Noise-Robust Speech Recognition Using Top-Down Selective Attention With an HMM Classifier

Chang-Hoon Lee and Soo-Young Lee, *Member, IEEE*

***Abstract*—For noise-robust speech recognition, we incorporated a top-down attention mechanism into a hidden Markov model classifier with Mel-frequency cepstral coefficient features. The attention filter was introduced at the outputs of the Mel-scale filterbank and adjusted to maximize the log-likelihood of the attended features with the attended class. A low-complexity constraint was proposed to prevent the attention filter from over-fitting, and a confidence measure was introduced on the attention. A classification was made to the class with the maximum confidence measure, and demonstrated 54% and 68% reduction of the false recognition rate with 15- and 20-dB signal-to-noise ratio, respectively.**

***Index Terms*—Hidden Markov model (HMM), selective attention, speech recognition.**

I. Introduction

ALTHOUGH noise-robust feature extractions based on auditory models have demonstrated improvements in the recognition performance [1], current speech recognition systems continue to require significant performance improvements in order to be utilized in practical applications in noisy real-world environments. One of the approaches to improve noise-robustness is to model top-down selective attention in higher brain functions.

Human beings utilize top-down selective attention from pre-acquired knowledge to improve the confidence level of recognition when encountering confusing patterns. Broadbent introduced the "early filtering" theory, in which the brain temporarily retains information about all stimuli but the information fades soon, unless attention is turned quickly to a particular memory trace [2]. Treisman later modified Broadbent's model and proposed that the filter merely attenuates the input rather than completely eliminating it [3]. The "spot light" attention model puts the filter at a small region of the input space, while Fukushima's Neocognitron model results in one adaptive attention filter at each input pixel and intermediate complex features for binary pattern recognition tasks [4].

Recently, top-down selective attention, which utilizes the multilayer perceptron (MLP) for pre-acquired knowledge has shown not only good recognition rates but also improved out-of-vocabulary rejection performance [5]. Moreover, it has been reported that the retraining independent component analysis (ICA) based on selective attention associated with MLP can result in a more accurate de-mixing matrix [6]. However, these models utilize MLP for pre-acquired knowledge in the higher brain whereas a hidden Markov model (HMM) is more widely used in speech recognition.

In this letter, the top-down selective attention model is extended to HMM classifiers. In addition, a low-complexity restriction is imposed on the attention filter in order to avoid over-fitting, and a new confidence measure is introduced for the classification.

II. Top-Down Selective Attention Model With HMM Classifiers

The proposed speech recognition algorithm using top-down selective attention is summarized as follows.

Step 1: Train HMMs from the training data and obtain HMM parameters λ_i's for each class.

Step 2: For a given testing speech, calculate the log-likelihood of each HMM.

Step 3: For the top N_C candidate classes.

1) Calculate the expected input and the attended log-likelihood for the class by the top-down attention process.

2) Calculate the confidence measure of the class.

Step 4: Choose the class with the maximum confidence measure.

Since the characteristics of a class are modeled as an HMM, the top-down selective attention algorithm at Step 3 estimates the expected input $\widetilde{\mathbf{x}}_i$ from the test speech $\mathbf{x}$ for the pre-trained attended class model λ_i as

$$\widetilde{\mathbf{x}}_i = \arg\max_{\mathbf{x}} P(\mathbf{x}|\lambda_i) \tag{1}$$

within a reasonable proximity from the original input $\mathbf{x}$.

Fig. 1 shows the signal flows with (solid line) and without (dashed line) the top-down attention for the popular MFCC features. In the absence of the top-down attention, the Mel-filterbank outputs go through a logarithmic nonlinearity and discrete cosine transform (DCT) to form the MFCC. With the top-down attention, a multiplicative attention filter may be placed at the Mel-filterbank output, which is equivalent to an additive adjustment at the log power as

$$\hat{b}_{tf} = b_{tf} + a_{tf}, \quad \begin{matrix} 1 \le t \le T \\ 1 \le f \le F \end{matrix} \tag{2}$$

Manuscript received May 15, 2006; revised October 16, 2006. This work was supported through the Brain NeuroInformatics Research Program by the Korean Ministry of Commerce, Industry, and Energy. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Steve Renais.

C.-H. Lee is with the Brain Science Research Center, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea (e-mail: chlee@neuron.kaist.ac.kr).

S. -Y. Lee is with the Department of BioSystems, Department of Electrical Engineering and Computer Science, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea (e-mail: sylee@kaist.ac.kr).

Digital Object Identifier 10.1109/LSP.2006.891326

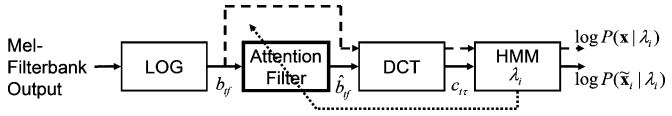

Fig. 1. Classification model with top-down selective attention filter.

where $a_{tf} \equiv \ln A_{tf}$ is the logarithm of the multiplicative attention filter A_{tf}, and b_{tf} and $\hat{b}_{tf}$ are the logarithm of the filterbank power of the actual and expected input, respectively. Additionally, T is the number of frames and F is the number of Mel-scale filters. $\hat{b}_{tf}$ is then transformed into cepstral coefficients by DCT. The cepstral coefficients with energy, and their delta and acceleration coefficients form the input feature $\mathbf{x}$ with 39 coefficients, and are applied to HMM for speech recognition.

The adaptation of the selective attention filter is done by maximizing the log-likelihood of the attended pattern $L = \log P(\tilde{\mathbf{x}}|\lambda)$ with the gradient ascent algorithm as

$$a_{tf}[n+1] = a_{tf}[n] + \eta \frac{\partial L}{\partial a_{tf}[n]} \tag{3}$$

where a_{tf} is the attention filter at time t for the fth Mel-filter. The gradient of (3) is calculated as

$$\frac{\partial L}{\partial a_{tf}} = \sum_{\tau} \sum_{t'} \frac{\partial L}{\partial x_{t'\tau}} \frac{\partial x_{t'\tau}}{\partial c_{t\tau}} \frac{\partial c_{t\tau}}{\partial \hat{b}_{tf}} \frac{\partial \hat{b}_{tf}}{\partial a_{tf}} \tag{4}$$

where $c_{t\tau}$ is the τth cepstral coefficient at time t. A summation over τ is required for DCT, and a summation over t' is needed for the delta and acceleration coefficients. The derivative of the log-likelihood on the speech features is derived in [7] for continuous density HMM.

As the selective attention process continues, the attended input pattern $\tilde{\mathbf{x}}$ may move toward the most likely pattern of the attended class, which is independent of the actual input pattern. To prevent this over-fitting, the attention filter needs be regularized by imposing low-complexity constraint. It is worth noting that this regularization is helpful and also biologically plausible for the filterbank output, but not for the MFCC features.

We propose to represent the attention filter as a linear summation of bilinear basis functions, i. e.,

$$a_{tf} = \sum_{t'f'} g_{t'f'} \Phi_{t'f'}(t,f) = \sum_{t'f'} g_{t'f'} \Phi\left(t' - \frac{t}{N_t}, f' - \frac{f}{N_f}\right) \tag{5}$$

where the basis function is locally defined as

$$\Phi(t,f) = \begin{cases} (1-|t|)\,(1-|f|), & \text{for } |t|<1 \text{ and } |f|<1 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

and $g_{t'f'}$ is the attention filter value at the low-resolution grid (t', f'). The original time-frequency space and low-resolution space are related as $t = N_t t'$ and $f = N_f f'$. Therefore, it is possible to reduce the complexity of the attention filter by increasing N_t and N_f. Now the adaptation is complete for $g_{t'f'}$.

In this letter, we define the confidence measure M_i as

$$M_i = (1-\gamma)\log P(\mathbf{x}|\lambda_i) + \gamma \log P(\tilde{\mathbf{x}}_i|\lambda_i), \tag{7}$$

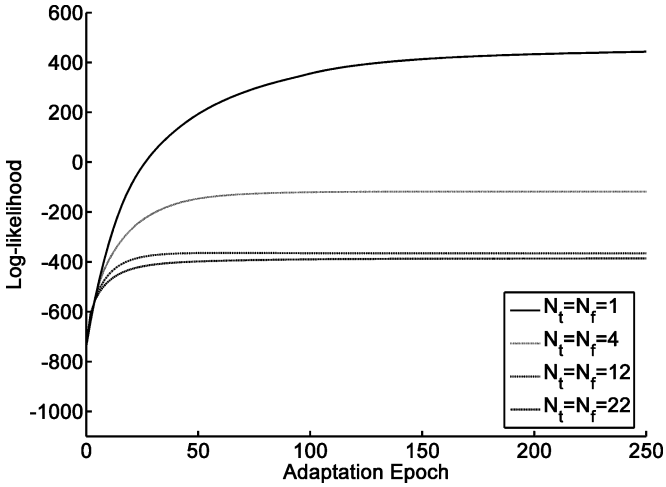

Fig. 2. Log-likelihood of an attended pattern for different grid sizes.

where $\log P(\mathbf{x}|\lambda_i)$ is the log likelihood of the original input pattern and $\log P(\tilde{\mathbf{x}}_i|\lambda_i)$ is that of the attended pattern to class i. Here, γ controls the relative importance between the two.

III. Experiments and Results

Speaker-independent isolated word recognition experiments were performed using isolated digit section of the AURORA database [8]. The database, which has a vocabulary of 11 words (the digits 1–9, "oh", and "zero"), contains 2412 utterances for training and 1144 utterances for testing.

After pre-emphasis with a coefficient of 0.97, an input speech signal was framed with a 25-ms Hamming window with 10-ms shifting. With fast Fourier transform (FFT), 23 Mel-scale filterbanks were formed at each time frame. Left-right continuous density HMMs were used. Each HMM had nine states with four Gaussian mixtures with diagonal covariance, and each was trained with the Baum-Welch algorithm with clean training data.

The attention filter was adapted until the log-likelihoods of the attended pattern converged. In these experiments, the optimal γ for the best recognition performance was different according to the level of the noise. As the noise level increased, larger γ resulted higher performance which gave more importance to the likelihood of the attended signal than the original signal. It was determined that $\gamma = 0.7$ gives the best performance in overall level of noise, and that $N_C > 7$ does not improve the performance.

Fig. 2 shows the log-likelihood of the attended input pattern during the top-down attention process with different values of N_t and N_f. For smaller N_t and N_f, the attention filter has a tendency of over-fitting and resulting in much higher log-likelihood values.

In Table I, the classification rates for the top-down attention are compared to the baseline HMM classifier. The recognition rates of the testing data with the smaller grid sizes were inferior compared to those of the baseline results, which clearly illustrate over-fitting. When the proposed low-complexity constraint is added, the recognition performance under a noisy condition is greatly improved. In addition, the performance is not sensitive

TABLE I
RECOGNITION RATES (%) WITH DIFFERENT GRID SIZES. ($\gamma = 0.7, N_C = 6$)

	Grid Size $N_t \times N_f$	SNR with White Gaussian Noise				
		clean	20dB	15dB	10dB	5dB
Baseline		**99.8**	97.8	93.5	73.3	43.4
The Same Grid Sizes with Time and Frequency	1x1	85.9	53.5	43.4	30.1	19.1
	2x2	85.0	60.8	49.8	31.6	21.1
	4x4	99.4	96.5	92.1	78.7	47.1
	8x8	99.6	98.9	96.9	89.7	62.1
	12x12	**99.8**	98.7	**97.0**	**91.0**	65.8
	16x16	**99.8**	99.0	96.3	90.6	63.2
	20x20	**99.8**	99.0	96.9	90.0	64.1
	22x22	**99.8**	98.8	**97.0**	90.9	**66.3**
Grid Sizes with Time Only (Same Attention Filter for All Frequency)	04x∞	**99.8**	**99.3**	96.9	87.1	60.6
	08x∞	**99.8**	99.2	96.9	86.5	59.1
	12x∞	**99.8**	99.2	96.6	87.0	57.1
	16x∞	**99.8**	99.2	96.5	86.0	55.8
	20x∞	**99.8**	99.0	96.9	86.7	56.1
	24x∞	**99.8**	99.2	96.9	86.5	53.9

to grid size when the grid size is big enough. The false recognition rates decrease from 2.2% to 0.7% for 20-dB SNR and from 6.5% to 3.0% for 15-dB SNR. These dramatic reductions, 68% and 54%, respectively, of the false recognition rate become smaller as the speech becomes noisier. This is similar to findings in cognitive science in which the effects of the top-down attention are significant only with familiar input patterns.

In (1), our attention filter is adjusted to maximize the likelihood of the input pattern. However, it maximizes a posterior probability $P(\lambda_i|\mathbf{x})$ only when $P(\mathbf{x})$ is fixed or not much changed. This is possible reason of worse performance with small grid size where the attended pattern moves far from the original input. In this case, the normalization of (7) by $P(\mathbf{x})$ may improve the classification performance.

The system showed good performance with stationary white Gaussian noise. In addition, the model can be applied to the nonstationary noise, because the multiplicative attention filter is not adjusted by the stationarity of the noise.

IV. CONCLUSION

A top-down selective attention model with an HMM classifier along with a low-complexity constraint is proposed. The top-down attention model iteratively determines the most-likely input pattern from noisy or corrupted input within the proximity of the input pattern. By introducing regularization on the attention filter in addition to a confidence measure, the proposed top-down attention model can greatly improve recognition rates in moderately noisy environments.

REFERENCES

[1] D.-S. Kim, S.-Y. Lee, and R. M. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 1, pp. 55–69, Jan. 1999.

[2] E. Broadbent, *Perception and Communication*. New York: Pergamon, 1958.

[3] A. Treisman, "Contextual cues in selective listening," *Quart. J. Exper. Psych.*, vol. 12, pp. 242–248, 1960.

[4] K. Fukushima, "Neural network model for selective attention in visual pattern recognition and associative recall," *Appl. Opt.*, vol. 26, pp. 4985–4992, 1987.

[5] K.–Y. Park and S.–Y. Lee, "Out-of-vocabulary rejection based on selective attention model," *Neur. Process. Lett.*, vol. 12, pp. 41–48, 2000.

[6] U.-M. Bae, H. M. Park, and S.-Y. Lee, "Top-down attention to complement independent component analysis for blind signal separation," *Neurocomputing*, vol. 49, pp. 315–327, 2002.

[7] S. Moon and J. N. Hwang, "Robust speech recognition based on joint model and feature space optimization of hidden Markov model," *IEEE Trans. Neural Netw.*, vol. 8, no. 2, pp. 194–204, Mar. 1997.

[8] D. Pearce and H. G. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ICSLP*, 2000, pp. 29–32.