estimation from the noisy signal. As shown in this correspondence, the clean speech derived codebook constrained approach is very effective. In addition, we have shown the enhancement performance limit using the clean speech parameters and accordingly, very low SNR segments can be improved only up to 0 dB SNR. Other segments with moderate and high SNR can only be slightly improved using further improvement in parameter estimation.

APPENDIX

The iterative Wiener filtering can be expressed as

$$S_i(\omega) = Y(\omega) \frac{\hat{P}_s(\omega)}{\hat{P}_s(\omega) + g^2\hat{P}_d(\omega)}; \quad i = 1, 2, \ldots$$

where

$$\hat{P}_s(\omega) = \frac{G_{i-1}^2}{|A_{i-1}(\omega)|^2}$$

and

$$A_{i-1}(z) = \left[1 + \sum_{k=1}^{a} a_k z^{-k}\right]$$

is estimated from $\hat{a}_{i-1} \Leftrightarrow \hat{S}_{i-1}(\omega)$ (Fourier transform pair). In addition, $Y(\omega) \Leftrightarrow y = s + g d$, where $s_a^2 = 1$ and $SNR = s_a^2/g^2$.

Considering such a low SNR case and clean speech LPC vector Wiener filtering where iterations become redundant, we can write

$$S_i(\omega) = S(\omega) = Y(\omega) \frac{\hat{P}_s(\omega)}{\hat{P}_s(\omega) + g^2\hat{P}_d(\omega)}$$

Defining the estimation error signal spectrum as $E(\omega) = S(\omega) - \hat{S}(\omega)$ and substituting for $Y(\omega) = S(\omega) + gD(\omega)$, we get

$$E(\omega) = S(\omega) \left[1 - \frac{\hat{P}_s(\omega)}{g^2\hat{P}_d(\omega)}\right] = \frac{gD(\omega)\hat{P}_s(\omega)}{g^2\hat{P}_d(\omega)}$$

$$\approx S(\omega) - \frac{D(\omega)\hat{P}_s(\omega)}{g\hat{P}_d(\omega)}.$$
speech recognition. If these two portions are modeled by the conventional HMM, then this information can be smeared by inaccurate modeling of state durations and state transitions, which represent the timing information and the temporal changes in the spectra. Some attempts have been made to incorporate correlations between successive observations. One solution is to use a high-order HMM, which has a conditional output probability density function (pdf) [2]. Its output probability of the current frame depends on the previous frame feature vector. Another is to use an observation pdf of a state that changes as a function of time during the time spent in the state [3], [4] or a template derived from an ensemble of segments corresponding to a state [5]. However, in a real situation, it is very difficult to obtain conditional output pdf’s directly from the training data because numerous parameters must be estimated.

To remedy this problem, we propose a modified HMM (MHMM) based on the assumption that each frame is dependent on the previous frame. With this proposed model, the temporal changes and the timing information in the spectra can be modeled accurately. Our modeling assumption is essentially based on the fact that the temporal changes and the acoustic effects of timing differences in the spectra characterize the time-varying vocal tract and consequently play an important role in human perception.

This paper is organized as follows. In Section II, we first describe the structure of the MHMM and compare it with the conventional HMM. Next, we show that its parameters can be trained by the segmental k-means algorithm. Then, in order to provide a solution to the problem that arises due to the insufficient training data, we propose two solutions. In Section III, we show recognition results for the MCDHMM and compare them with those obtained by the conventional CDHMM. Finally, we draw conclusions in Section IV.

II. MODIFIED HMM

A. Structure of the MHMM

For every state, let \( a_{ij} \) denote the transition probability of moving from state \( s_i \) at any time \( t \) to \( s_j \) at time \( t+1 \). In addition, let \( P_l(\tau) \) be the probability of having self-transition in state \( s_i \) over \( \tau \); for self-transition in \( r \)th state, the observation probability is given by \( B_{ir}(O_t) \); and for transition, the state duration-dependent observation probability combined with state duration probability \( P_l(\tau) \) is given by \( B_{ir}^{sz}(O_t) \). To explain how the MHMM is obtained, we show the structure of the MHMM, together with that of the conventional HMM in Fig. 1. Note that if \( B_{ir}^{sz}(O_t) \) is replaced by \( B_{ir}(O_t) \), the MHMM is the same as a hidden semi-Markov model (HSMM) algorithm [2], and if \( P_l(\tau) \) is set to 1 and \( B_{ir}^{sz}(O_t) \) is replaced by \( B_{ir}(O_t) \), the MHMM is the same as the conventional HMM.

For the spectral density within the \( r \)th state of the MCDHMM, the pdf in state \( s_i \), \( b_{is}(O_t) \) is given by

\[
b_{is}(O_t) = \sum_{m=1}^{M} c_{im} N(O_t, \mu_{im}, \Sigma_{im})
\]  

(1)

and the observation pdf combined with \( P_l(\tau) \) in transition from state \( s_i \) to \( s_j \), \( b_{ij}(O_t) \) is given by

\[
b_{ij}(O_t) = \sum_{m=1}^{M} c_{ijm} N(O_t, \mu_{ijm}, \Sigma_{ijm})
\]  

(2)

where

- \( M \) number of Gaussian mixture components
- \( c_{im} \) mixture gain for the \( m \)th mixture
- \( N(O_t, \mu_{im}, \Sigma_{im}) \) Gaussian probability density function for an observation vector \( O_t \) with the mean vector \( \mu_{im} \) and the covariance matrix \( \Sigma_{im} \).

In addition, the state duration-dependent terms \( c_{ir,m}, N(O_t, \mu_{ir,m}, \Sigma_{ir,m}), \mu_{ir,m}, \) and \( \Sigma_{ir,m} \) denote the mixture gain, the Gaussian pdf, the mean vector, and the covariance matrix, respectively, for the state duration-dependent \( m \)th mixture parameters.

The segmental k-means training algorithm can be used for estimating the MCDHMM parameters in the proposed algorithm [6]. To estimate the parameters of the MCDHMM, the Viterbi decoding algorithm can be used, and the optimal path can be determined according to

\[
p_l(j) = \max_t \max_{\tau} \left[ p_{l-\tau}(j) a_{ij} P_l(\tau) b_{ij}(O_t) \prod_{k=1}^{T-\tau} b_{is}(O_{t+k}) \right]
\]  

(3)

where \( b_{is}(O_t) \) is the state duration-dependent observation pdf in transition from state \( s_i \) to \( s_j \), and \( p_l(j) \) denotes the highest pdf taken in state \( s_j \). The assumption of duration-dependent observations results in the \( P_l(\tau) b_{is}(O_t) \) term of the output pdf’s.

B. Training of MCDHMM Using the Segmental k-Means Algorithm

There are two issues concerning the training procedure that need further explanation. The first issue is concerned with the initialization of the state duration-dependent parameters, and the second one is the smoothing algorithm during the training procedure.

In this work, during the predefined iteration, the conventional CDHMM is used for estimating the MCDHMM parameters. The MCDHMM terms \( P_l(\tau), b_{is}(O_t), \) and \( b_{ij}(O_t) \) are estimated by the equation of the optimal path. After the predefined iteration, the parameters of the MCDHMM are estimated and trained by the optimal path decoded from (3). The state duration-dependent parameters are smoothed after the parameters of the MCDHMM are estimated because the training utterances are far from sufficient to accurately estimate those parameters. Note that we apply the smoothing method only to the state duration-dependent probabilities but not to the parameters.
If \( P_r(\tau) \) is equal to zero, the state duration-dependent probabilities \( P_r(\tau) \) and \( b_r(\Omega) \) are smoothed by the following equations:
\[
P_r(\tau) = 0.5 \times \max(P_r(\tau - 1), (P_r(\tau + 1))
\]
\[
b_r(\Omega) = 0.5 \times \max(b_{r-1}(\Omega), b_{r+1}(\Omega)).
\]

In the MCDHMM, the state duration-dependent parameters \( P_r(\tau), b_r(\Omega) \), and \( h_r(\Omega) \) must be estimated for all possible combinations of \( s_i \) and \( \tau \), which require a large amount of training utterances. In the case of \( \tau \geq D_{max} \), the parameters \( P_r(\tau) \) and \( (\mu_{r,im}, \Sigma_{r,im}, \Theta_{r,im}) \) are replaced by \( P(D_{max}) \) and \( (\mu_{D_{max},im}, \Sigma_{D_{max},im}, \Theta_{D_{max},im}) \), respectively. Sufficient training data, automatic training algorithms, and detailed modeling are three important factors for a successful speech recognition system. In fact, consideration of these three aspects is the main point of this work. As a solution to the problem of insufficient data for training, we will use a different number of mixtures for \( h(\Omega) \) and \( b(\Omega) \).

### III. Speech Recognition Experiments

#### A. Task and Data Base

To investigate the performance of the proposed algorithm in speech recognition, a vocabulary with isolated 1036 Korean words produced by 48 speakers (34 males and 14 females) was used as the speech material. Words of 39 speakers (28 males and 11 females) were used for training data and those of the other nine for test data. The total number of training words was 12227 and that of the test word was 2810. They were used for estimating parameters of 32 phonemes and testing of the proposed algorithm. The phonemes used here were represented by a single first-order, left-to-right model of three states, with self and forward transitions without skipping.

The speech signal was sampled at 16 kHz and segmented into 20 ms frames with each frame advancing every 10 ms. Each frame was parameterized by 24-D feature vectors consisting of 12-LPC derived liftered cepstral coefficients and their corresponding time derivatives.

#### B. Experiment Results of the Proposed MCDHMM

In this experiment, the maximum state duration \( D_{max} \) is set to 10, i.e., \( \tau \geq 1 \). Table I shows recognition results for the CDHMM and MCDHMM with state duration probabilities only, using cepstral coefficients \( \tilde{C}_1 \) and cepstral coefficients combined with differential cepstral coefficients \( \tilde{C}_1 + \Delta \tilde{C}_1 \). The performances of the conventional CDHMM and the CDHMM with state duration probabilities only were also evaluated when the numbers of mixtures were 1, 2, 3, 4, 6, 9, 12, and 15, respectively. The recognition rates of the conventional CDHMM were similar to that of the CDHMM with state duration probabilities only. This is shown in Table I, where the value in the parenthesis shows the recognition accuracy including the second candidate. The improvement appears to be insignificant. The reason is that the CDHMM with state duration probabilities only treats spectral and duration models as separate or loosely connected. This can be expected as the two models differ only in state duration probability.

Tables I and II show that when the parameter sets are limited to \( \tilde{C}_1 \), the improvement obtained by the MCDHMM is relatively large for 1, 2, 3, and 4 Gaussian mixtures. With \( \tilde{C}_1 \), the performance of the MCDHMM with three mixtures is better than that of the conventional CDHMM with 15 mixtures. In addition, in the case of \( \tilde{C}_1 + \Delta \tilde{C}_1 \), the MCDHMM with three mixtures leads to similar performance of the CDHMM with 12 mixtures. This is because the CDHMM assumes that each frame is independent of the previous frame, but the proposed MCDHMM assumes that it is dependent on the previous frame. In addition, when the parameter sets include the \( \tilde{C}_1 + \Delta \tilde{C}_1 \), the improvement becomes significant in comparison with the CDHMM.

#### IV. Conclusion

We have proposed an MHMM to model state transitions and to have accurately temporal structures and timing informations. The main concept behind the MHMM is that, by adjusting local transition constraints, ordinarily relaxed constraints may be given to the steady portions of speech signal, which is susceptible to variations in the speaking rate, whereas tighter constraints may be given to the transient portions of speech signal in which timing information is known to be important for speech perception.
We have found that when compared with the conventional HMM, the proposed MHMM yields improvements in recognition performance. It has indirectly been shown that the transient and timing information in the spectra not only characterize the time-varying vocal tract but also play an important role in human perception. It has been shown that MHMM's can estimate the state duration-dependent parameters by Viterbi decoding in a straightforward manner. In the MCDHMM, detailed acoustic modeling will increase the number of parameters that will be estimated from observations and require a large amount of training uttersances. Thus, we have proposed a solution to the conflict between detailed acoustic modeling and insufficient training data. The solution is the use of a different number of mixtures between state duration-independent and state duration-dependent observation probability. In case of insufficient training data, the duration-dependent parameters can be estimated by postprocessing.

### REFERENCES


### TABLE III

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>Proposed MHMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>1</td>
<td>C + αC</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>C + αC</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>C + αC</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>C + αC</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>C + αC</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>C + αC</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>C + αC</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>C + αC</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>C + αC</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>C + αC</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>C + αC</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>C + αC</td>
</tr>
</tbody>
</table>

**A:** The number of mixtures in state duration-independent observation probability
**B:** The number of mixtures in state duration-dependent observation probability

**C:** Observation probability

### I. INTRODUCTION

We consider the problem of restoring speech that has been degraded through addition of multiple echoes. This problem appears very often in hands-free telephony when the microphone is placed far from the speaker. The sound wave at the microphone location consists of the direct path wave and multiple delayed and attenuated versions of it due to reflections on the room walls and other surfaces. Depending on the microphone location, the energy of the echoed speech might be large enough to degrade the intelligibility of the speech the far-end listener hears. Reverberation heavily affects the performance of any automatic speech recognition system. Real applications demand that the performance of a speech recognition system not be affected by changes in the environment. However, it is well known that when a recognition system is tested under conditions different than those used to train it, its recognition rate drops dramatically.

There are many sources of distortion that can degrade the accuracy of a speech recognition system. Generally, the sources of degradation are clustered into two categories: 1) additive noise sources that could be due to machinery, interference from other speakers, etc. and 2) convolutional noise sources due to the acoustical properties of the environment or due to the impulse response of the microphones used. Several types of array processing strategies have been applied to speech recognition systems. One approach is the delay-and-sum beamformer [2], where steerings delays are applied at the outputs of the microphones to compensate for arrival time differences at the two microphones, reinforcing the desired signal over other background signals. Another approach is based on adaptive minimization of the mean square error [11]. These algorithms provide nulls in the direction of undesired noise sources and reinforce sensitivity in the direction of the desired signal. The key assumption in these algorithms is that the desired speech signal is statistically independent of all sources of degradation, which means that the distortion cannot be due to delayed versions of the desired signal, as is the case in a reverberant room. All these techniques require multiple microphones and calibration of the microphone array.

Manuscript received November 14, 1994; revised April 1, 1996. The associate editor coordinating the review of this paper and approving it for publication was Dr. James H. Snyder.

The authors are with the Electrical and Computer Engineering Department, Drexel University, Philadelphia, PA 19104 USA.

Publisher Item Identifier S 1063-6676(96)06711-9.