WORDSPOTTING USING A PREDICTIVE NEURAL MODEL FOR THE TELEPHONE SPEECH CORPUS

Suhardi

Institute for Telecommunication and
Theoretical Electrical Engineering
Technical University of Berlin, Germany
suhardi@ftsu00.ee.tu-berlin.de

Klaus Fellbaum

Communication Engineering
Brandenburg Technical University of
Cottbus, Germany
fellbaum@kt.tu-cottbus.de

ABSTRACT

We describe a wordspotting algorithm based on a predictive neural model for a telephone speech corpus. Each keyword is modeled as a whole word. For keyword detection scoring we used a minimum accumulated prediction residual. We computed empirically a threshold value for rejecting non-keyword speech in place of building non-keyword models. We tested the algorithm with the TUBTEL telephone speech corpus and compared it with other algorithms like the standard DTW-based wordspotting algorithm and the two-stage wordspotting algorithm based on a DTW and a multilayer perceptron.

1. INTRODUCTION

Recently, novel results indicated that neural networks can improve the speech recognition performance if they are embedded into hybrid speech recognition systems; they can be incorporated with a DTW framework[1, 2, 3, 4] or combined with HMMs [5, 6, 7].

Applications of neural networks to improve the detection rate of wordspotting algorithms are proposed, too. In [8, 9, 10, 11, 12] neural networks are used as a keyword/false alarm discriminator to improve DTW-based or HMM-based wordspotting algorithms.

Another way for improving wordspotting algorithms is modeling of non-keyword speech as garbage models [13, 14, 15].

We have been developing wordspotting algorithms based on a multilayer perceptron (MLP). The MLP is embedded either into DTW framework or into HMM framework. In this paper we will describe our wordspotting algorithm based on a predictive neural model (PNM). The PNM consists of a number of multilayer perceptrons (MLPs). The algorithm has several advantages. Its structure is simple and easy to be trained. The training can be done flexibly based on a word, syllable or phone level. There is no need for a non-keyword speech training. In spite of that we empirically computed a theshold value of the accumulated prediction residual as a criterium for rejecting non-keyword speech. In our experiment we modeled keywords as a whole word and used the telephone speech corpus (TUBTEL) [16] for training and test purposes.

This paper is subdivided into five sections. After this short introduction, the second section introduces the PNM-based wordspotting algorithm. The third section describes an experiment for which the speech corpus TUBTEL was used. The fourth section presents some results of the experiment. Finally, we give some conclusions.

2. PNM-BASED WORDSPOTTING

The PNM was firstly proposed by K. Iso [1] for speaker-independent isolated and continuous speech recognition. We adapted the PNM to the wordspotting problem. We applied the PNM for modelling keywords [17]. The PNM representing a keyword sw consists of a number (N_{sw}) of MLPs applied as a speech pattern predictor. Each MLP n $(1 \leq n \leq N_{sw})$ predicts an actual speech feature vector at every time t ($\mathbf{x}_t$) based on previous speech feature vectors ($\mathbf{x}_{t-\tau_l}, \ldots, \mathbf{x}_{t-2}, \mathbf{x}_{t-1}$) and the preceding speech feature vectors ($\mathbf{x}_{t+1}, \mathbf{x}_{t+2}, \ldots,$ $\mathbf{x}_{t-\tau_r}$). τ_l is the left prediction order, and τ_r is the right one. The predicted speech feature vector at time t ($\hat{\mathbf{x}}_t$) is computed as follow :

$$\hat{\mathbf{x}}_t = \mathcal{F}(\mathbf{x}_{t-\tau_l}, \ldots, \mathbf{x}_{t-2}, \mathbf{x}_{t-1}, \mathbf{x}_{t+1}, \mathbf{x}_{t+2}, \ldots, \mathbf{x}_{t+\tau_r}) \quad (1)$$

$\mathcal{F}$ is the input-output relation of a MLP predicting the vector $\hat{\mathbf{x}}_t$. A squared Euclidean distance between the

actual vector ($\mathbf{x}_t$) and the predicted one ($\hat{\mathbf{x}}_t$) gives a measure of a local prediction residual of keyword sw at time t for the MLP n (equation 1).

$$d_{sw}(t,n) = \| \mathbf{x}_t - \hat{\mathbf{x}}_t \|^2; \quad (2)$$

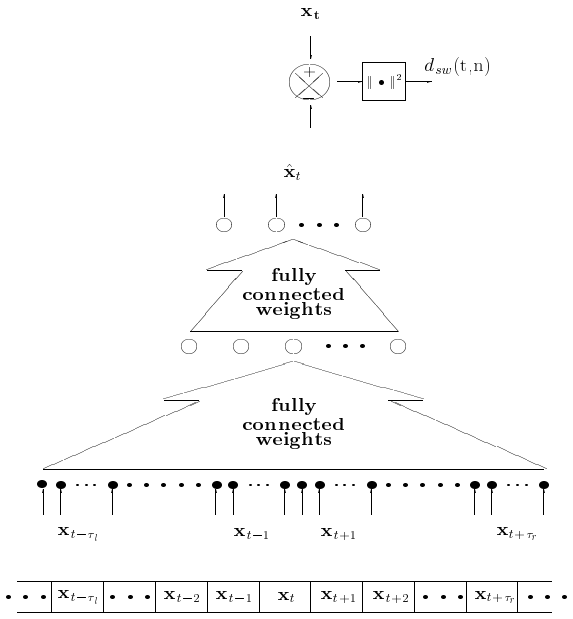

Figure 1: MLP as predictor

An accumulated prediction residual of keyword sw is defined by [1] :

$$D_{sw}(t) = \min_{n(t)} \sum_{n(t)} d_{sw}(t,n) \quad (3)$$

n = n(t) is a warping function on the t-n plane (figure 2) starting at t_a and finishing at t_e along the t axis, where

$$n(t_a) = 1 \quad (4)$$

$$n(t_e) = N_{sw} \quad (5)$$

$$0 \le n(t) - n(t-1) \le 1; \qquad t_a \le t \le t_e \quad (6)$$

$$\tau_l < t_a \le T - N_{sw} - \tau_r \ ; \quad (7)$$

$$(N_{sw} + \tau_l) \le t_e \le T - \tau_r \quad (8)$$

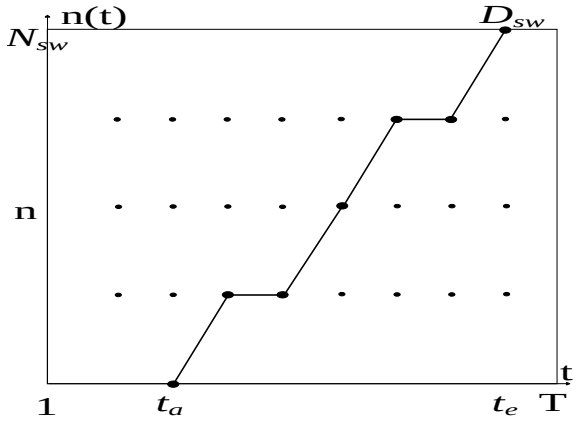

Figure 2: An example of a warping function in the t-n plane

This accumulated prediction residual is calculated by the DP equation [1] :

$$g_{sw}(t,n) = d_{sw}(t,n) + min \left\{ \begin{array}{c} g_{sw}(t-1,n) \\ g_{sw}(t-1,n-1) \end{array} \right\} \quad (9)$$

$$D_{sw}(t) = g_{sw}(t_e, N_{sw}) \quad (10)$$

The keyword detection is conducted by a continuous pattern matching. We used the optimum accumulated prediction residual $D_{sw}(t)$ starting at every time t_a along a test utterance. A putative keyword sw lies at t_e with the score $D_{sw}(t)$.

$$D_{sw}^{min} = \min \{D_{sw}(t)\} \quad (11)$$

$$t_e = arg \min_t \{D_{sw}(t)\} \quad (12)$$

If the vocabulary consists of K keywords, sw = 1, 2, 3, ... , K, after the continuous pattern matching we have K D_{sw}^{min}. If $D_{sw}^{min} \le \delta_{threshold}$ then sw is a putative keyword. We used just a simple decision rule for detecting the keyword among the putative keywords.

$$keyword = arg \min_{sw} \{D_{sw}^{min}\} \quad (13)$$

Each PNM is trained based on DP time alignment and back-propagation using a set of training patterns. The training goal is to find a set of MLP predictor weights, which minimize the accumulated prediction residuals for a training data set. We used the training procedure proposed by K. Iso *et. al.* [1].

3. EXPERIMENTS

We trained and tested our wordspotting algorithm on the telephone speech corpus TUBTEL [16]. The speech corpus was recorded directly from the telephone network at the Institute for Telecommunication, Technical

University of Berlin. A part of this speech corpus is designed for the developement of a wordspotting system. There are five types of sentences,

- Ich brauche .....
- Gibt es ..... bei Ihnen
- Ich möchte ..... kaufen
- Ich möchte ..... bestellen
- Könnten Sie mich vielleicht informieren, ob Sie ..... haben

and nine keywords : Rock, Bluse, Kleid, Hose, Shorts, Sandalen, Schuhe, Strümpfe, Gürtel.

We combined the five sentences and nine keywords to 45 sentences. The speech corpus consists of circa 550 speakers, but we used only 378 of them. Every speaker spoke one of the 45 kinds of sentences. The training patterns were segmented from continuous utterances. Each keyword was trained with circa 30 patterns. For testing we used continuous utterances in which one keyword of the vocabulary exists [12].

We used cepstral coefficients as feature representation of the speech signal. The cepstral coefficients are computed according to the JRASTA method [18], where the J value was adapted to the SNR of the speech signal. We used a Hamming window, its length is 32 ms. The window step is 8 ms. For comparison we took the MFCCs, where its window analysis is the same as those used for JRASTA. For both methods we computed 10 coeffients for every frame.

Table 1 shows keywords, the number of MLPs for every PNM, the number of training patterns and the number of test utterances of every keyword.

The MLP for this experiment has three layers : input layer, one hidden layer and output layer. We set $\tau_l = \tau_r = 3$, therefore the input layer had 60 nodes. The hidden layer had 15 nodes and the output layer consisted of 10 nodes.

Experiments with other methods, but the same training and test data set, were carried out for comparison. The methods are the standard DTW [12, 19] and the two stage wordspotting algorithm based on DTW and MLP [12]. In the two stage wordspotting algorithm the MLP is used as a secondary processing. The first processing was standard DTW that produced putative keywords. In the secondary processing the keyword with the higest probability was selected.

Table 1: Data for our experiments

No.	keyword	number of MLPs (N_{sw})	number of training patterns	number of test sentences
1.	Rock	7	30	21
2.	Bluse	12	30	20
3.	Kleid	12	30	13
4.	Hose	14	31	24
5.	Shorts	10	28	18
6.	Sandalen	20	27	20
7.	Schuhe	15	19	16
8.	Strümpfe	20	27	14
9.	Gürtel	15	35	15

4. RESULTS

Table 2 shows our experimental results. The detection rate (DR) was calculated as a number of sentences in which a keyword was correctly detected/recognized divided by the total number of test sentences. No. 1 is the DR of the PNM-based wordspotting algorithm. For comparison we presented our experimental results with the same data base using other algorithms, i.e. two stage wordspotting (No. 2) and standard DTW (No. 3). The DTW-based wordspotting algorithm is described in [18]. The third column of table 2 are the results for 10 MFCC coefficients and the fourth column contains the results using 10 cepstral coefficients with JRASTA method in which the J factor was adapted to every speech signal.

Table 2: Detection rate (DR)

No.	method	DR (%) MFCC	DR(%) JRASTA
1.	PNM	89	92
2.	DTW + MLP	62	66
3.	DTW	60	61

5. CONCLUSION

The PNM-based wordspotting algorithm has a better detection rate than the standard DTW and the two stage wordspotting algorithm (DTW + MLP). The algorithm has a high recognition rate to detect a keyword in a continuous utterance under the real telephone channel environment, though it was trained with only a few training patterns. It needs no modeling of the non-keyword speech.

The computation for a detection process is still time consuming. We are now improving the algorithm to speed up the detection process. We are developing a method to find word hypotheses, so that the computation for keyword searching can be reduced.

For the future work we are going to investigate the algorithm with subword modeling for a task-independent wordspotting system and we will intensively adapt the algorithm to the telephone channel environment.

REFERENCES

[**1**] Iso, K., Watanabe, T.; Speaker-Independent Word Recognition Using a Neural Prediction Model, IEEE Proc. of ICASSP-90, pp. 441-444.

[**2**] Iso, K., Watanabe, T.; Large Vocabulary Speech Recignition Using Neural Prediction Model, IEEE Proc. of ICASSP-91,

[**3**] Mellouk, A., Gallinari, P.; A Discriminative Neural Prediction System for Speech Recognition, International Conference on Artificial Neural Networks 1993, pp. 383-388.

[**4**] _____; Global Discrimination for Neural Predictive Systems Based on N-Best Algorithm, IEEE Proc. of ICASSP-95, pp. 465-468.

[**5**] Morgan, N., Bourlard, H.; Neural Networks for Statistical Recognition of Continuous Speech. Proceedings of the IEEE, Vol. 83, No. 5, May 1995, pp. 742 - 770.

[**6**] Bourlard, A., Morgan, N.; Connectionist Speech Recognition - A hybrid Approach. Amsterdam : Kluwer, 1994.

[**7**] Kershaw, D., Hochberg, M., Robinson, A.; Context-Dependent Classes in a Hybrid Recurrent Network-HMM Speech Recognition System, Technical Report, CUED/F-INFENG/TR217, Cambridge University Engineering Department.

[**8**] Morgan, D., Scofield, C. and Adcock, J.; Multiple Neural Network Topologies Applied to Keyword Spotting, IEEE Proc. of ICASSP-91, pp. 313 - 316.

[**9**] Naylor, J. Rossen, M.; Neural Network Word/False-Alarm Discriminators for Improved Keyword Spotting, International Joint Conference on Neural Network (IJCNN), 1992, pp. II-296 - II-301.

[**10**] Lippmann, R., Singer, E.; Hybrid Neural Network/HMM Approaches to Wordspotting, Proc. IEEE of ICASSP-93, pp. I-565 - I-568.

[**11**] Lippmann, R., Chang, E. Jankowski, C.; Wordspotter Training Using Figure-Of-Merit Back Propagation, IEEE Proc. of ICASSP-94, pp. I-389 - I-392.

[**12**] Suhardi, K. Fellbaum; Zur Schlüsselworterkennung Fliessender Sprache unter Verwendung Neuronaler Netze, Sechste Konferenz, Elektronische Sprachsignalverarbeitung, Wolfenbüttel, 1995, ISSN 0940-6832, pp. 48 - 55.

[**13**] Rose, R. C.; Keyword Detection in Conversational Speech Utterances Using Hidden Markov Model based Continuous Speech Recognition, Computer Speech and Language (1995) 9, pp. 309-333.

[**14**] Mukarami, J.; New Word Spotting Algorithm based on forward Decoding. IEEE Proc. European Conference on Speech Communications and Technology, pp. 2153-2156, Madrid, September 1995.

[**15**] Klemm, H., Class, F., Kilian, U.; Word and Phrase Spotting with Syllable-based Garbage Modelling. European Conference on Speech Communications and Technology, pp. 2157-2160, Madrid, September 1995.

[**16**] Schürer, T., Ahrling, S., Fellbaum, K., Hardt, D., Klaus, H., Mengel, A., Sahm, O., Suhardi; TUBTEL - Eine deutsche Telefon-Sprachdatenbank, Sechste Konferenz, Elektronische Sprachsignalverarbeitung, Wolfenbüttel, 1995, ISSN 0940-6832, pp. 183 - 187.

[**17**] Suhardi, Fellbaum, K.; Zur Schlüsselworterkennung unter Verwendung prädiktiver neuronaler Modelle, Siebente Konferenz, Elektronische Sprachsignalverarbeitung, Berlin, 1996, ISSN 0940-6832, pp. 108 - 114.

[**18**] Hermansky, H., Morgan, N.; RASTA Processing of Speech. IEEE Transaction on Speech and Audio Processing, Vol.2 , No. 4, Oct. 1994, pp. 578 - 589.

[**19**] Wang, D.; Speech Recognition with Word Spotting Techniques. Dissertasion, Institute for Telecommunication, Technical University of Berlin, 1993.