Interspeech 2018
2-6 September 2018, Hyderabad

Empirical analysis of score fusion application to combined neural networks for open vocabulary spoken term detection

Shi-wook Lee[1], *Kazuyo Tanaka*[2], *Yoshiaki Itoh*[3]

[1]National Institute of Advanced Industrial Science and Technology, Japan
[2]University of Tsukuba, Japan
[3]Iwate Prefectural University, Japan

s.lee@aist.go.jp, tanaka.kazuyo.gb@u.tsukuba.ac.jp, y-itoh@iwate-pu.ac.jp

Abstract

System combination, which combines the outputs of multiple systems or internal representations, is a powerful method to improve the performance of machine learning tasks and has been widely adopted in recent knowledge transfer learning. In this study, to describe how to extract effective knowledge from an ensemble of neural networks, we first examine several score fusions from an ensemble of neural networks tasked with open vocabulary spoken term detection, where the class probability of the neural network is utilized as a similarity metric; then, we investigate the trade-off between confusion and dark knowledge. From the experimental evaluation on open vocabulary spoken term detection, we obtain 2.09% absolute gain as compared to the best result from single systems. Furthermore, the performance gains achieved via score fusion of class probabilities exactly match the mathematical inequality for sum and power means results, and that the gain achieved via summation of class probabilities is consistently better than that achieved via score fusion of power means. The experimental analysis confirms that summation, which enhances the discriminative capability of the superior class probability, can implement smoothed probability distribution to yield more effective dark knowledge, while adequately suppressing undesirable effects.

Index Terms: system combination, score fusion, dark knowledge, similarity metric, open vocabulary spoken term detection

1. Introduction

System combination is an effective method to improve the performance of machine learning tasks by combining their outputs from multiple systems or their internal representations. Consequently, in automatic speech recognition (ASR) and spoken term detection (STD), which is an ASR application, it has also been reported that performance improvements could be achieved via system combination, which combines the outputs of multiple systems with diverse ASR components, such as an acoustic model, a decoding framework, and audio segmentation [1–5]. Furthermore, with the increased implementation of deep learning, system combination has been more aggressively adopted in ASR and its various applications; significantly diverse models can be generated by setting different initial parameters and neural network structures, e.g., deep neural networks (DNNs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs) [6, 7]. [6] tested linear and log-linear stacking methods for learning ensemble parameters following learning of the weight parameters of low-level DNN, CNN, and RNN systems. [7] proposed a joint model that is trained via a DNN and CNN, i.e., two different types of neural networks. Such improvements in system combination may result from the complementary information, which preserves the relative confidence of different outputs provided by multiple systems [8–10]. This complementary information is referred to as "dark knowledge" in [9].

More recently, there has been growing interest in knowledge transfer learning [10–16], which is also known as teacher-student learning, is motivated by model compression [17], and is strongly related to dark knowledge [9]. The main objective of knowledge transfer learning is to train a small model (student) by transferring knowledge from the outputs of another large model (teacher), which is a single complex model or an ensemble of models. Thus, large complex ensembles can be compressed into a single smaller and faster model, usually with less performance degradation. The demand for a single, smaller model is increasing because high-accuracy DNNs are massive and computationally expensive.

[11] trained a small-size DNN from a standard large-size DNN by utilizing a large amount of untranscribed data. [10] employed temperature to extract a relatively larger amount of dark knowledge from multiple systems. [12] reported on the transfer learning between heterogeneous models, where the knowledge of an accurate RNN is transferred to a small DNN. In contrast, [13] detailed interesting research, demonstrating that knowledge learned by simple models (DNN) can be effectively used to guide the training of complex models (RNN) where the teacher DNN is assumed weaker than the student RNN. [14] linearly interpolated the class probabilities and proposed choosing weights of the oracle to make a soft label. [15] transferred the knowledge from ensembles of multilingual models to the model of low-resource language. [16] presented two strategies to leverage multiple teacher labels, switching teacher labels at the mini-batch level and multiple teacher distribution by data augmentation. In order to make a smaller student model close to or outperform the performance of teacher model, the aim of early approaches was how to distill richer knowledge from the teacher model, where the approach of adopting temperature in [10] was a representative method. More recently, the trend has been to create various targets that comprise 1) an original hard label, 2) a forced alignment hard label, which is the output class from the teacher model, and 3) a soft label, which is the class probability distribution of the teacher model. Then, the enlarged targets are linearly combined or used in the framework of data augmentation learning, where transfer learning is based on the loss function, which is typically a linear combination of Kullback-Leibler divergence for the soft label with probability distribution, and cross-entropy for the hard label.

Motivated by these recent works, we examine a direct way to use the dark knowledge from multiple neural networks in open vocabulary STD tasks in which several score fusions of

10.21437/Interspeech.2018-1776

the class probabilities are used as a similarity metric. When the STD task is performed via acoustic-level dynamic time warping (DTW) matching, the class probability-based similarity metric does not require a lexicon or language model; thus, the dark knowledge, which is quantified by score fusion, can be independently evaluated.

The remainder of this paper is organized as follows: Section 2 describes how the class probability can be used as a similarity metric for an open vocabulary STD task that is based on acoustic-level DTW. Section 3 describes score fusion methods that are based on sum or power means on logit and class probability. Section 4 presents the results of experimental evaluations. Finally, Section 5 concludes this paper.

2. Class probability as a similarity metric

The open vocabulary STD task to locate all occurrences of a specified word/phrase in the search audio database [18, 19] is usually implemented by using subword-based detection [20] or acoustic-level DTW matching [21, 22]. Most current ASR tasks typically comprise multiple states of a subword, with the subword being a phoneme or its variant. The state is also identical to the class, which is the output of the final softmax layer in the DNN. In acoustic-level DTW matching for open vocabulary STD, a specified term is transformed into a sequence consisting of Q states $S = \{s^1, \cdots, s^q, \cdots, s^Q\}$, and the class (posterior) probability of $p(s^q|o_t)$ is calculated for the t-th frame of the target spoken database, in which $O = \{o_1, \cdots, o_t, \cdots, o_T\}$ denotes the sequence of acoustic observation vectors. For the observation vector at frame t, each class probability of $p(s^q|o_t)$ is calculated and then transformed into the negative logarithm below. The acoustic-level DTW to find optimal path for the state sequence is implemented on the following:

$$\begin{aligned} &-\log(p(s^{1:q}|o_{b:t})) \\ &= \min \left\{ \begin{array}{l} -\log(p(s^{1:q}|o_{b:t-1})) - \log(p(s^q|o_t)), \\ -\log(p(s^{1:q-1}|o_{b:t-1})) - \log(p(s^q|o_t)), \\ -\log(p(s^{1:q-1}|o_{b:t})) - \log(p(s^q|o_t)) \end{array} \right\} \end{aligned} \quad (1)$$

Here, $s^{1:q}$ refers to the partly matched states from s^1 to s^q of the specified term, and $o_{b:t}$ refers to the matched observations from o_b of the possible start frame b to o_t of the current t-th frame. Thus, the cumulative class probability of $p(s^{1:q}|o_{b:t})$ indicates the putatively optimal path up to the q-th state and t-th frame. The cumulative class probability of (1) is normalized at the last state s^Q by the frame number of the detected interval; this normalized value is used as the similarity metric in a ranked list.

$$\begin{aligned} &\langle Q^*, b^*, e^* \rangle \\ &= \arg\min_{\{Q,b,e\}} \left\{ \frac{-1}{e-b+1} \log(p(s^{1:Q}|o_{b:e})) \right\} \end{aligned} \quad (2)$$

As the value of (2) approaches zero, it becomes increasingly likely that the path of the observation sequence $o_{b:e}$ along the state sequence $s^{1:Q}$ includes the uttered interval of the specified term. When the value is less than a predefined threshold value, the path is regarded as a putative hit. The current STD (also known as keyword search) typically implements large vocabulary continuous speech recognition (LVCSR) in addition to some form of lattice post-processing, to generate an expeditious searchable index with reduced accuracy loss. As compared to the currently trending methods, the acoustic-level DTW method is expensive and time consuming. However, to further improve performance of open vocabulary tasks that require the system to detect an out-of-vocabulary term, the acoustic-level DTW method can be adopted to re-score the top results of the compact and high-speed system.

3. Score fusion methods

Recent studies showed that a large ensemble of models can be transformed into a single small model [10–17]. Using the knowledge distillation research in [10] as a reference, we examine the enrichment of class probability as a similarity metric via fusion of multiple neural networks, where the enriched class probability may be considered to represent dark knowledge. As compared to linear interpolation, power means implementation is a less empirical and more mathematically analytical method to investigate the effectiveness of score fusion for comparably strong models.

3.1. Fusion via sum and power means

In order to investigate the effectiveness of dark knowledge, we implement the power mean of class probability. Class probabilities of N systems, $(p_1, \cdots, p_N) \in [0, 1]$, are fundamentally positive real numbers; thus, the generalized power mean with exponent k of the class probabilities is

$$M_k(p_1, \cdots, p_N) = \left(\frac{1}{N} \sum_{n=1}^{N} p_n^k \right)^{\frac{1}{k}} \quad (3)$$

where the exponent k is real and non-zero. Note that M_1 is the arithmetic mean (AM) of N class probabilities, M_2 is the quadratic mean (QM), M_{-1} is the harmonic mean (HM), $\lim_{k\to 0} M_k$ is the geometric mean (GM), and $\lim_{k\to\infty} M_k$ is the maximum (MAX). These power means satisfy the inequality $M_l \geq M_m$ for all $l > m$, with equality if and only if $p_1 = \cdots = p_N$. In addition to the power means, we calculate the sum (SUM) of class probabilities, as follows:

$$\text{SUM} = N \cdot M_1(p_1, \cdots, p_N) = \sum_{n=1}^{N} p_n \quad (4)$$

Between SUM and MAX, the inequality of SUM $\geq$ MAX is also satisfied, with equality if and only if all $p_n = 0$ for $n = 1, \cdots, j-1, j+1, \cdots, N$ and $p_j \in [0, 1]$. Finally, we set the comparison experiments on the order of the inequality among the sum and power means, SUM $\geq$ MAX $\geq$ QM $\geq$ AM $\geq$ GM $\geq$ HM. Usually, the ensemble of multiple neural networks for a soft label is calculated by taking the AM as the average [10] or linearly interpolated from class probabilities [14]. Centered on the unbiased AM, SUM, MAX, and QM are biased to a higher probability, and GM and HM are biased to a lower probability. In other words, according to the order of the inequality, nearly all classes in which the probability is close to zero will become relatively much smaller, and some classes with higher probability can be far superior. Here, we set the following hypothesis: dark knowledge will become more effective with more aggressive, discriminative fusion.

3.2. Fusion on logit or class probability

For ASR as a multi-class classification task, each output neuron of a neural network represents a class $s \in R^C$, where C is the number of classes. The following softmax function is used to calculate class probability $p(s_c|o_t)$ from logits z_c, which is excitation for the c-th class.

$$p(s_c|o_t) = \frac{\exp(z_c/T)}{\sum_{i=1}^{C} \exp(z_i/T)} \tag{5}$$

where T is the temperature to control dark knowledge and is normally set to 1 [10]. A higher value of T produces a softer probability distribution over classes. Here, it is necessary to consider the trade-off between confusion and dark knowledge.

We examine score fusion on the levels of logit and class probability. By first performing score fusion on a logit level and then calculating the class probability via the softmax function of (5), the inequality resulting from a sum and power means method is eliminated as the logit z range is $(-\infty, +\infty)$. Conversely, fusion on the class probability maintains the inequality as the class probability range is $[0, 1]$. Additionally, when fusion is applied to the logit before applying the softmax function, the class probability distribution can be controlled by temperature. Here, in the case of applying fusion as the sum on the class probability, the sum, which is $f_{\text{SUM}} : [0, 1] \to [0, N]$, violates the probabilistic constraint. In other words, the domain of the sum extends linearly in the positive direction. However, the sum may be the bias of the AM as shown in (4) and provides a comparable performance to the AM; this has been confirmed in our preliminary experiments although it has not been discussed in this paper. We modified the summation to have the upper value of the sum fixed at 1 to satisfy the probabilistic constraint $\sum_{n=1}^{N} p_n(s_c|o_t) \in [0, 1]$ and still preserve the inequality. The modified summation function $f_{\text{SUM}'} : [0, 1] \to [0, 1]$ is defined as follows:

$$\text{SUM}' \overset{\text{def}}{=} \min\Big\{ \sum_{n=1}^{N} p_n, 1 \Big\} \tag{6}$$

4. Experimental evaluation

4.1. Spoken term detection task

In this section, the results of the experiments conducted on NTCIR10 STD task data [23] are presented. The data comprised 28.6 h of 40,746 utterances for the target speech database, along with 100 queries and their correct utterance spans. In the experiments, we implemented 40-dimensional log-mel filterbank features (FBANK) on 25-ms windows computed every 10 ms. The first and second derivatives of a total of 120 dimensions were extracted. A 186-h span of Corpus of Spontaneous Japanese data [24] was used for bidirectional long short-term memory (LSTM) RNN training [25, 26]. The LSTM-RNN comprised one hidden layer with 1,024 units for each gate. The number of classes for the output layer was 3,078 for the phonetic decision-tree-based tied triphone state. The specifications are summarized in Table 1.

Table 1: *Summary of LSTM RNN specifications*

Input layer	120 FBANK
Hidden layer	1 layer x 1,024 node Bidirectional LSTM
Output layer	3,078 class

4.2. Baseline performance of single systems

To combine multiple neural networks, we trained eight bidirectional LSTM-RNNs on two sets of training data and different initial parameter values; random initial weights were uniformly drawn from [-0.04, 0.04] by singular value decomposition. To prepare Dataset 1, we divided each utterance to 0.25 s. For a different training paradigm, the utterance length of Dataset 2 was padded to 2.22, 4.44, 8.35, and 13.9 s by bucketing silence at the end of the utterance. All LSTM-RNNs were trained using up to 20 epochs, with a learning rate of 1.0e-4, and as based on the cross-entropy criteria. To evaluate the single-value STD performance, we calculated the mean average precision (mAP; %) [27]. Table 2 summaries mAP(%) of eight single systems as the baseline STD performance. With a small standard deviation of 0.56, the results show the comparable strong models of different initial parameters and datasets.

Table 2: *mAP(%) of eight single STD systems trained with diverse bidirectional LSTM-RNNs.*

Average	Best	Worst	Std. dev.
80.67	81.73	80.08	0.56

4.3. Fusion on logits with temperature

The STD experiments were performed on logit-level fusion with temperature. The fusion was first executed on logit z; the class probabilities were then calculated by using (5). Because the logit takes values between negative and positive infinity, and thus the fusions of three power means, QM, GM, and HM violate the negative domain inequality, we only carried the score fusion on logit via the original SUM of (4), MAX, and AM.

Table 3: *mAP(%) results of applying different fusion methods and temperatures (T) to the two most preferred single systems.*

	T=1	T=2	T=3	T=4
SUM	82.92	82.08	79.32	71.15
MAX	81.95	71.36	48.31	28.85
AM	82.08	71.15	47.27	27.71

Table 3 shows that the best performance of 82.92 is achieved via SUM, and that all fusions with $T = 1$ lead to performance gain compared to the best result of 81.73 from single systems. However, a temperature increase resulted in markedly degraded performance. The experimental results show that there is a trade-off between the undesirable effects of smoothed probability distribution and the effectiveness of dark knowledge. The higher temperature-induced degradation associated with SUM is comparably less prominent because the SUM is a multiple of the AM on the logit-level fusion. As shown in Fig. 1, as the temperature increases, the variance of probability distribution from score fusion becomes dramatically smaller, resulting in lower discriminative capability. However, compared to the steeply smallized variance of MAX and AM, the variance of SUM becomes smaller gradually. This difference in variance accounts well for the performance degradation, as shown in Table 3, and the trade-off between confusion and dark knowledge.

4.4. Fusion on class probabilities

Next, the STD experiments on fusion of class probabilities were performed with a modified SUM′ (given as (6)) and five power

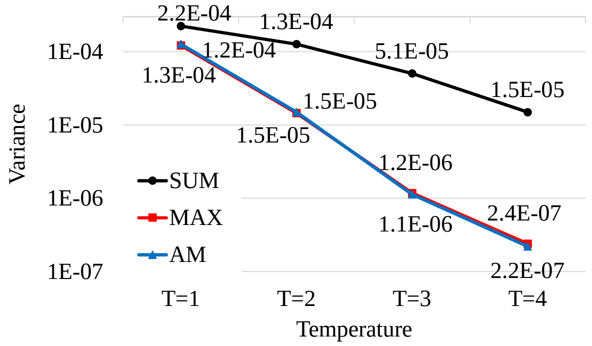

Figure 1: *Variance of probability distribution vs. temperature of score fusion on logits. A base-10 log scale is used for the vertical axis.*

means. Additionally, all 2-way, 4-way, 6-way, and 8-way combinations of eight single neural networks were tested to empirically examine the fusion methods and performance gain according to the number of combined networks.

Table 4: *Average mAP(%) for all combinations and different fusion methods. The values within the parentheses show the best mAP in each combination and fusion.*

	2-way	4-way	6-way	8-way
SUM′	81.54 (82.88)	82.36 (83.38)	83.51 (**83.82**)	83.77
MAX	81.70 (83.10)	82.18 (82.86)	82.28 (82.64)	82.16
QM	81.72 (82.85)	82.22 (82.81)	82.38 (82.86)	82.45
AM	81.66 (82.48)	82.16 (82.80)	82.34 (82.73)	82.49
GM	81.47 (82.30)	81.85 (82.44)	81.99 (82.39)	81.93
HM	81.11 (82.06)	81.37 (82.02)	81.43 (81.78)	81.41

As shown in Table 4, increasing the number of combined networks corresponded to linear performance improvements. Among the fusion methods, the SUM′ of class probabilities yielded the best performance. The results confirm that the classes become more effectively discriminative by summation. Regarding the fusion by power means, although the discriminative classes were achieved and robustly prevented the local minima or over-fitting, the undesired effects of smoothed distribution increased, thereby yielding comparably less significant gains.

Fig. 2 clearly shows that the performance gains via score fusion of class probabilities exactly match the mathematical inequality of the sum and power means, SUM′ $\geq$ MAX $\geq$ QM $\geq$ AM $\geq$ GM $\geq$ HM. By summing the class probabilities of multiple neural networks, dark knowledge can be more effectively utilized without the undesirable effects of smoothed distribution. From further observation of Fig. 2, large variances can be observed in the lesser combinations. However, increasing the number of combined networks increases performance stability. Particularly, the linearly increased gains are more obvious for fusion by SUM′. Because the number of classes is generally extremely higher than the number of combined systems in ASR, it is practically infeasible for most of the class probabilities achieved by SUM′ to be close to one.

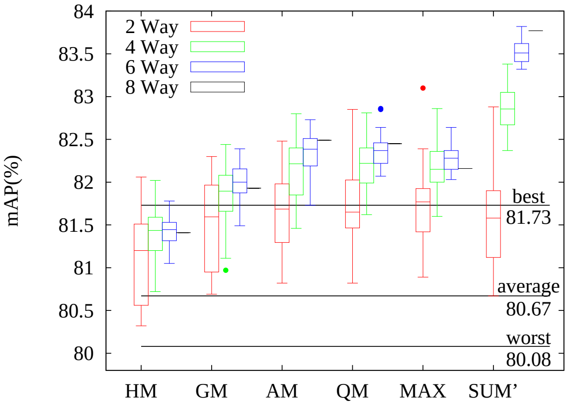

Figure 2: *Box-whisker plot of mAP(%) for various score fusions and combinations. The quartiles for the 2-, 4-, 6-, and 8-way combinations were calculated by using ${}_8C_2 = 28$, ${}_8C_4 = 70$, ${}_8C_6 = 28$, and ${}_8C_8 = 1$, respectively. The three black solid lines are single-system results presented for comparison.*

5. Conclusions

In this study, we examined several fusion methods including logits with temperature and class probabilities for an open vocabulary STD task. The aim was to empirically investigate fusion methods that more effectively use dark knowledge, and to clarify the effectiveness of dark knowledge via experimental evaluation. Performing fusion on logits with varying temperature revealed that there is a trade-off between the undesirable effects of smoothed probability distribution and the effectiveness of dark knowledge, with performance degradations in the case of fusion by sum being comparably less significant. We achieved a maximum of 2.09% absolute gain relative to the best result from single systems performing fusion on class probabilities. Among the fusion methods of the class probabilities, the performance gains achieved via summation are consistently better than those achieved via power means. Furthermore, the experimental analyses of fusions on logits and class probabilities confirmed that summation, which enhances the discriminative power of superior class probability, is proficient at using dark knowledge effectively, thus reducing the undesired effects of smoothed probability distribution. It is especially notable that the performance gains resulting from the fusion of class probabilities exactly match the mathematical inequality for sum and power means results. Based on the experimental evaluations, a more analytical approach to fusion can be designed for open vocabulary STD. Future work will include applying this technique to knowledge transfer learning.

6. Acknowledgements

This research was partially supported by a Grant-in-Aid for Scientific Research (C), KAKENHI Project Nos. 15K00241 and 15K00262.

7. References

[1] J. G. Fiscus, "A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER)," *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 347–354, 1997.

[2] L. Mangu, H. Soltau, H. K. Kuo, B. Kingsbury, and G. Saon, "Exploiting diversity for spoken term detection," *ICASSP*, pp. 8282–8286, 2013.

[3] S. Rath, K. Knill, A. Ragni, and M. Gales, "Combining tandem and hybrid systems for improved speech recognition and keyword spotting on low resource languages," *INTERSPEECH*, pp. 835–839, 2014.

[4] V. Soto, E. Cooper, L. Mangu, A. Rosenberg, and J. Hirschberg, "Rescoring confusion networks for keyword search," *ICASSP*, pp. 7088–7092, 2014.

[5] J. Trmal, G. G. Chen, D. Povey, and et al., "A keyword search system using open source software," *IEEE Workshop on Spoken Language Technology (SLT)*, pp. 530–535, 2014.

[6] L. Deng and C. Platt, "Ensemble deep learning for speech recognition," *INTERSPEECH*, pp. 1915–1919, 2014.

[7] H. Soltau, G. Saon, and T. N. Sainath, "Joint training of convolutional and non-convolutional neural networks," *ICASSP*, pp. 5572–5576, 2014.

[8] T. G. Dietterich, "Ensemble methods in machine learning," *Proceedings of the First International Workshop on Multiple classifier systems*, pp. 1–15, 2000.

[9] G. E. Hinton, O. Vinyals, and J. Dean, "Dark knowledge," *Presented as the keynote in BayLearn*, 2014.

[10] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[11] J. Li, R. Zhao, J. T. Huang, and Y. Gong, "Learning small-size dnn with output-distribution-based criteria," *INTERSPEECH*, pp. 1910–1914, 2014.

[12] W. Chan, N. R. Ke, and I. Lane, "Transferring knowledge from a rnn to a dnn," *INTERSPEECH*, pp. 3264–3268, 2015.

[13] Z. Tang, D. Wang, and Z. Zhang, "Recurrent neural network training with dark knowledge transfer," *ICASSP*, pp. 5900–5904, 2016.

[14] Y. Chebotar and A. Waters, "Distilling knowledge from ensembles of neural networks for speech recognition," *INTERSPEECH*, pp. 3439–3443, 2016.

[15] J. Cui, B. Kingsbury, B. Ramabhadran, and et al., "Knowledge distillation across ensembles of multilingual models for low-resource languages," *ICASSP*, pp. 4825–4829, 2017.

[16] T. Fukuda, M. Suzuki, G. Kurata, S. Thomas, J. Cui, and B. Ramabhadran, "Efficient knowledge distillation from an ensemble of teachers," *INTERSPEECH*, pp. 3697–3701, 2017.

[17] C. Bucila, R. Caruana, and A. Niculescu-Mizil, "Model compression," *ACM SIGKDD*, pp. 535–541, 2006.

[18] NIST, "The spoken term detection (std) 2006 evaluation plan," *http://www.itl.nist.gov/iad/mig/tests/std/2006/docs/std06-evalplan-v10.pdf*, 2006.

[19] Intelligence Advanced Research Projects Activity, *https://www.iarpa.gov/index.php/research-programs/babel.*

[20] K. Ng, "Subword-based approaches for spoken document retrieval," *PhD Thesis, MIT*, 2000.

[21] M. Silaghi and H. Bourlard, "A new keyword spotting approach based on iterative dynamic programming," *ICASSP*, pp. 1831–1834, 2000.

[22] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams," *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 398–403, 2009.

[23] T. Akiba, H. Nishizaki, K. Aikawa, X. Hu, Y. Itoh, T. Kawahara, S. Nakagawa, H. Nanjo, and Y. Yamashita, "Overview of the ntcir-10 spokendoc-2 task," *NTCIR Conference*, pp. 573–587, 2013.

[24] K. Maekawa, "Corpus of spontaneous japanese: its design and evaluation," *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, pp. 7–12, 2003.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[27] C. D. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008.