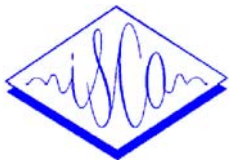

EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

EMPIRICAL COMPARISON OF TWO MULTILAYER PERCEPTRON-BASED KEYWORD SPEECH RECOGNITION ALGORITHMS

Suhardi

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

Klaus Fellbaum

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

ABSTRACT

In this paper, an empirical comparison of two multilayer perceptron (MLP)-based techniques for keyword speech recognition (wordspotting) is described. The techniques are the *predictive neural model (PNM)-based wordspotting*, in which the MLP is applied as a speech pattern predictor to compute a local distance between the acoustic vector and the phone model, and the *hybrid HMM/MLP-based wordspotting*, where the MLP is used as a state (phone) probability estimator given acoustic vectors. The comparison was performed with the same database. According to our experiments, the hybrid HMM/MLP-based technique excels the PNM-based techniques (~6.2 %).

1. INTRODUCTION

Recently, wordspotting techniques are in the focus of attention of speech recognition researchers, because it gives users the flexibility to speak naturally in a man-machine dialog application. It screens only pre-defined keywords in a continuous input speech, so that the user can utter continuous sentences that consist of unconstrained words. There are many potential applications for wordspotting, e.g. information retrieval based on natural speech, machine control by voice and context-dependent speech understanding.

Therefore, we need wordspotting techniques that can reject non-keyword speech and detect keywords in a continuous utterance. Research and development activities for wordspotting techniques have two directions. The first one is to model the non-keyword speech optimally and the second one is to use keyword detection techniques and their appropriate scoring measure. This paper investigates wordspotting techniques based on the MLP - predictive neural model (PNM) and hybrid HMM/MLP - and reports about an experimental comparison of both approaches.

2. MLP-BASED WORDSPOTTING ALGORITHMS

2.1. PNM-based Wordspotting

PNM is a class of acoustic-phonetic modeling methods for words/subwords that consist of MLPs embedded into a DTW framework [1]. A word can be modeled as a sequence of MLPs. The MLP is applied as a predictor to compute predicted acoustic vectors based on the previous and preceeding acoustic vectors. The predicted vector is then compared to the actual one to generate a local pattern variability score at every frame. Figure 1 is an MLP architecture that computes a local pattern distance in eq. 1. The local score is used by the DP method for pattern matching along the word model. The technique was succesfully applied for isolated word recognition [1], continuous speech recogniton [2] and whole word model-based wordspotting [3, 4].

Keyword models can be built as a concatenation of subword (phone) models. Each phone is presented by one MLP, so we have a set of L MLP weight parameters $\mathcal{W} = \{\Theta^1, \Theta^2, \ldots, \Theta^l, \ldots, \Theta^L\}$;

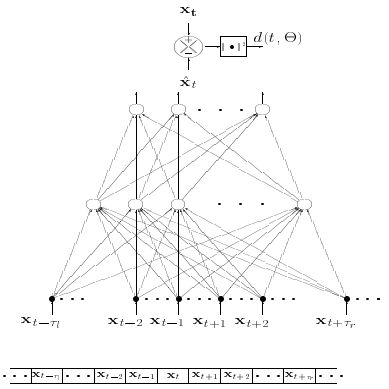

Figure 1: An MLP architecture with *fully conneted* weights to compute a local pattern distance in eq. 1.

L is the total number of phones in the vocabulary. Figure 2 illustrates a PNM for keyword kw. The

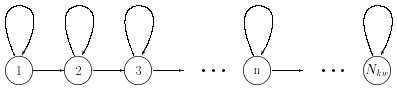

Figure 2: PNM of keyword kw that consists of N_{kw} nodes

keyword consists of N_{kw} phones, so that the PNM has N_{kw} nodes, where each node corresponds to one MLP that models one phone. The PNM can be presented as $\omega(kw) = \{\Theta_{kw}(1), \Theta_{kw}(2), \ldots, \Theta_{kw}(n), \ldots, \Theta_{kw}(N_{kw})\}$, where $\Theta_{kw}(n)$ is the weight parameter of the MLP that models the n^{th} phone of the keyword kw. Given acoustic input vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T\}$, a local distance between an acoustic vector at time t ($\mathbf{x}_t$) and the n^{th} phone of keyword kw can be computed according to

$$d_{kw}(t,n) = \parallel \hat{\mathbf{x}}(t, \Theta_{kw}(n), \tau_f, \tau_b) - \mathbf{x}_t \parallel^2 \qquad (1)$$

where $\hat{\mathbf{x}}(t, \Theta_{kw}(n), \tau_f, \tau_b)$ is the predicted acoustic vector at time t by the n^{th} MLP of the keyword model kw based on the τ_f frames of its previous acoustic vectors and the τ_b frames of its preceeding acoustic vectors.

Each frame of the acoustic input vectors can be assumed as an endpoint of keyword kw whose accumulated distance along the keyword $\mathcal{G}_{kw}(t)$ can be computed recursively according to the DP formula (eqs. 2 and 3).

$$g_{kw}(t,n) = d_{kw}(t,n) + min \left\{ \begin{array}{c} g_{kw}(t-1,n) \\ g_{kw}(t-1,n-1) \end{array} \right\} \qquad (2)$$

$$\mathcal{G}_{kw}(t) = g_{kw}(t, N_{kw}) \; ; \; t = \tau_f + 1, \ldots, T - (\tau_b + 1) \qquad (3)$$

A putative keyword was declared if a minimum value of this accumulated distance sequence is smaller than a pre-defined threshold value (δ_D). If the procedure is done for all K keywords, the detected keyword is decided according to the rule

$$detected\ keyword = \underset{kw}{argmin} \ \{\mathcal{G}_{kw}(t) < \delta_D\} \qquad (4)$$

where $kw = 1, 2, \ldots, K$.

2.2. Hybrid HMM/MLP-based Wordspotting

Hybrid HMM/MLP has been proposed to improve the standard HMM for acoustic-phonetic modeling in speech recognition by integrating MLPs into the HMM framework [5]. The MLP is trained under supervised mode as a classificator to estimate posterior probabilities of output classes (states) given acoustic vectors and previous state information [6]. Figure 3 shows a hybrid HMM/MLP, where the previous state information is coded in binary input vectors integrated to the input acoustic vectors. The trained MLP can be used to compute

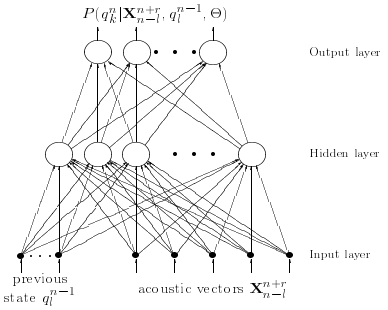

Figure 3: Hybrid HMM/MLP trained with the REMAP as described in [6]

state (phone) probabilities given the acoustic and previous state vectors.

A keyword kw consists of a concatenation of phones, thus the keyword kw can be presented as a sequence of output node numbers in a lexicon according to its phonetic transcription. Figure 4 illustrates a keyword model kw that consists of J_{kw} states $\{q(1), q(2), q(3), \ldots, q(j), \ldots, q(J_{kw})\}$.

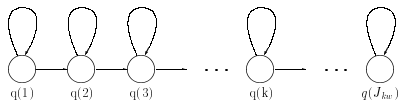

Figure 4: A keyword model kw based on hybrid HMM/MLP

Given a test sentence $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T\}$, the probability that the endpoint of keyword kw occurs at time t can be computed using the Viterbi algorithm by assuming that each frame of the test sentence was optimally aligned to the last state ($q(J_{kw})$) of the keyword model kw. Thus, calculation of the (log) probability can be accomplished recursively according to the DP method.

$$\phi_{kw}(t, q(j)) = max \left\{ \begin{array}{c} \phi_{kw}(t-1, q(j)) \\ \phi_{kw}(t-1, q(j-1)) \end{array} \right\} \times log\ P(q^t(j) | X_{t-l}^{t+r}, q_i^{t-1}, \Theta) \quad (5)$$

$$\Phi_{kw}(t) = \left\{ \begin{array}{ll} \phi_{kw}(t, q(j)) & for\ \phi_{kw}(t, q(J_{kw})) > \delta_P \\ -\infty & for\ \phi_{kw}(t, q(J_{kw})) \leq \delta_P \end{array} \right. \quad (6)$$

where $\Phi_{kw}(t)$ for $t = l+1, \ldots, T-(r+1)$ is a sequence of accumulated probabilities, l is the number of previous acoustic vectors correlated to the actual acoustic vector and r is the number of preceeding vectors correlated to the actual vector; $\mathbf{X}_{t-l}^{t+r} = \{\mathbf{x}_{t-l}, \ldots, \mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1}, \ldots, \mathbf{x}_{t+r}\}$. The maximum value of this sequence indicates the occurence of a putative keyword kw, if the value is greater than a pre-defined threshold value (δ_P).

If the above procedure is executed for all K keyword models, the decision on whether one of the keyword candidates is accepted as a detected keyword or all cadidates are rejected is accomplished as follows.

$$detected\ keyword = \underset{kw}{argmax}\ \{\Phi_{kw}(t) > \delta_P\} \quad (7)$$

3. DATABASE

Both wordspotting techniques were trained with the Phondat 1 database (CD-1 and CD-2) and tested on the Phondat 2 database. The Phondat 1 & 2 databases [7] are speech corpora recorded in the studio environment. They are partially labeled according to the SAMPA segmentation [7]. Phondat 1 consists of 4 CDs, but we used only CD-1 and 2 to train both wordspotting systems. The Phondat 2 database consists of 16 speakers. Each speaker spoke 200 sentences. Ten Speakers were used to test the systems. Six speakers were used as cross-validation during the training of the hybrid HMM/MLP. The vocabulary has 10 keywords: *Augsburg, Dortmund, Frankfurt, Hamburg, Hindelang, Mannheim, München, Nürnberg, Regensburg, Würzburg.* We have chosen test sentences that contain only one keyword. The test database has a total of 740 ontinuous sentences.

4. EXPERIMENTS AND RESULTS

We used two methods for the feature extraction, namely MFCC [8] and PLP [9]. In both cases, the Hamming window had a length of 20 ms and a frame rate of 10 ms. Each frame was represented by 10 cepstral coefficients. The subword vocabulary consisted of 53 phone models.

The first experiment investigated the performance of the PNM-based wordspotting technique. We used $\tau_f = 2$ and $\tau_b = 1$, so each MLP that models a phone had 30 input nodes and 10 output nodes. The MLP had one hidden layer with 15 nodes. Feature parameters were linearly normalized to [-1, +1], before they were applied as input and target vectors of the MLPs. We used a sigmoid activation function that is non-symmetric ($f(x) = \frac{1}{1+exp(-x)}$) for hidden nodes; but for the output nodes, the asymmetric sigmoid function (f(x) = 2 tanh (0.75 x)) was adopted, so we used only the (closed) linear region of the function for output nodes. For each acoustic training sentence and its phonetic label, we built a sequence of MLPs according to the given phonetic label, then each MLP in this sequence was trained with the given acoustic training vectors based on the error back-

propagation [10] to minimize accumulated distances along the modeled phone.

The second experiment was performed with the hybrid HMM/MLP-based wordspotting technique. We have chosen l = 4 and r = 4, thus the MLP had 143 input nodes and 53 output nodes. The MLP had one hidden layer with 1000 nodes. The hybrid HMM/MLP was trained according to the REMAP, as described in [6].

The detection rate DR (%)is equal to the number of sentences in which a keyword was correctly recognized, divided by the total number of test sentences as a function of the probability of false alarm PF (%). PF is computed as a probability that a putative keyword is a false alarm. The results are shown in figure 5. Our experiments with

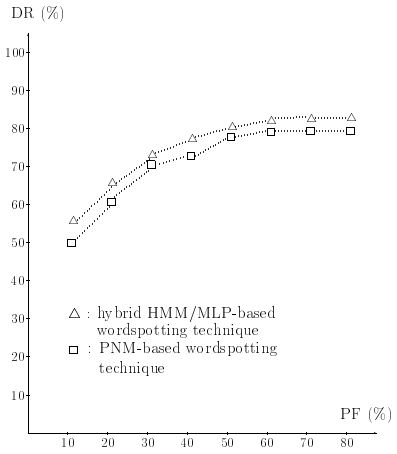

Figure 5: Detection rate of both wordspotting techniques using PLP

MFCCs gave results that have no significant difference ($\sim 0.1\%$ or smaller) in comparison to the results using PLP.

5. CONCLUSIONS

Two wordspotting techniques based on MLP were investigated and empirically compared. Our experiment results have shown that the HMM/MLP-based wordspotting technique is better than the PNM-based one. The difference in the detection rate is in the order of 6 %. Under studio conditions, there was no significant difference in the results for MFCCs and PLPs.

In our next experiments, we will integrate non-keyword models in the MLP-based wordspotting technique and we will adapt it to the telephone network environment.

6. REFERENCES

[1] Iso, K., Watanabe, T. Speaker-independent word recognition using a neural prediction model, *Proc. of IEEE Conference on Acoustic, Speech and Signal Processing*, 1990, pp. 441-444.

[2] Mellouk, A., Gallinari, P. Global discrimination for neural prediction system based on N-best algorithm, *Proc. of IEEE Conference on Acoustic, Speech and Signal Processing*, 1995, pp. 465-468.

[3] Suhardi, Fellbaum, K. Zur Schlüsselworterkennung unter Verwendung prädiktiver neuronaler Netze, *7. Konferenz Elektronische Sprachsignalverarbeitung*, Berlin-Hirschgarten, 1996

[4] Suhardi, Fellbaum, K.; Wordspotting using a predictive neural model for the telephone speech corpus, *Proc. of IEEE International Conference on Acoustic Speech and Signal Processing*, Munich, April 1997.

[5] Bourlard, H. A., Morgan, N. Connectionist speech recognition : A hybrid approach, Kluwer Academic Publishers, 1994.

[6] Bourlard, H.A., Konig, Y., Morgan, N. REMAP: recursive estimation and maximization of posteriori probabilities in connectionist speech recognition, in *Proc. Europ. Conf. Speech Commun., Technol. (EUROSPEECH)*, Madrid, Spain, Sept. 1995.

[7] Documentation of the Phondat 1 & 2 CDs

[8] Davis, S.B., Mermelstein, P.; Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. on ASSP, Vol. 28, No. 4, 1980, pp.367-366.

[9] Hermansky, H.; Perceptual linear predictive (PLP) analysis of speech, J. Acoustic Soc. America, Vol. 87, No.4, 1990.

[10] Rumelhart, D., *et al.* Learning internal representations by error propagation. In *Parallel Distributed Processing*, Rumelhart, D., *et al*, 1986.