Pergamon

Pattern Recognition, Vol. 28, No. 6, pp. 795–805, 1995
Elsevier Science Ltd

0031–3203/95 $9.50 + .00

0031–3203(94)00145–6

SPEECH RECOGNITION WITH HIERARCHICAL RECURRENT NEURAL NETWORKS

WEN-YUAN CHEN*, YUAN-FU LIAO and SIN-HORNG CHEN
Department of Communication Engineering, National Chiao Tung University, Hsinchu, Taiwan, Republic of China

(*Received* 26 *October* 1993; *in revised form* 29 *October* 1994; *received for publication* 30 *November* 1994)

Abstract—A hierarchical recurrent neural network (HRNN) for speech recognition is presented. The HRNN is trained by a generalized probabilistic descent (GPD) algorithm. Consequently, the difficulty of empirically selecting an appropriate target function for training RNNs can be avoided. Results obtained in this study indicate the proposed HRNN has the advantages of being capable of absorbing the temporal variation of speech patterns as well as possessing effective discrimination capabilities. The scaling problem of RNNs is also greatly reduced. Additionally, a realization of the system using initial/final sub-syllable models for isolated Mandarin syllable recognition is also undertaken for verifying its effectiveness. The effectiveness of the proposed HRNN is confirmed by the experimental results.

Speech recognition Hierarchical Recurrent neural networks
Generalized probabilistic descent Discriminative training

1. INTRODUCTION

Speech is unmistakably produced by a slowly moving vocal tract. Features extracted from speech signals are therefore distributed over time in a complex manner. The temporally distributed features of speech signal make it difficult to recognize. The hidden Markov model (HMM) is a conventional method applied in coping with this difficulty. In HMMs, a speech pattern is treated as if it is the output of a stochastic system with its associated internal state being a process governed by probabilistic laws. In this sense, HMMs do not directly deal with time warping; however, they learn statistical distribution of a training set which contains time warped patterns.[1] Consequently, HMMs recognizing a test pattern which is far away from the statistical distribution of the training set would be rather unlikely. Conducting within-class training without competition with hostile classes is yet another vulnerability of HMMs; that is, they are trained with a maximum likelihood criterion. Hence, elements of confusing words are not emphasized as essential cues when distinguishing between them. Some criteria, e.g. the maximum mutual information, actually provide an approach of discriminative training for HMMs; however, these criteria require much hypotheses to be solved.

Multi-layer perceptrons (MLPs), on the other hand, possess effective discrimination capabilities through competitive training for static patterns to produce outputs which discriminate between the classes to recognize. Since speech signal is inherently dynamic, a network which accepts a stream of speech requires recurrent connections for maintaining a representation of previous cues. Unfortunately, the structure of MLPs being a feedforward network does not directly conform to the speech recognition because of no memory to represent the process of state transitions over time. Some approaches toward MLP-based speech recognition have been studied in recent years to solve the problem caused by temporal distortion in speech signal.[2–6] The time delayed neural network[3] (TDNN) and the dynamic programming neural network[4] (DNN) are two typical examples of these approaches. TDNN deals with the time-alignment problem through the mapping of a temporal variation of the speech signal into interconnections which are present between neurons of different delay periods. The cells of the TDNN integrate activities from adjacent time-delayed vectors, which allows each vector to be separately weighted in time. Even so, TDNN does not work well for recognizing dynamic speech signals since the local features from adjacent time-delayed vectors do not directly contribute to the final classification. The DNN applies the conventional dynamic programming[7] (DP) algorithm toward optimistically aligning the MLP input cells with the input utterance. The DP, although an efficient strategy of searching for the optimum path among all L^L possible paths, still requires $\sim O(L^2)$ operations,[1] where L is the frame length of the input. Additionally, one operation here represents all of the calculations involved in evaluating the score of one path. The DP is later demonstrated here as not being required whenever using the proposed HRNN. This exclusion is a dramatic saving in the computation power.

* Author to whom all correspondence should be addressed.

The conventional neural network architectural configurations have obtained a high recognition rate for small vocabulary speech recognition. If these configurations are extended for large vocabulary speech recognition, however, the training time becomes excessively large according to the scale of the network. Moreover, searching for an optimal solution in the solution space of a large network configuration becomes increasingly difficult and causes the network to fall into local minima instead of the global minimum. A large network also requires a substantial amount of training data; otherwise, the network simply memorizes the training samples, thereby resulting in a poor generalization. Some approaches based on hierarchical neural networks have already been studied in light of the fact that scaling up a conventional network for large vocabulary speech recognition is not realistic.[8-13] In hierarchical neural networks, each individual network is only faced with a partial task of solving a large problem in its entirety and, consequently, can be trained with less training samples. Additionally, the experimenter has the option of assigning subproblems to individual networks as well as structuring communication between networks in a manner that reflects knowledge of the domain.[14] Matsuoka *et al.*[12] proposed an integrated neural network, which consists of a control network and several sub-networks, to recognize 62 syllables. Hampshire II and Wiabel[10,11] proposed the Meta-Pi network consisting of a multi-network TDNN which performs multi-speaker phoneme discrimination (/b, d, g/). The Meta-Pi architecture is a multi-source connectionist pattern classifier that is comprised of a number of source dependent sub-networks that are integrated by a combinational superstructure. Hild and Waibel[9] improved multi-state TDNN (MS-TDNN) for speaker independent connected letter recognition by exploring network architectures with "internal speaker models". However, all of these systems are implemented in MLP-based neural networks which have inherited the drawbacks from MLPs. Some of these systems are only appropriate in dealing with static patterns. Others deal with the temporal problem using either the TDNN structure or the DP algorithm.

The architectural configuration of RNN is one in which the state of the network at any time depends on a complex aggregate of previous inputs. Consequently, RNN-based systems more easily catch dynamic information than those MLP-based systems. Nossair and Zahorian[15] demonstrated that features extracted from dynamic spectrum are superior to features extracted from the static spectrum. RNNs are therefore potentially suitable for recognizing speech patterns. However, two main limitations involving application of RNNs to speech recognition still remain unsolved and require further studies, i.e. the selection of appropriate target functions for training the network as well as its applicability toward large vocabulary speech recognition. The desired outputs of RNNs must be expressed as functions of time, as indicated in previous studies.[16,17] The setting of proper target functions over time is rather critical for RNNs because it determines how the network's weights can be updated. The setting target functions for RNNs is totally different from the approach of training a static network, e.g. an MLP, by setting fixed desired targets for a specific input pattern. The selection of appropriate time-dependent target functions is unfortunately still an empirical process. The limitation of applying an RNN to large vocabulary speech recognition is a scaling problem. The RNN should be sufficiently large to discriminate all of the word patterns in the vocabulary. However, the learning time and the number of weights required for accurately distinguishing all word patterns would grow exponentially and become unacceptable as the network becomes large.

A hierarchical RNN (HRNN) system for speech recognition is proposed in this study. Speech signal of an utterance is first phonetically divided into sub-word units. Each subword unit is then separately discriminated using an RNN. Next, a sequence of RNNs formed by serially cascading these basic RNN recognizers is utilized as the syllable recognizer. Besides, an additional RNN is employed in generating weighting functions for softly segmenting the input utterance as well as for unequally combining outputs of the sequential network. The segmentation of the input utterance is notably a soft one. This system is suitable for large-vocabulary speech recognition applications in light of the fact that sub-word units are adopted herein. Additionally, a novel training scheme based on a generalized probabilistic descent (GPD) algorithm[18] is also introduced for training the HRNN. The difficulties of empirically selecting appropriate target functions for training RNNs can be avoided by using the GPD competitive training algorithm. The proposed HRNN has the advantages of absorbing the temporal variation of speech patterns as well as possessing efficient discrimination capabilities.

This paper is organized as follows. The proposed HRNN system for speech recognition is presented in Section 2. A realization of the system based on initial/final sub-word models for Mandarin syllables recognition is also discussed. Performance of the system is examined by simulations discussed in Section 3. Conclusions are finally made in Section 4.

2. HIERARCHICAL RECURRENT NEURAL NETWORKS

2.1. *Recurrent neural networks*

The architectural configuration of the basic RNN used in this study is shown in Fig. 1. In the RNN, outputs of hidden units are delayed and fed back as supplementary inputs of the network. The activation function of output neuron k at time n is defined as

$$O_k(n) = \sum_j w_{kj} O_j(n) \tag{1}$$

where w_{kj} is the feedforward connection strength from hidden neuron j to output neuron k. Additionally, O_j

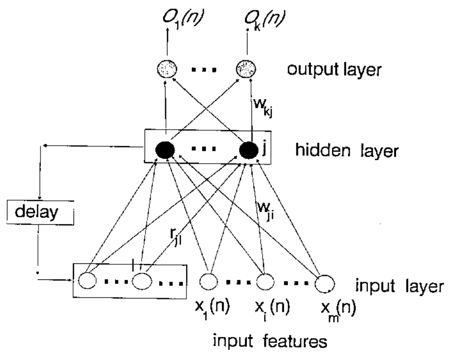


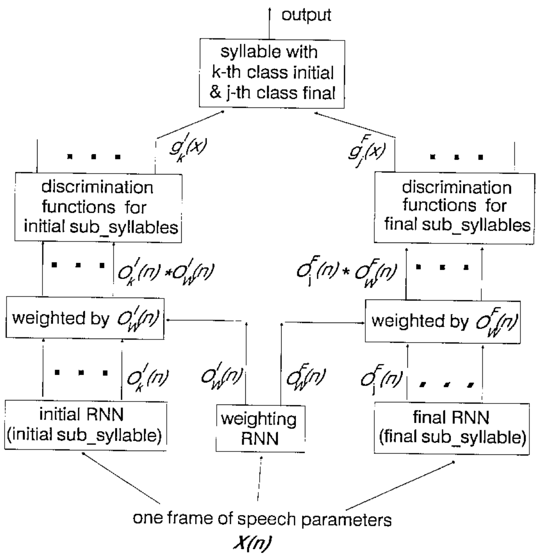
Fig. 1. The architectural configuration of a recurrent neural network. Previous outputs of hidden nodes are fed back to inputs. All nodes in the input layer are fully connected to the nodes in the hidden layer.

is the activation function of hidden neuron j, defined as

$$O_j(n) = \text{Sigmoid}(\text{net}_j(n)) \tag{2}$$

$$\text{net}_j(n) = \sum_i w_{ji}x_i(n) + \sum_l r_{jl}O_l(n-1)$$

where Sigmoid( ) is a sigmoidal function defined by $\text{Sigmoid}(x) = 1/(1+e^{-x})$; $O_l(n-1)$ is the activation value of hidden neuron l at time $n-1$; r_{jl} is the recurrent connection strength from hidden neuron l to hidden neuron j; and $x_i(n)$ is the input value of input neuron i at time n.



output
syllable with k-th class initial & j-th class final
$g^I_k(x)$
$g^F_j(x)$
discrimination functions for initial sub_syllables
discrimination functions for final sub_syllables
$O^I_k(n) * O^I_W(n)$
$O^F_j(n) * O^F_W(n)$
weighted by $O^I_W(n)$
weighted by $O^F_W(n)$
$O^I_k(n)$
$O^I_W(n)$
$O^F_W(n)$
$O^F_j(n)$
initial RNN (initial sub_syllable)
weighting RNN
final RNN (final sub_syllable)
one frame of speech parameters
$X(n)$

Fig. 2. The proposed hierarchical recurrent neural networks (HRNN) composed of three RNNs. The initial RNN and the final RNN are employed to distinguish initial sub-syllables and final sub-syllables, respectively. The weighting RNN is applied towards producing two weighting functions for segmenting the input utterance as well as unequally emphasizing the outputs of initial and final RNNs.

2.2. *The proposed HRNN for Mandarin speech recognition*

Each character in Mandarin speech is pronounced as a syllable. An isolated Mandarin syllable can be phonetically decomposed into initial and final sub-syllable units. Only 22 initials, including a dummy one, and 39 finals are available in Mandarin speech. The initial of a syllable is simply composed of a single consonant if it exists at all. As a result of the simple phonetic structure, all of the 408 Mandarin syllables form many confusing sets because many syllables have rather similar phoneme constituents. A HRNN speech recognition system is proposed here for discriminating between isolated Mandarin syllables. The HRNN is composed of a sequential network and a weighting RNN. The sequential network utilizes sub-syllables, i.e. initials and finals, as basic recognition units in this study. By decomposing each Mandarin syllable into an initial and a final, two separate RNNs are contained in the sequential network and employed so as to discriminate them, respectively. Also, the weighting RNN is applied towards producing two weighting functions for segmenting the input utterance as well as for unequally emphasizing the outputs of these two RNNs. By serially cascading these two weighted RNNs, a HRNN is formed and taken as the syllable recognizer. The block diagram of the proposed HRNN, as composed of two RNNs and a weighting RNN, is displayed in Fig. 2. Discriminant functions for initial and final sub-syllables can be calculated by

$$g^I_k(x) = \sum_{n=0}^{L-1} O^I_k(n)O^I_W(n), \quad \text{for initials} \tag{3}$$

$$g^F_j(x) = \sum_{n=0}^{L-1} O^F_j(n)O^F_W(n), \quad \text{for finals} \tag{4}$$

where L is the length of the input utterance; $O^I_k(n)$ and $O^F_j(n)$ are the kth output of the initial RNN and the jth output of the final RNN, respectively; and $O^I_w(n)$ and $O^F_w(n)$ are the weighting functions produced by the weighting RNN for the initial and the final RNNs, respectively. The final decision rule involves selecting candidates for the initial and the final with maximal discriminant functions. The input utterance is recognized as a syllable which has the kth class of initials if $g^I_k(x) > g^I_l(x)$ for all $l \neq k$ and the jth class of finals if $g^F_j(x) > g^F_l(x)$ for all $l \neq j$.

2.3. *Applying a GPD algorithm for training the HRNN*

Recently, Juang *et al.* have proposed a GPD algorithm for speech recognition.[18–20] The GPD algorithm, which is a systematic training algorithm employed for minimizing the recognition error rate, is adopted in this study for training the HRNN. The procedure of applying the GPD algorithm towards training the weights of the HRNN is stated as follows. Two misclassification measures for an input utterance x with the kth class of initial and the jth class of final are defined on the basis of the discriminant functions of equations (3) and (4):

$$d_k^I(x) = [-g_k^I(x) + g_p^I(x)]\frac{1}{L} \tag{5}$$

$$d_j^F(x) = [-g_j^F(x) + g_q^F(x)]\frac{1}{L} \tag{6}$$

where p and q are the most probable incorrect classes of initial and final, respectively. Next, a loss function $J(d)$ is defined for evaluating the costs of the current decisions for the initial and the final sub-syllables. The loss function should be a monotonically increasing, differentiable function. If a well approximating 0–1 cost function is used for $J(d), J^I = \sum_x J(d_k^I(x))$ and $J^F = \sum_x J(d_j^F(x))$ would approximately represent the total recognition errors of initials and finals, respectively. The sigmoidal function shown below

$$J(d) = \frac{1}{1+e^{-vd}} \tag{7}$$

is selected in this study as the loss function $J(d)$, where v is a scalar to control the rate of adjustment. The loss function $J(d)$ would clearly induce the training algorithm for emphasizing those utterances which are located at a short distance from the decision boundary; in addition, the scalar v serves to scale that distance. The objective of the GPD algorithm involves recursively adjusting the weights of the HRNN so as to achieve a minimum of J^I and J^F. The amount of weight change in the HRNN can be expressed through the GPD algorithm as

$$\Delta W^I = -\eta_n \frac{\partial J(d_k^I(x))}{\partial W^I}, \quad \text{for the initial RNN} \tag{8}$$

$$\Delta W^F = -\eta_n \frac{\partial J(d_j^F(x))}{\partial W^F}, \quad \text{for the final RNN} \tag{9}$$

$$\Delta W^W = -\eta_n \frac{\partial (J(d_k^I(x)) + J(d_j^F(x)))}{\partial W^W}, \quad \text{for the weighting RNN} \tag{10}$$

where η_n is the learning rate at the nth iteration. The scheme of selecting a proper learning rate can be found in Komori and Katagiri's study.[21] These three derivative terms can actually be computed via application of the chain rule.[22] The training of the HRNN can be accomplished by applying a bootstrap strategy. The training procedure is described as follows:

Step 1—randomly initialize all of the weights of these RNNs to small values;

Step 2—segment each training utterance into two parts, i.e. an initial and a final, by using the generalized minimum distortion segmentation (GMDS) method.[23] Modify these two segments such that they are overlapped by several frames. Data in both segments are used in training the weighting RNN by the standard error back propagation (EBP) training algorithm.[22] The target for the output node associated with initial (final) weighting function is set to 1.0 when the input signal lies in the initial (final) segment, and 0.0 otherwise. This procedure is repeated for several passes of the training set until it converges;

Step 3—maintain all of the weights of the weighting RNN fixed. Next, apply the GPD algorithm towards training the initial and the final RNNs by using equations (8) and (9);

Step 4—keep all of the weights of the initial and the final RNNs fixed. Next, retrain the weighting RNN by the GPD algorithm using equation (10); and

Step 5—steps 3 and 4 are iterated until a convergence is reached.

3. SIMULATIONS

3.1. *Database*

The performance of the proposed speech recognition method was examined by simulations on a multi-speaker speech recognition task. A database[24] containing utterances of 54 confusable Mandarin monosyllables with the first tone was employed in the test. These 54 syllables are the set of all possible combinations of 22 initials and four finals including /en/, /eng/, /in/, and /ing/. Table 1 lists these 54 syllables with a numerical label set to indicate a legal combination of initial and final. Each of these 54 syllables was pronounced three times by seven males and four females. Two repetitions of each speaker were used for training and the remaining one repetition for testing. One other speaker, a female, uttered each syllable 13 times, i.e. 10 times for training and three times for testing. There were a total of 1728 training utterances and 756 testing utterances. All speech signals in the database were digitized into 16-bit data format at a rate of 8 kHz and pre-emphasized with a high-pass filter, $1-0.98\,z^{-1}$. Next, a short-time spectral analysis by 256-point FFT was performed over every 32 ms Hamming-windowed frame with 8 ms frame shift. A bank of filters (in mel-scale) was then implemented to extract 12 log-compressed energies from the spectrum of each frame. 12 delta log-compressed energies were also calculated for each frame. The recognition features include these 24 parameters.

Table 1. 54 Mandarin monosyllables with the first tone. These 54 syllables are the set of all possible combinations of 22 initials and four finals

		b	*d*	*p*	*t*	*m*	*n*	*l*	*j*(*i*)	*ch*(*i*)	*s*(*i*)
en	1	2		3		4	5				
eng		6	7	8	9	10	11	12			
in	13	14		15		16	17	18	19	20	21
ing	22	23	24	25	26	27	28	29	30	31	32
	j	*ch*	*sh*	*dz*	*ts*	*s*	*h*	*f*	*g*	*k*	*r*
en	33	34	35	36	37	38	39	40	41	42	43
eng	44	45	46	47	48	49	50	51	52	53	54

The phonetic symbols are in the Yale system.

3.2. *Experimental results and analysis*

The effectiveness of the proposed HRNN speech recognizer was next examined by using the database. As mentioned previously, the HRNN is comprised of a sequential network and a weighting RNN. The sequential network is composed of two RNNs, one for initial sub-syllables and the other for final sub-syllables. The number of output nodes was set to 22 for the initial RNN, and four for the final RNN. Two output nodes were used for the weighting RNN so as to produce two weighting functions for initial and final sub-syllables, respectively. The number of hidden units was empirically selected to be 48 for the initial RNN, and 24 for the other two RNNs. Input features for all three RNNs consisted of 12 log-compressed energies and 12 delta log-compressed energies generated from frame-based spectra. Two other parameters were also set for training the HRNN. The learning rate η_n in equations (8)–(10) was initially set to 0.1 and then linearly decayed with time. The scalar v of the loss function used in the GPD training algorithm was set to 1.

Table 2 summarizes the recognition results attained by the proposed HRNN recognizer. A recognition rate of 73.5% was achieved. Notably, the values shown in parentheses indicate the numbers of correct classifications out of 756 testing utterances. For performance comparison, the continuous density hidden Markov model (CDHMM) method was also tested. Each syllable used in the CDHMM method was modelled by a six-state left-to-right network with a single transition. The observation features in each state were modelled by a five-mixture Gaussian distribution. As shown in Table 2, the recognition rate achieved by the CDHMM method was 67.2%. Obviously the proposed method performed much better in the study than the CDHMM method.

Detailed analyses of the HRNN are worthwhile since a better understanding regarding its behaviors can be obtained and may be conducive for further improvement in future studies. The learning curves of the HRNN for training data were first scrutinized. Figure 3 shows the learning curves of recognition rates for initial sub-syllables, final sub-syllables, and syllables. The weighting RNN had been initially trained with the EBP algorithm discussed in Steps 1 and 2 of the HRNN training procedure described in Section 2.3. This figure displays the training results of the GPD algorithm which alternately executes Steps 3 and 4 of the HRNN training procedure. This same figure reveals that all three learning curves increase gradually as the training procedure progresses until the 300th iteration. They all fall abruptly at the 301th iteration and then rise again thereafter. This phenomenon is accounted for as follows. At the end of the 300th iteration, one complete training process from Step 1 to Step 4 had just been performed for the HRNN. Both initial and final RNNs had been properly trained in Step 3 with the guidance of the initially-trained weighting RNN. Furthermore, the weighting RNN had been updated in Step 4. The training procedure then jumped to Step 3 at the 301th iteration to retrain both initial and final RNNs with the guidance of the updated weighting RNN. Due to the fact that the effective initial/final boundaries of many syllables determined by the old weighting RNN were inaccurate and had been corrected by the updated weighting RNN, a portion of training data which was previously guided by the old weighting RNN to train the initial (final) RNN was switched to train the final (initial) RNN with the guidance of the updated weighting RNN. This effect causes the abrupt fall off these three learning curves at the 301th iteration. Fortunately, the updated weighting RNN performed better than the original one so as to guide the retraining process in a correct direction. All three learning curves went up as the training procedure was continued and converged to better recognition rates.

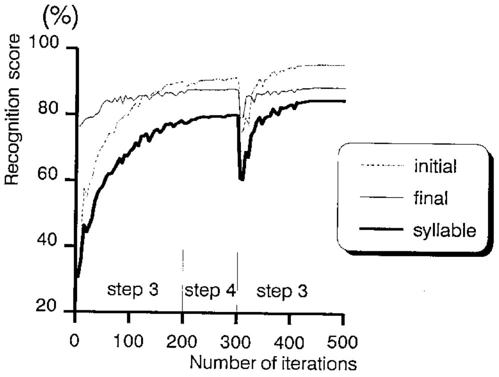


Fig. 3. Learning rates of the initial sub-syllable, the final sub-syllables, and the syllable corresponding to the number of learning iterations. The three RNNs had initially been trained by using the EBP training algorithm. Next, the training on the HRNN was directed by interlacing the training procedure of Step 3 or Step 4.

Table 2. The recognition results of 54 confusable Mandarin monosyllables. The figures in parentheses indicate the number of correct classification. The total utterance for testing is 756

	Initial	Final	Syllable
CDHMM			67.2% (508)
RNNs based	88.8% (671)	81.1% (613)	73.5% (556)

Next, the effectiveness of the weighting RNN in assisting the discrimination of confusing syllables and on softly segmenting the test utterance was examined by observing the weighting functions it produced. An example is shown in Fig. 4. The spectrogram of the utterance /sh-ēn/ is first displayed in Fig. 4(a). A spectrogram is a three-dimensional pattern showing the magnitude spectrum on grey-level display with time and frequency taken as the horizontal and the vertical axes, respectively. This figure indicates that the genuine initial-final boundary located at the 16th frame was accurately detected by the GMDS method. Figure 4(b)

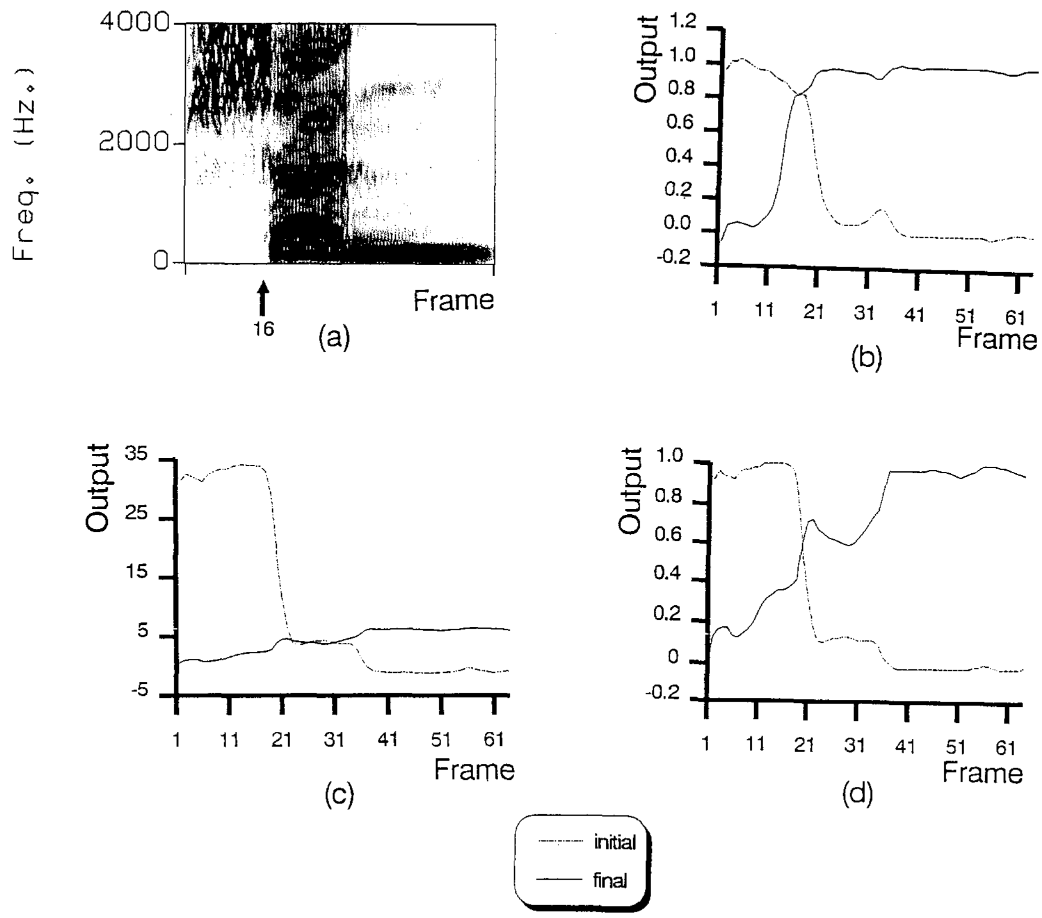


Fig. 4. Analysis of the weighting RNN. (a) the spectrogram of an input utterance /sh-ēn/ and the initial-final boundary segmented by a GMDS method is at the 16th frame. Outputs of the weighting RNN: (b) result of an initial training procedure by using the EBP training algorithm, (c) result of the complete training procedure by using the proposed GPD training algorithm, (d) normalized result of (c) such that peak values of two functions are equal to 1.

displays the weighting functions generated by the weighting RNN trained in Step 2 of the HRNN training procedure with targets set on the basis of the segmentation result from the GMDS method. This figure reveals that these two functions can be approximately regarded as 1–0 and 0–1 step functions to indicate the initial and the final parts of the utterance, respectively. Figure 4(c) displays the weighting functions generated by the updated weighting RNN trained in Step 4 of the HRNN training algorithm. This same figure reveals that both of these two functions give different weights to the three acoustic events of the utterance. Additionally, the initial weighting function gives a very heavy weight to the initial, light weight to the vowel, and a very light weight to the nasal. On the contrary, the final weighting function gives a heavy weight to the nasal, light weight to the vowel, and a very light weight to the initial. The overall effects of these two weighting functions specially emphasize the initial part of the input signal as well as emphasize the nasal part of the input signal. According to the phonetic structure of Mandarin syllable, initial and nasal are the two most important parts to distinguish these 54 syllables. Therefore, using these two weighting functions can assist the HRNN syllable recognizer in increasing its discrimination capability. To verify the effectiveness of the updated weighting RNN in segmenting the input utterance, the weighting functions displayed in Fig. 4(c) are normalized with peak values set to 1. Figure 4(d) shows these two normalized weighting functions. This figure indicates that the initial-final boundary can still be correctly detected. Another example using an utterance of /b-ēn/ is shown in Fig. 5. Similar results have been found from the figure except for that the initial/final boundary determined by the old weighting RNN is an incorrect one. This would result not only in providing unsuitable weighting functions to the initial and the final sub-syllables for final decision, but also in misguiding the training of the initial and the final RNNs. Fortunately, as shown in Fig. 5(d), the mis-segmentation had been corrected in the updated weighting function. Actually, in the training process, many mis-segmentations occurred for utterances with voiced plosive initials having very short durations had been corrected by the updated weighting

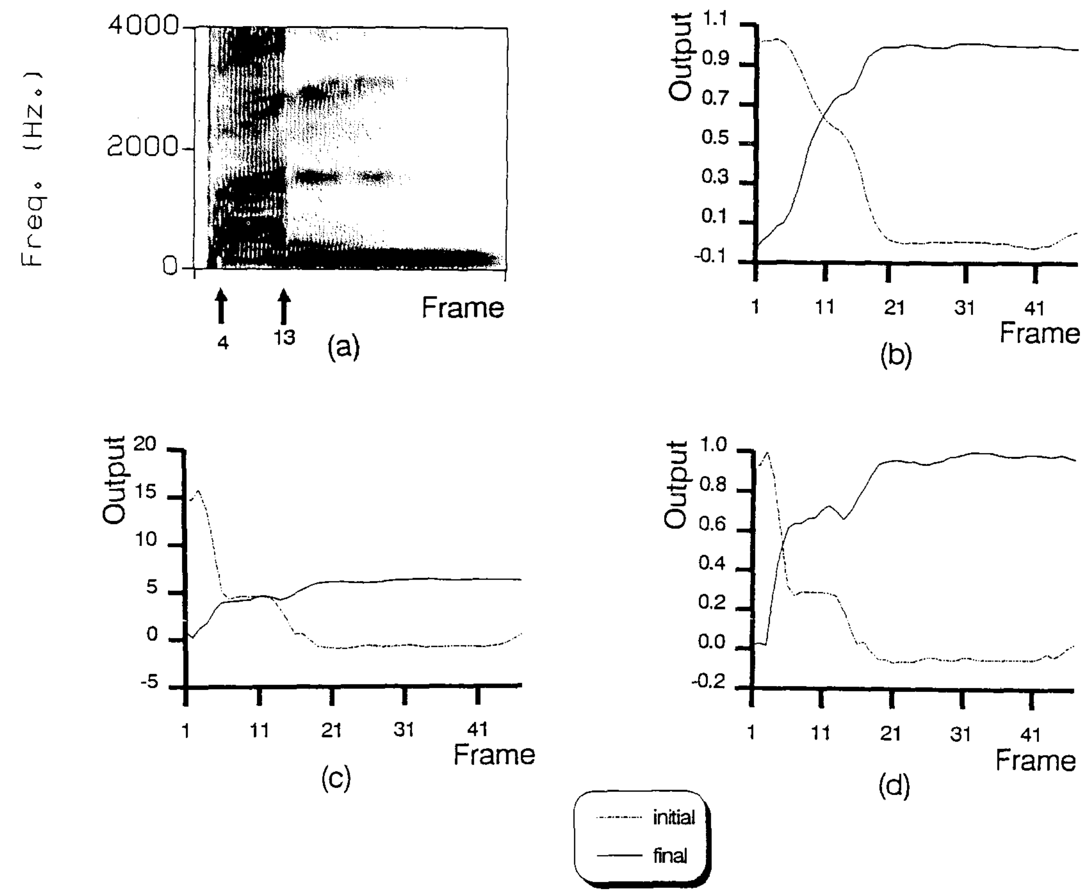


Fig. 5. Analysis of the weighting RNN. (a) the spectrogram of an input utterance /b-ēn/ and the boundaries segmented by the GMDS method and a manual vision are located at the 13th and fourth frames, respectively. Outputs of the weighting RNN: (b) result of an initial training procedure by using an EBP training algorithm, (c) result of the complete training procedure by using the proposed training procedure, (d) normalized result of (c) such that peak values of two functions are equal to 1.

RNN. The benefit of correcting mis-segmentations is two-fold. One advantage is to provide correct and appropriate weighting functions to unequally emphasize different parts of input signal for discriminating confusing syllables. The other is to correctly guide the training for both the initial and the final RNNs. Some utterances which were previously mis-classified can therefore be correctly recognized.

An examination is then made of the effect of combining the outputs of the initial and the final RNNs with the weighting functions generated by the weighting RNN for syllable recognition. An example to recognize an utterance of /sh-ēn/ is shown in Fig. 6. Weighted scores for initial and final sub-syllables are obtained by multiplying the initial and the final weighting functions to the corresponding outputs of the initial and final RNNs. Figure 6(a) and (b) display the weighted scores for the best four initial sub-syllables and for the four final sub-syllables, respectively. Their cumulative weighted scores are displayed in Fig. 6(c) and (d). Figure 6(a) and (c) indicate that the final part of the input utterance starting from the 20th frame to the ending frame makes almost no contribution to the cumulative weighted scores of initials. Similarly, as shown in Fig. 6(b) and (d), the contributions from the initial part to the cumulative weighted scores of finals are negligible. The discriminant functions of syllables for final recognition decision are calculated by simply combining the corresponding cumulative weighted scores of initial and final sub-syllables. In this example, the utterance was correctly recognized. Figure 7 provides yet another example which recognizes an utterance of /bēn/. Similar results can also be found in this figure.

From above analyses, we can conclude that a well-trained weighting RNN is capable of generating proper weighting functions to unequally emphasize acoustic events relevant to speech discrimination as well as softly segment the input utterance. Finally, no dynamic programming is notably required to be performed in the HRNN to optimally map the input testing utterance to the sequential network. Hence, the recognition process can be made more efficiently.

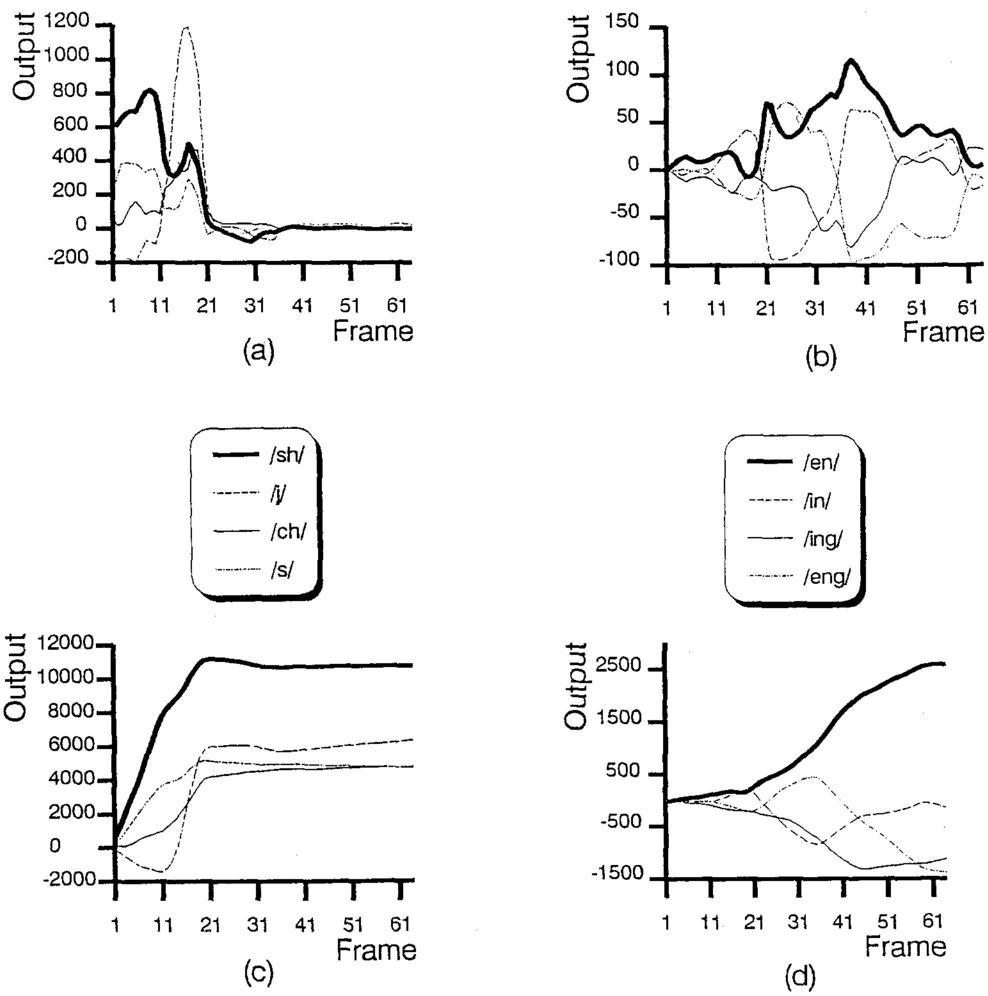


Fig. 6. For an input utterance /sh-ēn/, the best four weighted outputs of the initial RNN, (b) weighted outputs of the final RNN, (c) the best four discrimination functions for initials, (d) discrimination functions for finals.

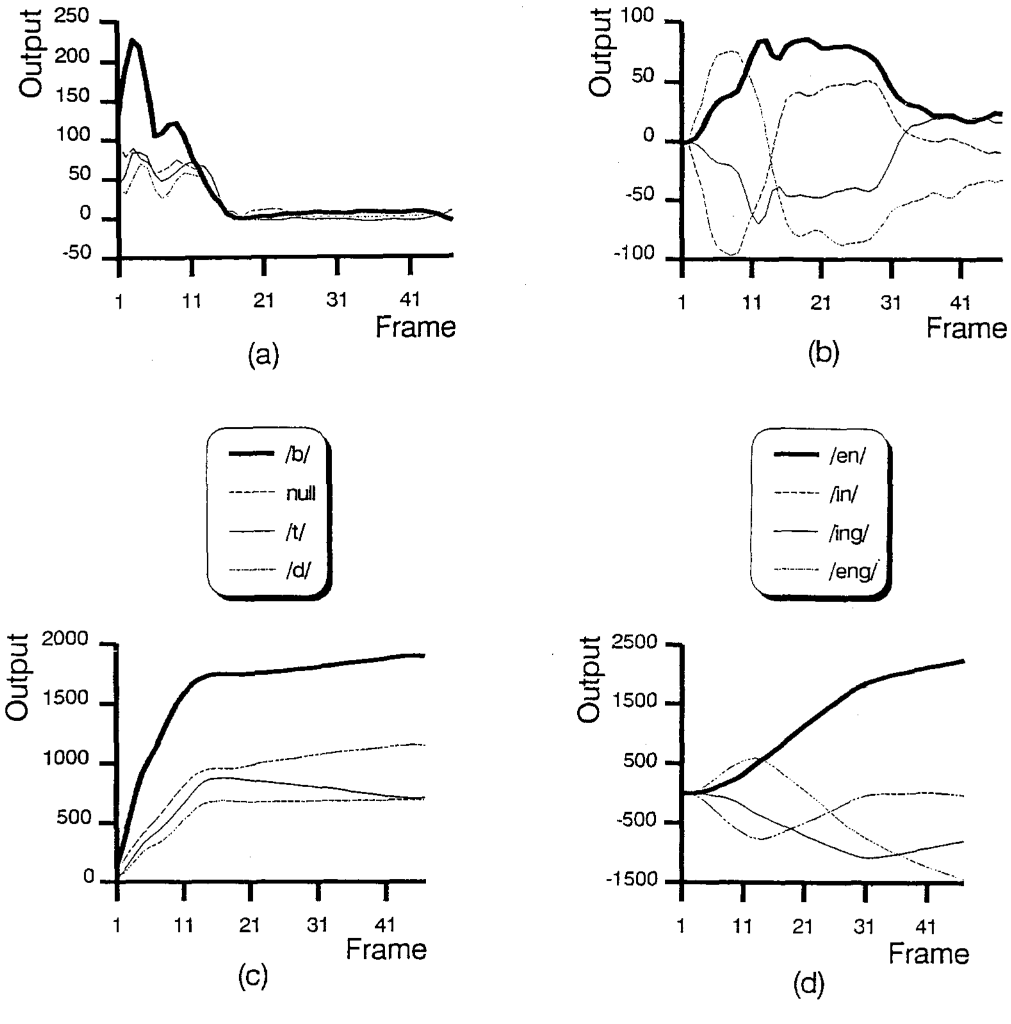


Fig. 7. For an input utterance /b-ēn/, (a) the best four weighted outputs of the initial RNN, (b) weighted outputs of the final RNN, (c) the best four discrimination functions for the initial RNN, (d) discrimination functions for the final RNN.

4. CONCLUSIONS

A novel hierarchical RNN-based approach for speech recognition was proposed in this study. The proposed method has successfully solved the time-alignment problem and retains the merit of competitive learning of artificial neural networks via using an HRNN network with a discriminative training algorithm. Besides, the application of sub-syllable recognition units has made it potentially suitable for scaling up to a large vocabulary application. Validity of the proposed approach was confirmed via simulations on a multi-speaker speech recognition of recognizing 54 confusable Mandarin syllables. Experimental results confirmed that the approach outperforms the CDHMM method. Extending this approach towards application of isolated speech recognition for all 408 Mandarin syllables would be a worthwhile task in future research efforts.

Acknowledgement—This authors wish to thank Telecommunication Labs, Ministry of Transportation and Communications, Republic of China, for their support of the database. They also wish to thank the anonymous reviewers for their insightful comments of the paper.

REFERENCES

1. G. Z. Sun, H. H. Chen, Y. C. Lee and Y. D. Liu, Time warping recurrent neural networks, *Proc. Int. Jt Conf. Neural Networks (IJCNN)*, Vol. I, 431–436 (1992).
2. W. Y. Chen and S. H. Chen, Word recognition based on the combination of a sequential neural network and the GPDM discriminative training algorithm, *Proc. IEEE Neural Networks for Signal Processing (NNSP)*. pp. 376–384 (1991).
3. K. J. Lang and A. H. Waibel, A time-delay neural network architecture for isolated word recognition, *Neural Network* **3**, 23–43 (1990).
4. H. Sakoe, R. Isotani, K. Yoshida, K. Iso, and T. Watanabe, Speaker-independent word recognition using dynamic programming neural networks, *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 29–32 (1989).
5. K. I. Iso, Speech recognition using dynamical model of speech production, *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Vol. II, pp. 283–286 (1993).
6. J. Tebelskis, Performance through consistency: connectionist large vocabulary continuous speech recognition, *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Vol. II, pp. 259–262 (1993).
7. H. Sakoe and S. Chiba, Dynamic programming algorithm optimization for spoken word recognition, *IEEE Trans. Acoust. Speech and Signal Process.* **26**, 43–49 (1978).
8. Y. Chen, B. Yuan and B. Lin, Real time Chinese syllable recognition with hierarchically structured neural network and transputer system, *Proc. Int. Jt Conf. Neural Networks (IJCNN)*, Vol. IV, pp. 743–748 (1992).
9. H. Hild and A. Waibel, Multi-speaker/speaker independent architectures for the multi-state time delay neural network, *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Vol. II, pp. 255–258 (1993).
10. J. B. Hampshire II and A. H. Waibel, The meta-pi network: connectionist rapid adaption for high performance multi-speaker recognition, *Proc. IEEE Intern. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 165–168 (1990).
11. J. B. Hampshire II and A. Waibel, The meta-pi network: building distributed knowledge representations for robust multisource pattern recognition, *IEEE Trans. Pattern Analy. Mach. Intell.* **14**, 751–769 (1992).
12. T. Matsuoka, H. Hamada and R. Nakatsu, Syllable recognition using integrated neural networks, *Proc. Int. Jt Conf. Neural Networks (IJCNN)*, Vol. I, pp. 251–258 (1990).
13. A. Waibel, H. Sawai and K. Shikano, Modularity and scaling in large phonemic neural networks, *IEEE Trans. Acoust. Speech Signal Process.* **37**. 1888–1898 (1989).
14. R. A. Jacobs, Initial experiments on constructing domains of expertise and hierarchies in connectionist systems, *Proc. Connectionist Models Summer School*, San Mateo, CA, pp. 144–153 (1988).
15. Z. B. Nossair and S. A. Zahorian, Dynamic spectral shape features as acoustic correlates for initial stop consonants, *J. Acoust. Soc. Am.* **89**(6), 2978–2991 (1991).
16. S. J. Lee, K. C. Kim, H. Yoon and J. W. Cho, Application of fully recurrent neural networks for speech recognition. *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 77–80 (1991).
17. G. Kuhn, Connected recognition with a recurrent network, *Speech Commun.*, **9**, 41–48 (1990).
18. S. Katagiri, C. H. Lee and B. H. Juang, New discriminative training algorithms based on the generalized probabilistic descent method, *Proc. IEEE Neural Networks for Signal Processing (NNSP)*, pp. 299–308 (1991).
19. P. C. Chang and B. H. Juang, Discriminative training of dynamic programming based speech recognizes, *IEEE Trans. Speech Audio Process.* **1**(2), 135–143 (1993).
20. B. H. Juang and S. Katagiri, Discriminative training, *J. Acoust. Soc. Jpn (E)* **13**(6), 333–339 (1992).
21. T. Komori and S. Katagiri, GPD training of dynamic programming-based speech recognizers, *J. Acoust. Soc. Jpn (E)* **13**(6), 341–349 (1992).
22. D. E. Rumelhart, G. E. Hinton and R. J. Williams, Learning internal representation by error propagation, in *Parallel Distributed Processing: Exploration in the Microstructure of Cognition.* The MIT Press, London (1986).
23. S. H. Chen and Y. R. Wang, Vector quantization of pitch information in Mandarin speech, *IEEE Trans. Comm.* **38**(9), 1317–1320, (1990).
24. J. S. Liou, R. G. Chen, S. M. Yu, J. R. Hwang and I. C. Jou, The speech database of Telecommunication Laboratories, Ministry of Transportation and Communications, R.O.C., *Proc. of Telecommunications Symp.*, Taiwan, pp. 128–132 (1990).

About the Author—WEN-YUAN CHEN received the B.S. and M.S. degrees from the National Taiwan Institute of Technology, Taipei, Taiwan, and the National Tsing Hua University, Hsinchu, Taiwan, in 1985 and 1987, respectively, both in electrical engineering. Currently he is working toward a Ph.D. in electronic engineering at National Chiao Tung University, Hsinchu, Taiwan. Since 1987, he has been with Industrial Technology Research Institute, Hsinchu, Taiwan where he is involved in research work on speech recognition and audio signal processing. His current interests include speech recognition, pattern recognition and audio signal processing.

About the Author—YUAN-FU LIAO received the B.S. and M.S. degrees in 1991 and 1993, respectively, and has been a Ph.D. student since 1993, all in the communication engineering of National Chiao Tung University, Taiwan. His current research interests include speech recognition, neural networks for signal processing.

About the Author—SIN-HORNG CHEN received the B.S. degree in communication engineering and the M.S. degree in electronics engineering from National Chiao Tung University, Hsinchu, Taiwan, Republic of China, in 1976 and 1978, respectively, and the Ph.D. degree in electrical engineering from Texas Tech University, Lubbock, in 1983. Form 1978 to 1980, he was an Assistant Engineering for Telecommunication Laboratories, Taiwan. He became an Associate Professor at the Department of Communication Engineering, National Chiao Tung University in August 1983, and a professor in August 1990. He also became the Chairman from August 1985 to June 1988, and from October 1991 to August 1993. He is currently doing research in the areas of digital communication and speech processing, specially concentrating on the problems of Mandarin speech recognition and text-to-speech.