obtained in the brackets in Eq. (5), represents a state similarity between states *s* and *t*. Because a local distance is required in CDP, Eq. (5) transforms the state similarity to a state distance *SD* (*s, t*) between state *s* and *t*.

$$SD(s,t) = -\log\left\{\frac{1}{N_s}\sum_{n=1}^{Ns} P_t(F_s(n))\right\} \tag{5}$$

All state distances are computed for any two states, and an acoustic-distance matrix between states is constructed.

3.2.2. Construction of Acoustic Distance between Subwords

In a similar manner as Eq. (3), Eq. (6) defines the acoustic distance *AD* (*p, q*) between subword *p* and *q*, where the state distances described in the previous section are averaged over *M* successive states in two subword HMMs.

$$AD(p,q) = \frac{1}{M}\sum_{i=1}^{M} SD_i(s,t) \tag{6}$$

All acoustic distances are computed for any two subwords, and an acoustic-distance matrix between subwords is constructed.

4. Evaluation Experiments

4.1. Experimental Conditions

For acoustic and language model training, we used 1255 speeches (about 287 hours of audio, so about 14 minutes per speech) included in the Corpus of Spontaneous Japanese (CSJ) [10,11]. 177 presentation speeches in the CSJ were excluded and used as testing data. We used a triphone acoustic model composed of a left-to-right HMM with three states, and we used tied-state triphone models with 3009 states and 32 mixtures per state. The input feature vectors were extracted under the conditions shown in Table 1. Five frame-feature vectors were added before and after the current frame as a feature vector for the DNN. The DNN was trained using a 418-dimensional feature vector under the conditions shown in Table 2. The alignments between speech signals and each state were obtained from the results of forced alignment. Syllable trigrams were used for language models. For syllable recognition, we used Julius [12,13], a large vocabulary continuous speech recognition engine. All spoken documents are transformed into triphone sequences after syllable recognition using a DNN-HMM-based Julius system.

To measure processing time, we used a personal computer with an Intel Core i7-4770 CPU, NVIDIA GeForce GTX TITAN GPU, and 16 GB of memory.

4.2. Test Collections

To evaluate the STD performance, we used the two open test collections that were used in the NTCIR-9 workshop [13] and the NTCIR-10 workshop [14]. As shown in Table 3, the test collection of NTCIR-9 contains 44 hours of CSJ presentation speeches (excluded for training data) and two query sets (dry run and formal run). Each query set includes 50 query terms. The test collection of NTCIR-10 contains 29 hours of spoken documents, distinct from CSJ, and two query sets. Four kinds of test collection were thus used in this study. We regarded all query as OOV queries. We used the mean average precision as the measure of STD accuracy.

Table 1: *Conditions for feature extraction*

Feature parameter	38 dimensions (dim) MFCC (12 dim) + Delta-MFCC (12 dim) + Delta-Delta-MFCC (12 dim)
Window length	25 ms
Frame shift	10 ms

Table 2: *Conditions for DNN*

Number of nodes	Input layer: 418 Hidden layer: 2048 Output layer: 3009
Number of hidden layers	3 layers

Table 3: *Two open test collections*

	NTCIR-9	NTCIR-10
Spoken documents	CSJ, 177 presentations, 44 hours, 53,892 utterances	SDPWS, 104 presentations, 29 hours, 40,746 utterances
Query sets and number	Formal run: 50 Dry run: 50	Formal run: 100 Dry run: 32

of queries		

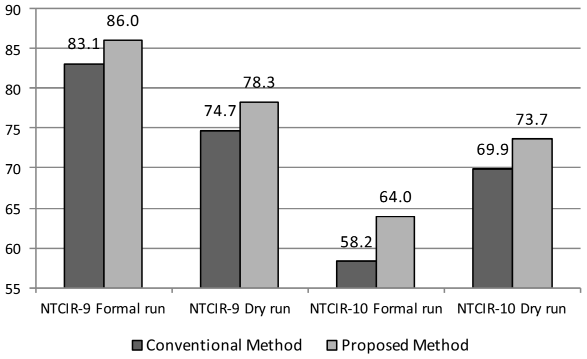

Figure 2: Retrieval accuracy using subword-level matching

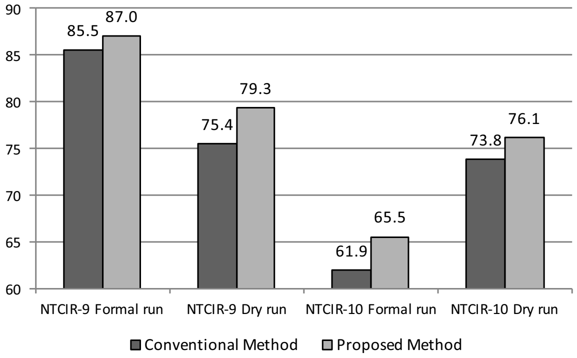

Figure 3: Retrieval accuracy using state-level matching

4.3. Results

We conducted experiments with four types of NTCIR-9 and NTCIR-10 formal- and dry-run test collection, shown in Table 3. The experimental results for the four test collections are shown in Figs. 2 and 3. These figures denote the retrieval result by subword-level matching and the retrieval result by state-level matching. In both figures, the “conventional method” used an acoustic distance constructed using the GMM described in section 2.1, and the “proposed method” used an acoustic distance constructed using the DNN described in section 3.

In subword-level matching (Fig. 2), the proposed acoustic distances improved STD accuracy in all test sets by between 2.91 and 5.71 points. In state-level matching (Fig. 3), the proposed acoustic distances improved STD accuracy in all test sets by between 1.52 and 3.90 points. These results demonstrate the effectiveness of the proposed acoustic distances obtained using DNN.

When comparing the retrieval results of subword-level matching with those of state-level matching, STD accuracy using state-level matching was slightly better than that using subword-level matching for all test sets. This is because state-level matching performs a more detailed search in a DP lattice. In this case, both the reference and the input become increasingly detailed three times. As a result, the retrieval time when using state-level matching was 0.30 seconds, and that when using subword-level matching was 0.04 seconds.

Acoustic distance between subwords and states were constructed using the CSJ as learning data. Because the retrieval target data of NTCIR-9 is the rest part other than the training data of CSJ, the effect can be expected. In fact, it showed an average of +3.26 points of STD accuracy improvement in subword-level matching and an average of +2.71 points of STD accuracy improvement in state-level matching. On the other hand, the retrieval target data of NTCIR-10 is SDPWS, which is spoken data from a different recording environment. In this test set, STD accuracy improved by an average of +4.75 points in subword-level matching and an average of +2.96 points in state-level matching. These results show the robustness of the acoustic distance constructed by the proposed method.

When spoken documents are transformed to triphone sequences using a GMM-HMM recognizer, not a DNN-HMM one, the accuracy improvements shown in Figs. 2 and 3 were not obtained, and STD accuracy degraded for some test sets. When using a DNN-HMM recognizer for spoken documents, the error tendency of each subword is similar to the acoustic distance matrix and substitution errors can be recovered by using the acoustic-distance matrix. In contrast, when using a GMM-HMM recognizer for spoken documents, the error tendency did not accord with the acoustic-distance matrix.

5. Conclusion

We proposed a method to define acoustic distance between subwords and states obtained using the posterior probability of a DNN. The experimental results were applied to four open test sets of NTCIR-9 and 10. The STD accuracy in terms of mean average precision was improved +5.71 points at maximum and +4.01 points on average when using subword-level matching, and +3.90 points at maximum and +2.84 points on average when using state-level matching. These results demonstrate the effectiveness of the proposed acoustic distances.

In the future, we will seek a method for obtaining theoretical distances, such as the Bhattacharyya distance [17] using a DNN, and compare the results with those using the confusion matrix proposed here.

6. Acknowledgements

This work was supported by JSPS (C), KAKENHI, Grant Number 15K00241.

7. References

[1] C. Auzanne, JS. Garofolo, JG. Fiscus, and WM Fisher, “Automatic Language Model Adaptation for Spoken Document Retrieval,” 2000TREC-9 SDR Track, 2000.

[2] A. Fujii, and K. itou, “Evaluating Speech-Driven IR in the NTCIR-3Web Retrieval Task,” Third NTCIR Workshop, 2003.

[3] P. Motlicek, F. Valente, and PN. Garner, “English Spoken Term Detection in Multilingual Recordings", INTERSPEECH 2010, pp.206-209, 2010.

[4] K. Iwata, Y. Itoh, K. Kojima, M. Ishigame, K. Tanaka and S. Lee, “Open-Vocabulary Spoken Document Retrieval based on new subword models and subword phonetic similarity,” INTERSPEECH, 2006.

[5] G. Hinton, S. Osindero, and Y. Teh, “A fast learning algorithm for deep belief nets,” Neural Computation, vol. 18, pp. 1527–1554, 2006.

[6] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, “Deep Neural Networks for Acoustic Modeling in Speech Recognition,” IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82–97, 2012.

[7] L. Mangu, H. Soltau, H.K. Kuo, B. Kingsbury, and G. Saon, “Exploiting diversity for spoken term detection,” in Proc. ICASSP, 2013.

[8] B. Kingsbury, J. Cui, X. Cui, M. J. F. Gales, K. Knill, J. Mamou, L. Mangu, D. Nolden, M. Picheny, B. Ramabhadran, R. Schluter, A. Sethy, and P. C. Woodland, “A high performance Cantonese keyword search system,” in Proc. ICASSP, 2013.

[9] K. Maekawa, “Corpus of Spontaneous Japanese: Its design and evaluation,” Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003), 2003.

[10] National Institute for Japanese Language and Linguistics, Corpus of Spontaneous Japanese, http://www.ninjal.ac.jp/corpus_center/csj/

[11] A. Lee and T. Kawahara, “Recent Development of Open-Source Speech Recognition Engine Julius,” Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2009.

[12] Open-Source Large Vocabulary CSR Engine Julius, http://julius.osdn.jp/

[13] T. Akiba, H. Nishizaki, K. Aikawa, T. Kawahara and T. Matsui, “Overview of the IR for Spoken Documents Task in NTCIR-9 Workshop,” Proc. of NTCIR-9 Workshop Meeting, pp. 223-235, 2011.

[14] T. Akiba, H. Nishizaki, K. Aikawa, X. Hu, Y. Itoh, T. Kawahara, S. Nakagawa, H. Nanjo, Y. Yamashita, “Overview of the NTCIR-10 SpokenDoc-2 Task,” Proc. of the 10th NTCIR Conference, pp. 573-587, 2013.

[15] H. Nishizaki, H. Furuya, S. Natori, and Y. Sekiguchi, “Spoken term detection using multiple speech recognizers' outputs at NTCIR-9 SpokenDoc STD subtask,” Proc. of NTCIR-9 Workshop Meeting, 2011.

[16] K. Kon'no, H. Saito, S. Narumi, K. Sugawara, K. Kamata, M. Kon'no, J. Takahashi and Y. Itoh, “An STD System for OOV Query Terms Integrating Multiple STD Results of Various Subword units, Proceedings of the 10th NTCIR Conference,” 2013.

[17] Y. Kashiwagi, C. Zhang, D. Saito, N. Minematsu, “Divergence estimation based on deep neural networks and its use for language identification,” to appear at ICASSP, 2016.

[18] N. Mareau, H.G. Kim and T. Sikara, “Phonetic Confusion Based Document Expansion for Spoken Document Retrieval,” INTERSPEECH, 2004.

[19] P. Zhang, J. Shao, J. Han, Z. Liu, Y. Yan, “Keyword Spotting Based on Phoneme Confusion Matrix,” ISCSLP, 2006.

[20] O.C. Marales and S. Cox, “Modelling Confusion Matrices to Improve Speech Recognition Accuracy, with an Application to Dysarthric Speech,” INTERSPEECH, 2007.

[21] D. Xu, Y.Wang, F. Metze, “EM-based Phoneme Confusion Matrix Generation for Low-resource Spoken Term Detection,” SLT, 2014.