A NEURAL NETWORK USING ACOUSTIC SUB-WORD UNITS FOR CONTINUOUS SPEECH RECOGNITION

Ha-Jin Yu, Yung-Hwan Oh

Dept. of Computer Science, KAIST
(Korea Advanced Institute of Science and Technology)
373-1, Kusong-dong, Yusong-gu, Taejon Korea
hjyu@bulsai.kaist.ac.kr

ABSTRACT

A subword-based neural network model for continuous speech recognition is proposed. The system consists of three modules, and each module is composed of simple neural networks. The speech input is segmented into non-uniform units by the network in the first module. Non-uniform unit can model phoneme variations which spread for several phonemes and between words. The second module recognizes segmented units. The unit has stationary and transition parts, and the network is divided according to the two parts. The last module spots words by modeling temporal representation. The results of speaker independent word spotting of 520 words are described.

1. INTRODUCTION

Recently, although hidden Markov model (HMM) has been regarded as the most powerful model for speech recognition, neural networks are still under consideration because of their merits over other models such as higher parallelism and robustness in hardware faults. Most neural network models regard the units of speech as homogeneous static patterns, and they do not represent the internal structure of the units. For example, in models such as TDNN, recurrent network, and SOFM [1], the input speech is divided into frames of uniform length, and a phoneme or a word is represented by a series of frames. Since phonemes or words of human utterances are not of uniform acoustic structure, a network would be more efficient if it could model the acoustic structure of the units.

In this research, the speech input is first divided into non-uniform acoustic sub-word units [2] which can be divided into stationary and transition parts. The network structure is designed to cope with the structure of the unit. In continuous speech, the effect of co-articulation spreads for several phonemes and words, and some phonemes tend to be missed or weakened. For example, in TIMIT speech data, the transcription of the word "morning" is /m ao1 r n ix ng/, but in real speech data, one of the utterances is segmented and hand-labeled as /m ao en/. In that case, the last four phonemes are reduced to one phoneme /en/. Using non-uniform units in speech recognition, we can model such phoneme variations successfully.

The concept of a non-uniform unit is introduced as a speech synthesis unit by Sagisaka [3]. It is then used for speech recognition [4] [5]. In [5], the long-unit candidates are selected using text data by calculating the frequency of all combinations of neighboring phonemes, and the phoneme HMMs are concatenated and retrained. In our approach [4], the units are selected using speech data. The speech segments are selected as units, and they are cut in the middle of relatively stationary points where the values of spectral transition measure [6] are local minima. Since the segmentation is straightforward, it takes much less time to classify the units, and the recognizer can be simple. In this research, the recognizer was built by using simple structures of neural networks.

Our system consists of three modules. The input speech is segmented by the first module, and is classified by the second module. In previous work [2], we implemented the first two modules, and reported that the system can classify the sub-word units. In this research, a module is added to detect words from the result of sub-word unit recognition. The units are trained by the result of word detection, rather than the result of unit recognition itself.

2. STRUCTURE OF THE SYSTEM

The system is composed of three modules: segmentation, unit, and word modules, as shown in Figure 1. The segmentation module segments the input feature stream into acoustic sub-word units, and the unit module classifies the segmented units. The word module detects words from the recognized unit.

2.1. Segmentation and Unit Module

The segmentation module segments the continuous speech input into non-uniform units. Figure 2 shows the structure of the module and the example of the units it produces. It gets input from a time window of seven frames which is shifted by one frame. The first layer of the segmentation module is

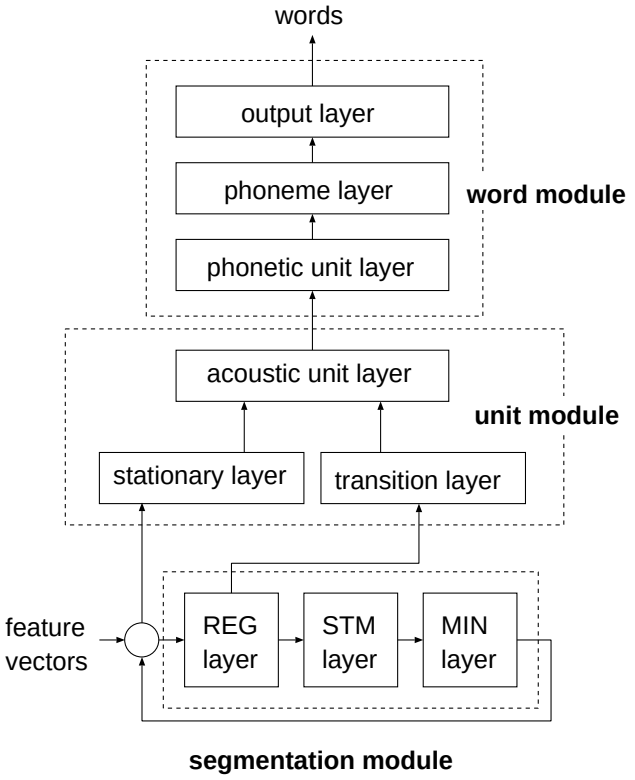

Figure 1: Structure of the system

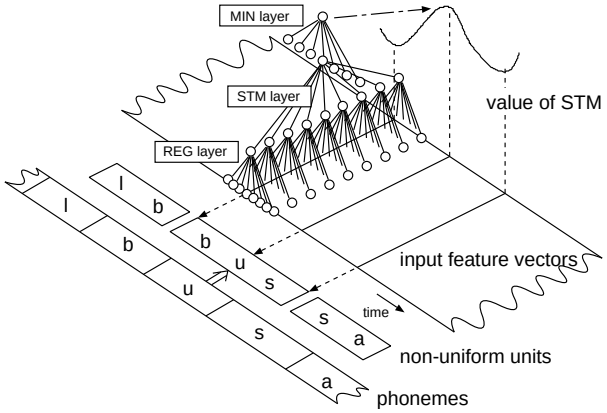

Figure 2: Segmentation module and unit definition

the REG layer, which has 14 output neurons. The output of the mth neuron at time t, $r_m(t)$ is

$$r_m(t) = \sum_{n=-N}^{N} x_m[t+n] \cdot w_n^r, \qquad 1 \le m \le p \qquad (1)$$

where the input $x_m[t+n]$ is the mth component of the mel-cepstral coefficient at time t+n, N = 3 which decides the size of the window, p = 14 is the analysis order, and the weights are fixed to w_n^r = n. The outputs of the REG layer are

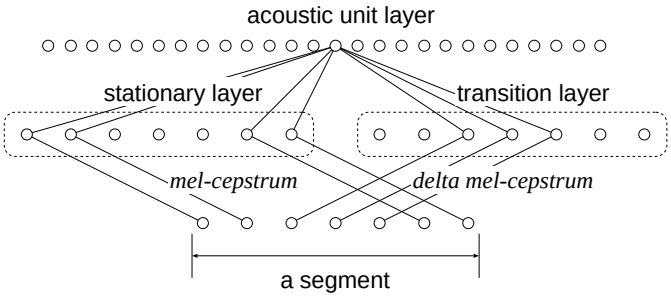

Figure 3: Unit module

regression coefficients of input parameters which represent the slope of the time function [6].

The second layer of the segmentation module is the STM layer, which has the output value at time t,

$$s(t) = \sum_{m=1}^{p} \mid r_m(t) \mid w_m^s \qquad (2)$$

where w_m^s is the weight between mth input neuron and the output neuron. The output value s(t) is the spectral transition measure [6] weighted to reduce the segmentation variation of the units [2]. A high value of the output implies that the spectrum changes rapidly at that point.

The third layer is MIN layer, which detects points where the outputs of the STM layer are local minima. The output value M_t at time t is

$$M_t = f_h(\theta_M - \sum_{n=-N}^{N} f_h(s(t+n) - s(t))) \qquad (3)$$

where the function $f_h()$ is the hard limiting nonlinearity, and $\theta_M = 2N + \alpha$, $0 < \alpha < 1$ is a threshold. The output is active when s(t) is the minimum of s(t+n), $-N < n < N$, N = 3, that is, the value of the weighted spectral transition measure at time t is the local minimum.

The input vector stream is cut out in the middle of the two points where the output of the MIN layer is active. This means that the input is segmented at the point where spectral change is relatively slow. The position can also be regarded as the stationary point of the input, and the region bounded by the points can be regarded as the transition part, since the spectrum changes rapidly between the two points. In the example shown in Figure 2, the segment includes three phonemes 'b', 'u', and 's', so a unit labeled as 'bus' is defined.

The units segmented by segmentation module are classified by the unit module shown in Figure 3. Seven vectors are selected from a segment. Two vectors are selected at the two stationary points which are the ends of the segment, one at the point where the value of the spectral transition measure is the peak, and the rest vectors between the center and the boundaries. The vectors close to the central point in time are

fed to the transition layer, and the vectors close to the ends of the segment are fed to the stationary net. The features for the stationary layer are mel-cepstral coefficients, and those for the transition layer are delta mel-cepstral coefficients, which are the output of the REG layer.

The activation A_j for jth output neuron of acoustic unit layer is

$$A_j = f_h(\sum_i \|\vec{w_{ij}^s} - \vec{m_i}\| + \sum_i \|\vec{w_{ij}^t} - \vec{d_i}\| - \theta_U) \quad (4)$$

where $\vec{w_{ij}^s}, \vec{w_{ij}^t}$ are the weight vectors between jth neuron of acoustic unit layer and ith neuron of stationary and transition layer, respectively. $\vec{m_i}, \vec{d_i}$ are ith vectors of mel-cepstrum and delta mel-cepstrum, respectively, in the segment. The neurons of acoustic unit layer are connected to those of the phonetic unit layer in the word module.

2.2. Word Module

The word module detects words from the output of the unit recognition module. Figure 4 shows the structure of the word detection module. The word net has four layers: the acoustic unit layer, phonetic unit layer, phoneme layer, and output layer, as shown in Figure 1. The first layer, the acoustic unit layer, is composed of neurons which correspond to acoustic units. The layer is also the output of the unit module. The second layer is the phonetic unit layer, and the links between the first two layers map acoustic units to phonetic units. The phonetic units are classified by the phonemes included in the units, and the acoustic units are discriminated by acoustic distances between the units. The mapping is needed because the words can be described by phonetic units, and the input signal can be easily transformed to acoustic units. An acoustic unit can be mapped to several phonetic units, and vice versa. Several neurons of the two sub-word unit layers can be activated from one input segment.

The neurons in the output layer correspond to words, which are the final goal of this system. A neuron in the output layer has its own set of neurons in the phoneme layer. The neurons in the phoneme layer are linked to only one neuron in the output layer, and the group of neurons represents phonemes that constitute a word. A neuron in the phoneme layer is linked to several neurons in the phonetic unit layer which include the corresponding phonemes. In Figure 4, the output neuron for the word "itemize" is linked to a set of six neurons in the phoneme layer, which represent the phonemes 'ay', 't', 'ax', 'm', 'ay', 'z', and the six neurons represent only the word.

A neuron in the phonetic unit layer is activated whenever one or more of the neurons in the acoustic unit layer which are linked to the neuron in the phonetic unit layer. The value of ith neuron $P_i^w[t]$ in the phoneme layer of a word w according to the input from the phonetic unit layer at time t is

$$P_i^w[t] = \begin{cases} \max(P_i^w[t-1], \max_k(U_k[t])) & \text{if } G_i^w = 1 \\ P_i^w[t-1] - d & \text{otherwise.} \end{cases} \quad (5)$$

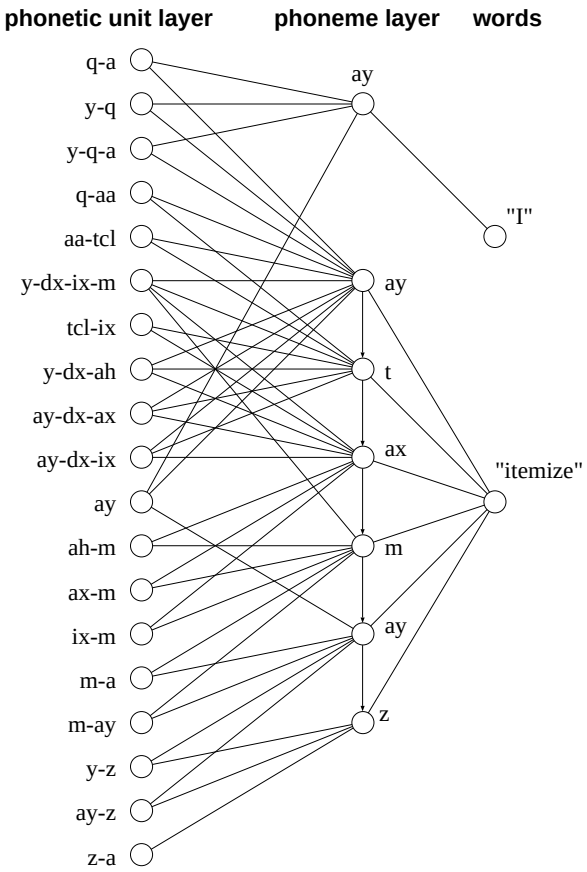

Figure 4: Word module

$$G_i^w = \begin{cases} 1, & \text{if } P_{i-1}^w > 0 \text{ or } i = 0, \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

where $P_i^w[t]$ is the value of ith neuron in the phoneme layer which is linked to wth neuron in the word layer at time t, $U_k[t]$ is the kth neuron in the phonetic unit layer at time t, where there exists a link between P_i^w and U_k, $1 \le i \le L_w$, L_w is the number of phonemes in word w, and d is a constant. The neurons in the phoneme layer are activated only when they are activated in order. Otherwise, the value of P_i^w is decreased by a constant d.

Finally, the output $O^w[t]$ which corresponds to word w of word module at time t is

$$O^w[t] = f_h(\sum_{i=1}^{L_w} P_i^w[t] - \theta_w) \quad (7)$$

where θ_w is the threshold for the word w, The number of phonemes in the word w is L_w, that is, the number of the neurons in the phonetic layer that linked to output neuron O^w. The activation of an output neuron means the detection of the word w.

3. TRAINING

The networks are trained by using the phoneme labeled continuous speech, after the whole process of recognizing words in a sentence is ended. If a word is not spotted in an appropriate region, the segments related with the word are used for training. If a segment yields a unit that has not been seen, a new neuron in each acoustic unit layer and phonetic unit layer is created, and the links between the layers are added. If there already exists a neuron for the unit, the links are updated by supervised Hebbian learning. If a new word is to be added, a neuron is added to the output layer, and neurons are added in the phoneme layers in correspondence with the number of phonemes in the word. Then links between the phoneme layer and phonetic unit layer are added, and the links between acoustic unit layer and phonetic unit layer are created or updated. The threshold θ_w is lowered if the detection rate of a word w is lower than that of other words.

4. EXPERIMENTS

The system is simulated on a SPARC-20 workstation. A subset of TIMIT data is used to test the system. In the data, the SX sentences are uttered by seven speakers. We select 150 sentences in the SX sentences, and select 520 keywords in the sentences. Six utterances for each sentence, a total of 900 sentences are used for training, and the remaining 150 sentences are used for testing, which include 593 keywords. A 16 msec Hamming window is applied to the speech every 8 msec, and 14th-order LPC cepstral coefficients are extracted. As a result, 5068 units are defined in training, and 59.3 % (2413) of the total 4077 units are classified correctly in tenth candidates. 10.1 % (411) of the units are unseen units, and 13.5 % (80) of the words are missed because of the unseen unit, when word detection rate is 77.2 % at fa/kw/hr of 23.6. That means 59 % of the error is due to the unseen units. The system can be trained with the unseen units for improvement. We also evaluated the system when input speech is segmented at points where the values of spectral transition measure are peaks rather than valleys, to show that using stationary points is effective. 5339 units are defined and 48.3 % of total 5710 units are classified correctly. Figure 5 shows the receiver operating characteristic (ROC) curve, for the two cases. The result shows that using the valley of the spectral transition can acquire about 150 % of word detection performance of that using peak. In the simulation, the units are pre-selected by using only the two vectors at the ends, since it takes too much time to evaluate all the activation values of the acoustic unit layer. The number of neurons that need full calculation is reduced to about 10 % of the total number of neurons in the acoustic unit layer, but the word detection rate is not changed.

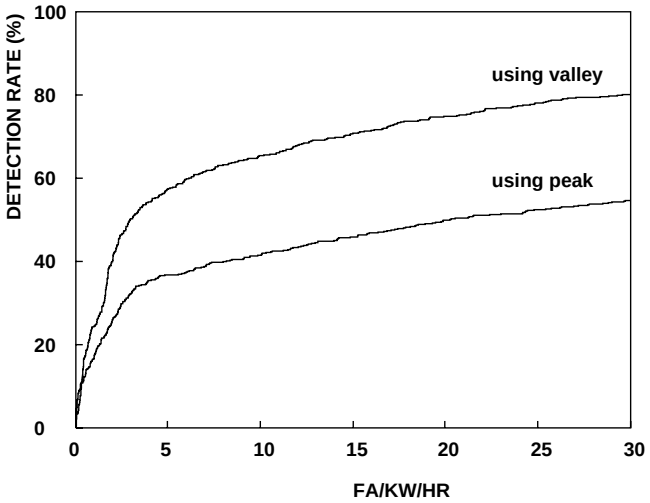

Figure 5: Word detection result (ROC curve)

5. CONCLUSIONS

We have introduced non-uniform units to a neural network model for continuous speech recognition. The system consists of three modules which are aggregations of simple neural networks. The first module segments the continuous speech input into non-uniform units, and the second module recognizes the units. The third module detects words in the sentence. Using non-uniform unit, the errors in segmentation will not be a critical problem, because we can define enough units to cover the acoustic variations. We conclude that using the non-uniform units we proposed is effective for spotting 520 keywords. The system is currently being compared with other systems, and an unsupervised training algorithm is being devised to train the system with more data.

6. REFERENCES

1. Morgan, D.P., and Scofield, C.L., *Neural Networks and Speech Processing*, Kluwer Academic Publishers, Boston, 1991
2. Yu, H.-J., and Oh, Y.-H., "A Neural Network using Non-Uniform Unit for Continuous Speech Recognition," EUROSPEECH'95, Madrid, Spain, Vol 3, pp. 1677-1680, September 1995
3. Sagisaka, Y., "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," Proceedings of International Conference on Acoustics, Speech, and Signal ICASSP '88, pp. 679-682
4. Yu, H.-J., and Oh, Y.-H., Yamashita, Y., and Mizoguchi, R., "Expert system for continuous speech recognition with non-uniform recognition unit," IEICE Technical Report, SP93-56, pp.57-64, August 1993
5. Matsumura, T. and Matsunaga, S., "Towards Non-uniform Unit HMMs for Speech Recognition," 2nd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications (IVTA94) pp. 89-92, 1994
6. Furui, S., "On the Role of Spectral Transition for Speech Perception," Journal of Acoustic Society of America 80(4), pp. 1016-1025, 1986