# A SUBWORD NEURAL TREE NETWORK APPROACH TO TEXT-DEPENDENT SPEAKER VERIFICATION

*Han-Sheng Liou, Richard J. Mammone*

CAIP center, Rutgers University,
Piscataway, New Jersey 08855, USA

## ABSTRACT

In this paper, a new algorithm for text-dependent speaker verification is presented. The algorithm uses a set of concatenated Neural Tree Networks (NTN's) trained with subword units for speaker verification. The conventional NTN has been found to provide good performance in text-independent tasks. In the new approach, two types of subword unit are investigated, phone-like units (PLU's) and HMM state-based units (HSU's). The training of the models includes several steps. First, the predetermined password in the training data is segmented into subword units using a Hidden Markov Model (HMM) based segmentation method. Second, an NTN is trained for each subword unit. The new structure integrates the discriminatory ability of the NTN with the temporal models of the HMM. This new algorithm was evaluated by experiments on a TI isolated-word database, and YOHO database. An improvement of performance was observed over the performance obtained using a conventional HMM.

## 1. INTRODUCTION

In general, text-dependent speaker verification systems achieve better performance than text-independent systems. In addition, text-dependent speaker verification systems typically require shorter training and testing utterances than that required by text-independent systems. Furthermore, the passwords used to verify speakers can be chosen arbitrarily by the speakers themselves. This provides one more level of protection for the users. Consequently, most speaker verification services currently in the marketplace are text-dependent ones.

In a text-dependent speaker verification system, a speaker is usually asked to select a fixed password. However, The drawback of using a fixed password is that it could be defeated by mimicking or playing a recorded voice of the true speaker. The problem can be overcome by randomly prompting a temporary password which can not be predicted beforehand. In this case, it is not possible to train on the word models. Instead, subword models based on a small amount of training data must be used. Once the set of subword models of a speaker is obtained, the model of an arbitrary password is constructed by concatenating a sequence of subword models together. In this approach, the password changes with time, and the speakers have to speak the prompted word correctly to be verified. Therefore, it enhances the security of the system. Recently, the speaker verification systems based on characterizing the password of a speaker as a sequence of concatenating subword units represented by Hidden Markov Models (HMM's) has been investigated [7, 5]. The subword based model was shown effective in speaker verification tests for password of connected digits or a randomly prompted sentence.

The speaker verification system described in this paper uses subword approach. The discriminative training is performed in each subword model. Thus each subword unit is modeled for the speaker with respect to other speakers. The differences between two speaker can be discriminated only when their utterances are time aligned. When the speech waveforms corresponding to the same context are aligned, the different way of pronouncing phonemes by two speakers can be differentiated. In this paper, a new approach using the Neural Tree Network (NTN) classifier is applied to text-dependent speaker verification. The training algorithm of the NTN is based on discriminant learning, which is a learning algorithm applied to classifiers that minimize the classification error rate. Traditional classifiers are based on minimizing the approximation error which is not directly related to classification performance. In contrast to traditional classifiers, the NTN not only models the statistical distribution of the training data for a speaker accurately, but also has the ability of discriminating the speaker from impostors. The NTN has been shown to yield substantial improvement over conventional methods on text-independent speaker verification

This work was sponsored by Rome Laboratory, Contract No. F30602-91-c-0120. The authors would like to thank Joseph Campbell and Chiwei Che for their useful suggestions and help in the experiments.



0-7803-2431-5/95 $4.00 

[3]. Usually, the text-independent speaker verification system extracts features from speech, and classifies feature vectors in a static phonetic space. However, the temporal information is critical for the text-dependent speaker verification. A number of hybrid algorithms based on HMM and artificial neural networks (ANN's) have been proposed to enhance the performance of conventional HMM classifier [4]. They have been applied to speech and speaker recognition and achieve various improvement over a benchmark performance of HMM. Motivated by these studies, a new system that integrates the NTN with an existing HMM framework for speaker verification is proposed. In this system, an NTN is trained for each subword unit. The scoring method that combines the confidence measure of the subword NTN is described. To evaluate the performance of this algorithm, we conducted text-dependent speaker verification experiments using a TI isolated-word database and YOHO database. Experimental results show that the proposed hybrid method can achieve better performance than that obtained by either NTN or HMM classifiers alone.

## 2. NEURAL TREE NETWORK MODELS

The neural tree network is a tree-structured classifier that combines the properties of the neural network and the decision trees [8]. Discrimination at each node is implemented by a simple neuron that can be trained to have the minimum classification error. Recently, a modified neural tree network (MNTN) has been applied to text-independent speaker verification [3]. As described in [3], each speaker is modeled by a binary NTN which is trained by the feature vectors of that speaker and the all the other speakers. During training, the feature vectors of the speaker are labeled '1', and those of the other speakers are labeled '0'. The NTN is recursively trained in the following way. Given a set of training data at a particular node, the neuron is trained to split the feature vectors into two subsets that minimizes the classification error. These subsets are subsequently passed to children of the node. This algorithm recurrently proceeds until the subset contains the feature vectors of the same class, or the growth to the prespecified level is reached. The leave at the terminal nodes are labeled by the majority class, and the confidence measure of each leaf is also computed. During testing, the speaker likelihood is computed as the ratio of the accumulated speaker confidence measure to the sum of the accumulated speaker and antispeaker confidence measures. Hence, the likelihood of speaker $k$ for the NTN is computed as

$$P(\mathbf{x}|S_k) = \frac{\sum_{i=1}^{M} c_i^1}{\sum_{j=1}^{N} c_j^0 + \sum_{i=1}^{M} c_i^1}, \qquad (1)$$

where $M$ and $N$ are the number of vectors classified as the speaker, or the antispeaker, respectively, and $c_1$ and $c_0$ are the confidence measures, given that a vector was classified as the speaker, or the antispeaker, respectively.

## 3. SUBWORD NEURAL TREE NETWORK MODELS

In this section, the subword NTN model is presented for text-dependent speaker verification. The subword NTN differs from the multiple-word NTN in the sense that they use a different training data set. Instead of using all the words of training data to train a large NTN, the new training algorithm only takes vectors assigned to particular subword units in the training speech to train a subword NTN.

To obtain good training of NTN for subwords, the speech segmentation is an essential process. The error in segmentation tends to corrupt the speaker verification by matching the testing subword to the incorrect subword NTN. Previously, the reliability of phonetic detection was inadequate to support a good speaker verification performance, so the phonetic matching approaches were usually not used in the speaker verification [2]. Currently, the HMM has been successfully applied to speech recognition and phoneme segmentation for its ability to handle the temporal variation of speech. We use HMM based segmentation scheme as the first step in defining the subword model.

In this paper, two types of subword units are used to model speech, phone-like units (PLU's) and HMM state-based units (HSU's). The PLU's are based on phonetic transcriptions of a spoken utterance. The HSU's are subword units based on segmentation of the HMM. To extract the HSU's from an utterance, the whole word is modeled by a fixed-state HMM. The utterance is then segmented into subword units by the HMM based segmentation. These speech segments correspond to states of the HMM, which are referred to as HSU's. The major differences between the two subword units are as follows:

1. The HSU's are based on each single state in word based HMM's, but the PLU's are modeled by phoneme based HMM's which consist of multiple states.

2. In training of the HSU's, A context-dependent HMM is trained by specific passwords, but the training of phoneme based HMM for PLU's is in context-independent mode.

NTN models are then trained on either HSU or PLU data. Speaker verification systems based on these two models are described in the next section.

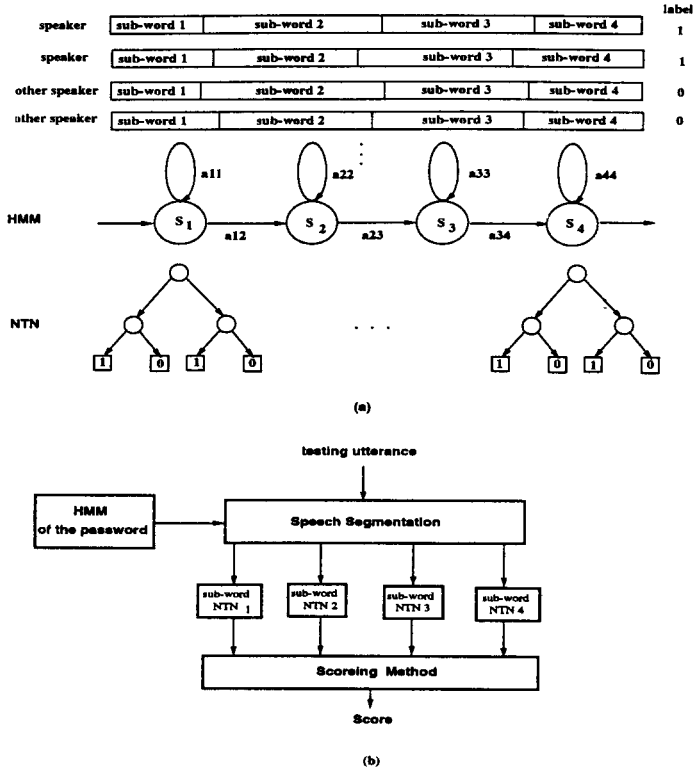


Figure 1: (a) Training of HMM state-based NTN's. (b) Testing of HMM state-based NTN's.

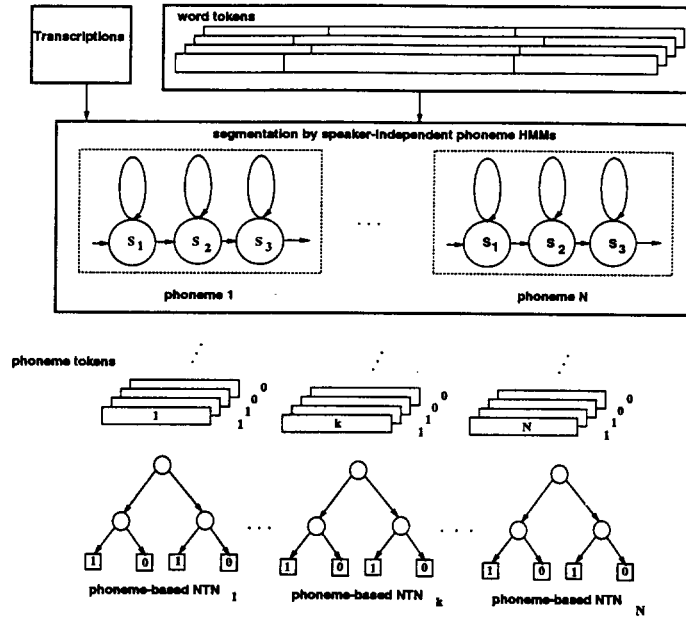


Figure 2: The training of phoneme-based NTN's.

### 3.1. HMM state-based NTN

In the HSU based subword model, the password is modeled by a word HMM. A subword unit is defined as a speech segment that is clustered into a particular state in the HMM. The training procedures are illustrated in Fig. 1(a). First, a speaker-independent word-based HMM is trained for each password. Second, the training utterances are segmented into subword units by the corresponding HMM's. Finally, a subword NTN is trained by the speech segments associated with each subword unit. During testing, as shown in Fig. 1(b), the unknown utterance is segmented by the HMM of the password. Then, the score is measured by the subword NTN's. Since the subword NTN is trained by the same speech event, it should be more powerful in discriminating the speaker from impostors than a multiple-word model.

### 3.2. Phoneme-based NTN

The PLU's are subword units extracted from utterances according to the transcriptions. The procedure for training phoneme-based NTN is illustrated in Figure 2. A speaker-independent phoneme-based HMM is used to model each subword. The parameters of a set of HMM's is initially given by existing models trained by the Resource Management database. Then they are re-estimated by the training utterances in the YOHO database using the Baum-Welch algorithm. For each training utterance, a composite model is synthesized by concatenating phoneme models given by the transcription. After the re-estimation, all the utterances are segmented into subwords and labeled by a Viterbi decoding technique based on the composite models. A speaker-specific phoneme-based NTN is trained for each PLU using the subword tokens labeled as this phoneme. The NTN trained for this phoneme can provide the ability to discriminate between the speaker and impostors. During testing, the utterances are first segmented by the concatenation of the subword models given by the prompted words. Then the subword units are applied to the corresponding phoneme-based NTN, the score is then calculated by the above equation. The speaker verification systems using PLU model have the advantage that the testing passwords are not restricted to fixed passwords. Hence, the security of the system is enhanced.

## 4. EXPERIMENTAL RESULTS

### 4.1. Testing the HMM State-based NTN

The database for evaluating the HMM state-based NTN is a subset of the TI 46-word database. This database is recorded in a clean environment, and sampled at a frequency of 12.5 kHz. It consists of utterances of ten digits spoken by 16 speakers (8 male, 8 female). The feature

vectors used in training and testing are the $28th$ order LP cepstral coefficients. The feature frames are extracted every 25 ms with 10 ms shift. Four experiments are performed to evaluate the text-dependent speaker verification systems, which are the multiple-word NTN, word-specific NTN, subword NTN, and HMM classifiers. In all experiments, eight utterances of all the digits in the training data set are concatenated to train the models, and two other utterances are used to determine the cohort for each speaker [6]. The 16 utterances of each digit in the testing data set which are recorded in 8 different sessions are used individually for testing. The result of this experiment is shown in Table 1. In the isolated digit speaker verification, the word-specific NTN performs better than multiple-word NTN, and the subword NTN performs even better than both of the above methods.

### 4.2. Testing the Phoneme-based NTN

The phoneme-based NTN classifier was evaluated by using the YOHO database [1]. The YOHO voice verification corpus was designed particularly for testing text-dependent speaker verification systems. It consists of 138 speakers enrolled (106 males, and 32 females). There are 4 enrollment sessions with 24 utterances each, and 10 verification sessions with 4 utterances each. The syntax used in the YOHO database incorporates "combination lock" phrases, and the phrases used for enrollment and verification are different. The utterances in YOHO are sampled at a rate of 8 kHz, and limited to a 3.8 kHz bandwidth. In the experiment, the speech signal is pre-emphasised using a first order digital filter. The feature vectors, $12th$ order MFCC's, are extracted from speech signal every 25 ms with 10 ms shift. In training the HMM and speech segmentation, the $\triangle$ and $\triangle^2$ MFCC are argumented to the feature vectors. In training the phoneme-based NTN, and testing, only the $12th$ order MFCC's are used. The PLU is modeled by a 3-state left-to-right HMM with no skip between states. A total of 20 phonemes were found enough to transcribe the spoken numbers in YOHO database. The scores are cohort normalization with a set of 5 closest cohort speakers, then compare with a global threshold. The testing is conducted in a set of 138 speakers. Those who in the cohort set of each speaker are excluded in testing, and there is no testing between speakers of different gender. The results in Table 2 show that the phoneme-based NTN performs better than the HMM.

## 5. SUMMARY

We have proposed a new classifier based on subword NTN's. The new classifier has been evaluated by the text-dependent speaker verification experiments, and it demonstrates the new method's effectiveness in improving the performance over both the conventional HMM and multiple-word NTN. Since the subword NTN classifier is trained with the discriminant error measure for each speaker, it is shown to provide better discriminant ability than the conventional HMM classifier. A subword NTN classifier also outperforms the multiple-word NTN, because it integrates the segmentation ability of HMM which can catch the temporal variation of speech and segment the speech into phonetically homogeneous data for NTN training.

| classifier | Error | |
|---|---|---|
| | FA | FR |
| multiple-word NTN | 1.06% | 7.89% |
| word-specific NTN | 0.36% | 5.82% |
| subword NTN (8 subwords/digit) | 0.17% | 3.28% |
| HMM (8 states/digit) | 1.44% | 11.72% |

Table 1: speaker verification Performance on TI DB

| classifier | Equal Error Rate | |
|---|---|---|
| | 4 utterances(10s) | 1 utterance(2.5s) |
| phoneme NTN | 0.36% | 0.76% |
| HMM | 1.66% | 4.02% |

Table 2: speaker verification Performance on YOHO DB

## 6. REFERENCES


[1] J.P. Campbell. Testing with the yoho cd-rom voice recognition corpus. In *Proceedings ICCASP*, 1995.

[2] G.R. Doddington. Speaker recognition - identifying people by their voices. *Proceedings of IEEE*, 73(11):1651–1664, Nov. 1985.

[3] K.R. Farrell, R.J. Mammone, and K.T. Assaleh. Speaker recognition using neural networks and conventional classifiers. *IEEE Tran. on Speech and Audio Proc.*, 2(1 part II):194–205, Jan. 1994.

[4] S. Katagiri and C.-H.Lee. A new hybrid algorithm for speech recognition based on hmm segmentation and learning vector quantization. *IEEE Tran. on Speech and Audio Proc.*, 1(4):421–430, October 1993.

[5] T. Matsui and S. Furui. Concatenated phoneme models for text-variable speaker recognition. In *Proceedings ICCASP*, pages 391–394, 1993.

[6] A.E. Rosenberg, J. DeLong, C-H. Lee, B-H. Juang, and F. Soong. The use of normalized score for speaker verification. In *Proc. of ICSLP*, October 1992.

[7] A.E. Rosenberg, C.-H. Lee, and F.K. Soong. Sub-word unit taker verification using hidden marcov models. In *Proc. ICASSP*, 1990.

[8] A. Sankar and R.J. Mammone. Growing and pruning nural tree networks. *IEEE Tran. on Comp.*, 42(3):1–9, 1993.