Department of
Veterans Affairs

JRRD Volume 42, Number 3, Pages 363–372
May/June 2005
Journal of Rehabilitation Research & Development

Effect of high-frequency spectral components in computer recognition of dysarthric speech based on a Mel-cepstral stochastic model

Prasad D. Polur, PhD;* Gerald E. Miller, PhD
Department of Biomedical Engineering, Virginia Commonwealth University, Richmond, VA

Abstract—Computer speech recognition of individuals with dysarthria, such as cerebral palsy patients, requires a robust technique that can handle conditions of very high variability and limited training data. In this study, a hidden Markov model (HMM) was constructed and conditions investigated that would provide improved performance for a dysarthric speech (isolated word) recognition system intended to act as an assistive/control tool. In particular, we investigated the effect of high-frequency spectral components on the recognition rate of the system to determine if they contributed useful additional information to the system. A small-size vocabulary spoken by three cerebral palsy subjects was chosen. Mel-frequency cepstral coefficients extracted with the use of 15 ms frames served as training input to an ergodic HMM setup. Subsequent results demonstrated that no significant useful information was available to the system for enhancing its ability to discriminate dysarthric speech above 5.5 kHz in the current set of dysarthric data. The level of variability in input dysarthric speech patterns limits the reliability of the system. However, its application as a rehabilitation/control tool to assist dysarthric motor-impaired individuals such as cerebral palsy subjects holds sufficient promise.

Key words: cerebral palsy, communication disorder, dysarthric speech, hidden Markov model, high-frequency spectral components, human-machine interfaces, Mel-frequency cepstral coefficients, rehabilitation, speech and motor disabilities, speech recognition.

INTRODUCTION

Speech articulatory disability or dysarthria can arise from a number of conditions, including cerebral palsy, multiple sclerosis, Parkinson's disease, and others [1]. In this study, we investigated 3 subjects exhibiting dysarthria due to cerebral palsy, since this condition accounts for a large number of dysarthric patients. But the results of this work may be equally relevant to noncerebral palsy individuals with dysarthria. Hence, to emphasize a generality, we shall use the term "dysarthric speech" to describe speech that is difficult to understand as a result of the speaker's disability and that is characterized by distortions, substitutions, and omissions. Dysarthric errors result from a disruption of muscular control due to lesions of either the central or peripheral nervous systems. Some people with dysarthric speech, like cerebral palsied individuals, are also severely motor-impaired, with limited or no control of their local environment [2].

Abbreviations: DXD = digit data for subject X, DXW = word data for subject X, HMM = hidden Markov model, MFCC = Mel-frequency cepstral coefficient, ND = digit data for normal speech, NW = word data for normal speech, RR = recognition rate.

This material was based on work supported in part by the Whitaker Foundation (Rosslyn, Virginia) special opportunity award in Rehabilitation Engineering and Human-Computer Interfaces and by the School of Engineering at Virginia Commonwealth University (Richmond, Virginia).

*Address all correspondence to Prasad D. Polur, Department of Biomedical Engineering, Virginia Commonwealth University, 1112 E. Clay St., Room 220B, Richmond, VA 23298; 804-852-4624; fax: 804-828-4454. Email: pbpolur@vcu.edu

DOI: 10.1682/JRRD.2004.06.0067

In general, cerebral palsied individuals lack articulatory precision. Simple "steady-state" phonemes, such as vowels, are physically the easiest to produce since they do not require dynamic movement of the articulatory structures. However, phonetic transitions, i.e,. consonants, are most difficult to produce since they require fine motor control to precisely move the articulators. Severely impaired speakers and mildly impaired speakers differ in degree of disability rather than quality [2].

Application of speech-recognition technology to dysarthric speech would enable the people affected by such speech to electronically enhance their intelligibility. It can also benefit them in control applications. Such applications would be immensely useful to these individuals, since they seem to prefer a natural mode of communication and control as represented by speech [3]. The problem of computer recognition of dysarthric speech involves several challenges that are not encountered in normal speech. The enormous variability and nonconformity of such speech imposes high constraints on the recognition system and, hence, methods to improve recognition either through signal modification or model variation, among other means, is an ongoing process.

Several investigations have identified the applicability of speech-recognition technology to dysarthric speech. Sy and Horowitz evaluated the degree of speech impairment and the utility of computer recognition to such speech using a statistical causal model [4]. They presented a case study of a dysarthric speaker compared against a normal speaker serving as a control. They reported that dysarthric speech is perceived as articulatory error patterns in comparison with normal speech. Further, they identified that the sources of error in terms of the manner of articulation primarily originate from either stops or fricatives. They concluded that a recognition system could effectively serve as an augmentative communication/control device when the system appropriately exploits the characteristic of dysarthric speech patterns by using words that require less dynamic movement of speech articulators for such speakers. Patel investigated the application of such technology to severely dysarthric speakers by examining prosodic parameters like frequency and intensity contours [5]. This technology may provide additional channels of communication since dysarthric speakers are unable to use devices such as keyboards or mice because of impaired motor control. Patel reported that commercial speech-recognition technology was more applicable to, or can be more reliably applied to, mild and moderate dysarthric speakers than to severely dysarthric speakers. Goodenough and Rosen performed a similar investigation and reported that speech-recognition performance rapidly deteriorated for vocabulary sizes greater than 30 words, even for persons with mild to moderate dysarthria [6]. Jayaram and Abdelhamied also investigated the commercial IntroVoice speech-recognition system and reported that it had a low recognition rate (RR), while a dedicated small vocabulary system developed and trained on the same data produced better results [7]. Further, they reported that large multisyllable words with higher consonant content produced greater recognition error than words with low number of syllables owing to inconsistency in articulation.

In view of these previous investigations, the current study used the speech of cerebral palsy patients who were subjectively classified as moderately dysarthric by a trained clinician. We restricted the vocabulary to fewer than 30 utterances for the subjects' convenience and also to limit the amount of data to be analyzed. The words were chosen such that they were recognizable by the subjects and had only one normal pronunciation. The utterances had a limited number of syllables with low complexity to reduce articulation inconsistency. Another important consideration was that words had a real-world application for the dysarthric individuals, such as environmental or mobility controls. In mobility control applications, such as a wheelchair and in appliance control such as television, radio, or telephone, a menu structure with simple command words as well as digits would be highly practical. Some other simple words like "my," "is," etc., were chosen for initial simplification of analysis. The utterances chosen were a subset of the words used in previous research.

The most popular and successful tool/model in speech recognition applications is the hidden Markov model (HMM). Several investigations have adapted HMMs toward dysarthric speech recognition and identified methods to enhance recognition of such speech using signal modifications, among other techniques. Deller et al. used a vector quantized HMM approach to model isolated dysarthric speech, in which they reported that a full structured HMM, along with clipping of transitional regions of speech, improved recognition [8]. They reported that transitional regions of speech might be detrimental to the recognition process, indicating that dysarthric speech characteristics might be treated differently than normal speech from a computer recognition point of view. Deller et al. suggest that the use of Mel-cepstral coefficients and inclusion of high-frequency spectra may hold promise for improved recognition of such speech. Chen and Kostov used a statistical approach and reported that the direction of the formant transitions might provide important acoustic cues/contrasts, which permit enhanced discrimination [9]. They also indicate that lower energy high-frequency components might play a positive role in the recognition process. Kain et al. investigated voice transformation systems for dysarthria, with the goal of implementing intelligibility-enhanced speech in a wearable device [10]. They estimated the formants and energies of a dysarthric speech set and modified the trajectories to more closely approximate desired targets. Kain et al. performed this initial investigation to enhance speech intelligibility through signal modification, where they reported that removal of vocal fry improved perceived quality. They identified that such speech exhibited both distortions and very high levels of variability, thereby showing significant spectral deviation from normal speech.

One could infer that in addition to the low-band, the high-band components above 4 kHz might also play a role in enhancing resynthesis accuracy and in recognizing this type of speech. Moreover, we investigated the spectrum of the current data using Sound Forge 6.0 professional audio software (Sony Media Software, Madison, WI), revealing that although most of the high-energy components were present in the frequency band below 5.5 kHz, significant frequency components of lower energy for most utterances were still present above this band. These ranged from 6.5 kHz for some utterances to 14 kHz for a few others and were not regularly encountered to the same extent in the normal speech set/calibration set.

Most commercial systems, e.g., Dragon Naturally Speaking, band limit the signal to approximately 5 kHz for normal speech applications. This ensures reduced processing power and faster response. Since some dysarthric speech characteristics were different from normal speech from a computer recognition point of view, we felt that an investigation into the utility of these higher frequency components for the current data was necessary. Even though a subset of the vocabulary and not all the words were expected to have potentially useful high-frequency information, such investigation using the current limited data set might provide sufficient direction at this preliminary stage. We adopted a Mel-cepstral HMM approach for this investigation, since it could provide good statistical representation for this data type. This study is thus part of an early effort to develop an artificially intelligent communication/control tool for speech- and motor-impaired individuals.

METHODS

Data Acquisition and Signal Processing

Three male cerebral palsy patients were chosen whom a trained clinician subjectively classified as moderately dysarthric. Their speech was recorded in multiple sessions after obtaining informed consent from each. The subjects were asked to read aloud a set of 10 digits and a set of 15 words in a normal manner. Each utterance was repeated 12 times in a low-noise environment to reduce acoustic interference. The recording was performed at a sampling rate of 44.1 kHz with an Audio-Technica (Stow, OH) AT3525 cardioid condenser microphone that was connected to the TASCAM (Montebello, CA) DA-P1 digital tape recorder. The microphone had a flat frequency response ranging from 30 Hz to 20 kHz. The recorded speech was loaded into the computer through an M-Audio (Irwindale, CA) 24-bit DIO 2448 input/output card. The data were segregated and individually stored as .wav files with the use of Sound Forge 6.0.

To study the effect of high-frequency components on recognition, we investigated the signal at two levels. One set had a sampling rate of 44.1 kHz band-limited to 15 kHz, referred to henceforth as “set H.” Another set had a down-sampled rate of 11.025 kHz band-limited to 5.5 kHz, referred to henceforth as “set L.” The MATLAB-generated (The MathWorks, Natick, MA) spectrogram of the speech samples is shown in **Figure 1** for illustration of spectral content. Of these sample files, only the first eight repetitions per utterance were used in the training/testing phase, because we noticed that the last few repetitions introduced excess variability caused by physical fatigue and frustration of the dysarthric individual. Hence, we deemed these inappropriate for use. Four randomly chosen repetitions of each digit or word were used as training data and the remaining four for testing/recognition. The test data were designated digit data for dysarthric subject X (DXD) and word data for dysarthric subject X (DXW). Thus, set H had DXD and DXW with higher frequency components, and set L had DXD and DXW with frequency components limited to 5.5 kHz. A vocabulary of 15 words and 10 digits, with 8 repetitions each, was thereby created for each subject. The set of utterances are

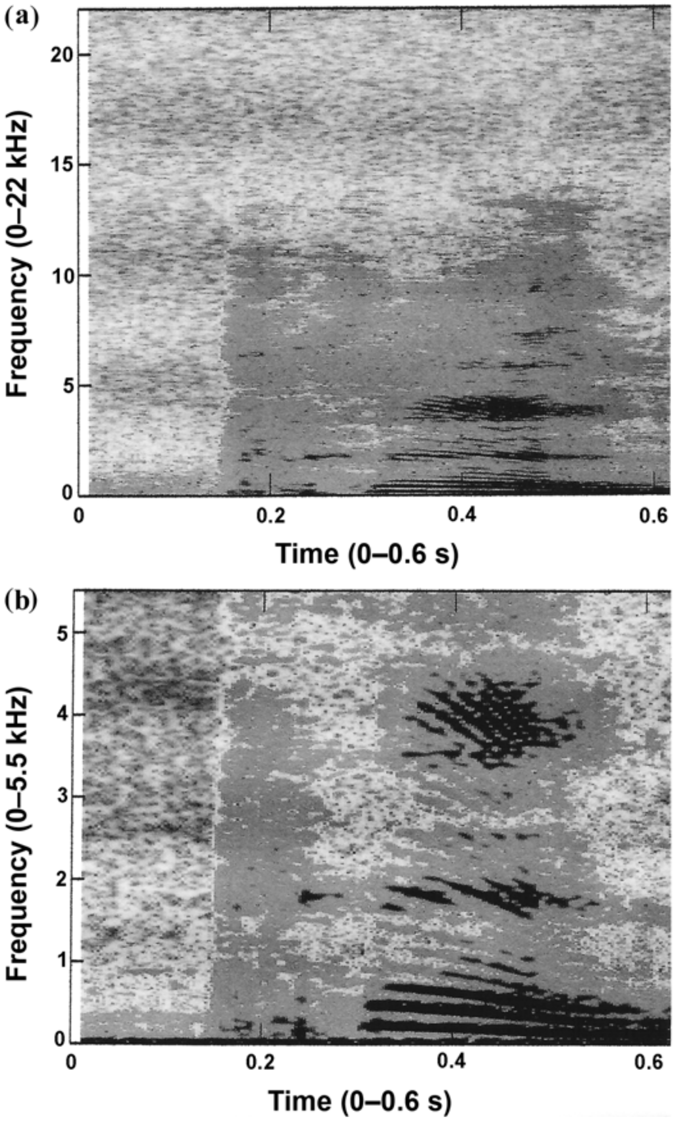

Figure 1.
Spectrogram of word “ten”’ from word data for subject 1 (D1W): **(a)** from set H and **(b)** from set L. Significant spectral components (lower energy) above 5.5 kHz can be seen.

listed in **Table 1**. Additionally, we created a similar test set using normal speech, i.e., digit data for normal speech (ND) and word data for normal speech (NW) as a reference to verify the proper functionality of the recognition system.

Table 1.
List of utterances used in experiments.

Utterance	Digits	Words
1	One	Back
2	Two	Front
3	Three	Go
4	Four	Good
5	Five	Is
6	Six	Left
7	Seven	My
8	Eight	Name
9	Nine	No
10	Ten	Ok
11	—	Right
12	—	Start
13	—	Stop
14	—	Very
15	—	Yes

Feature Extraction

The recognition system was implemented in MATLAB. The speech signal was divided into short 15 ms frames segments with the use of a Hamming window for further analysis. We experimentally determined in our previous pilot study that 15 ms frames generated better recognition than 10 ms frames for the available dysarthric data. In general, dysarthric speakers have a slow rate of articulation. Therefore, a larger frame size might provide more nontransitional information to the statistical model that enhances its ability to learn the characteristics of the signal. However, we noted that no further significant improvement was obtained for frame sizes larger than 15 ms, and hence, we retained the frame size as such. The average RR versus frame size is illustrated in **Figure 2(a)** for set L. Further, we identified from the pilot study that a 10-state ergodic HMM exhibited a high level of robustness in its ability to handle the large amount of variability present in dysarthric speech. An illustration of the effect of number of states on average RR for DXW data of the Mel-cepstral model is shown in **Figure 2(b)** for set L.

Mel frequencies are based on the known variation of the human ear’s critical bandwidths with frequency filters spaced linearly at frequencies below 1 kHz and logarithmically at higher frequencies. These scales have been used to capture the phonetically important characteristics of speech. Feature extraction involves the use of a filter bank with center frequencies and bandwidths determined by the Mel-frequency scale just described. The steps involved in Mel-frequency cepstral coefficient (MFCC) extraction are illustrated in **Figure 3**.

The MFCCs are calculated with

$$C_i = \sum_{k=1}^{N} X_k \cos\{[\pi_i(k-0.5)]/N\}, \text{ for } i = 1, 2....P, \qquad \textbf{(1)}$$

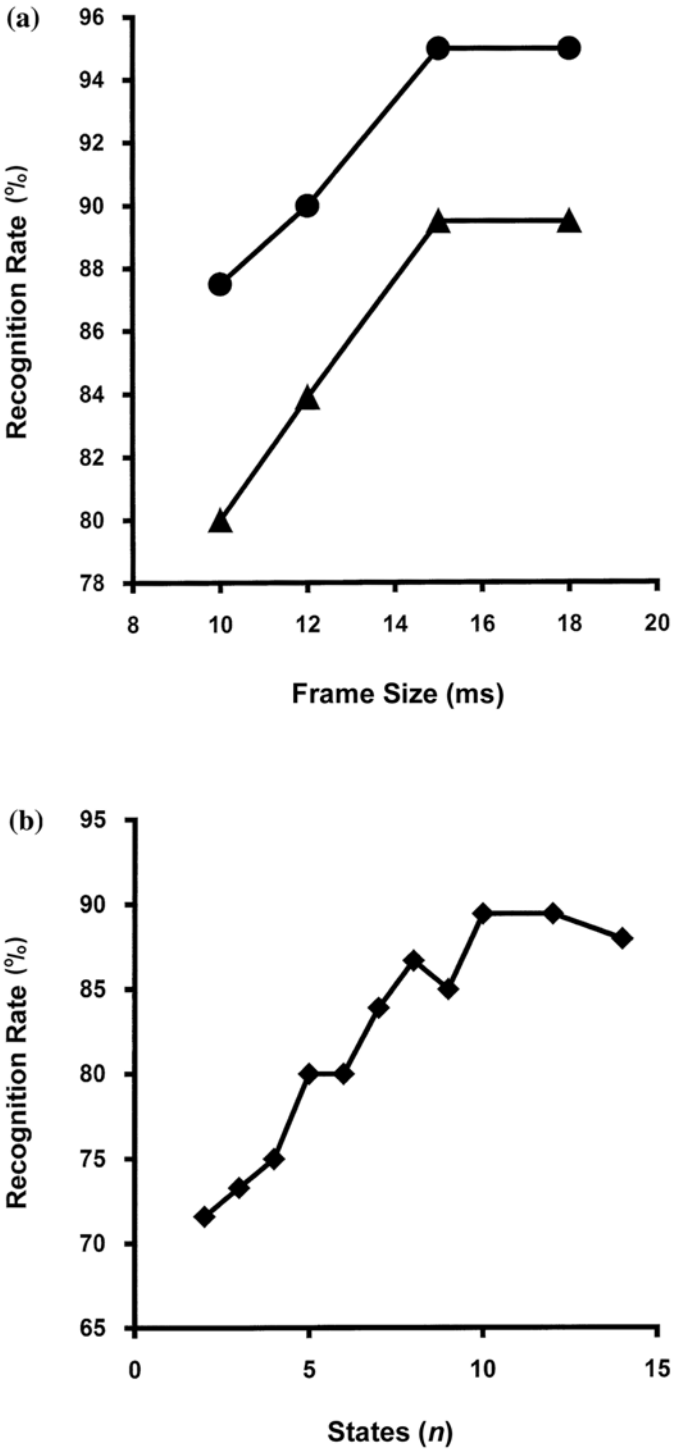

Figure 2.
(a) Average recognition rate (RR) vs. frame size with Mel-cepstral model for digit data for subject X (DXD) (circular points) and word data for subject X (DXW) (triangular points) for set L, where one can see that 15 ms frame model generates higher RR. **(b)** Effect of number of states on average RR of model for set L DXW, where one can see that 10-state model produces higher RR.

where C_i is the cepstral coefficients, P is the order, k is the number of discrete Fourier transform magnitude coefficients, X_k is the kth order log-energy output from the filter bank, and N is the number of filters (usually 20). Thus, 14 coefficients and an energy feature were extracted, generating a vector of 15 coefficients per frame. Additional information on cepstral coefficients can be obtained from Noyes and Frankish [3] and Mak et al. [11].

Acoustic Model (HMM)

The feature vectors serve as input to an acoustic model, namely, the HMM. A simplified description of an HMM follows. Detailed descriptions of HMMs are available from several sources [8,12–13]. Markov models are mathematical models of stochastic processes, i.e., processes that generate random sequences of outcomes according to certain probabilities. A simple example of a stochastic process is a sequence of coin tosses, the outcomes being heads or tails. We can construct Markov models as a simple case using state diagrams, such as the one shown in **Figure 4**.

In **Figure 4**, $S1$ and $S2$ represent the possible states of the process we are trying to model, i.e., coin tossing, and the arrows represent transitions between states. The label on each arrow represents the probability of that transition. At each step of the process, the model generates an output, or emission, depending on which state it is in.

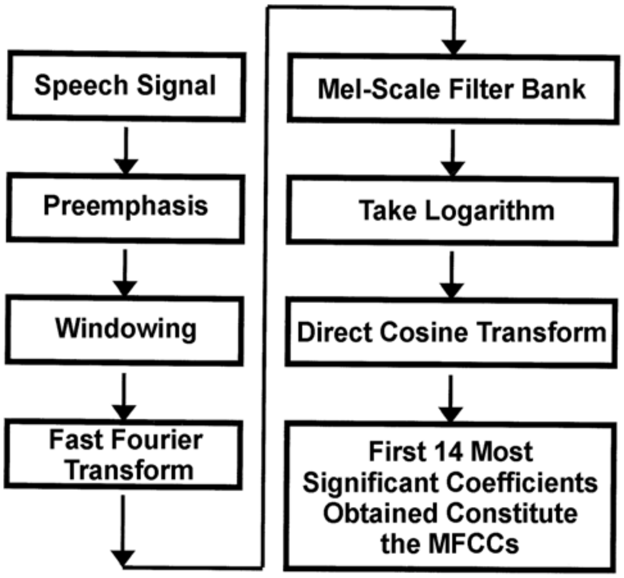

Figure 3.
Steps involved in Mel-frequency cepstral coefficients (MFCCs) extraction.

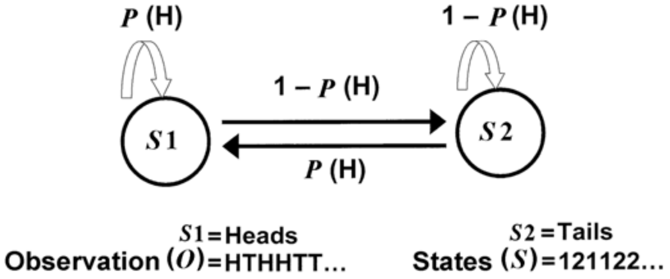

Figure 4.
Markov model in coin-toss experiment. *P* = probability.

In this example of a sequence of coin tosses, the two states are heads (*S*1) and tails (*S*2). The most recent coin toss determines the current state of the model and each subsequent toss determines the transition to the next state. If the coin is fair, the transition probabilities are all 0.5. Thus there is a correspondence between the observations, i.e., heads or tails, *O* = H T H H …, and the states, i.e., *S*1 or *S*2, *S* = 1 2 1 1…. In this example, the emission at any moment in time is simply the current state. However, in more complicated models, the states themselves can contain random processes that affect their emissions.

An HMM is a model in which we observe a sequence of emissions, but we do not know the sequence of states the model went through to generate the emissions. In this case, the goal is to recover the state information from the observed data. An HMM is characterized by the following elements:

- *S*, the number of states in the model. The previous coin-toss example contained two states, namely, *S*1 (heads) and *S*2 (tails). In speech application, the number of states usually approximates the number of phonemes or subphonemes in the word. Thus, for an utterance such as “yes,” which may be viewed to contain three phonemes, we can construct a three-state HMM model. Generally, the states are interconnected in such a way that any state may be reached from any other state, i.e., an ergodic model, or some restrictions may be imposed on how transitions can take place from one state to another.
- *O*, the number of distinct observation symbols per state, i.e., the discrete alphabet size. The observation symbols correspond to the physical output of the model. In the coin toss example, the observation symbols were simply heads or tails. In the case of speech, the observation symbols are the 15 vectors per frame, since they are the parametric representation of the phonemes in the word model.
- *T*, the state transition probability distribution. This governs the probability of transitioning from one state to another in the model.
- *B*, the emission probability. This governs the likelihood of emitting each possible sound (phoneme) while in a particular state. For example, in a three-state HMM with states *S*1, *S*2, and *S*3 modeling the word “yes,” the emission probability for the phoneme /s/ would be much higher in state 3 (*S*3) than in the other two states. This is because it is the terminal state that corresponds to the terminal phoneme in the word. In other words, a much higher probability exists that the final state is responsible for emitting the phoneme /s/.
- *P*0, the initial state distribution/probability. This determines the initial probability for each state in the model.

Thus, an HMM can be appropriately defined by specifying *S*, *O*, *T*, *B*, and *P*0. Usually, we may specify the number of states (*S*) and the observation vector (*O*) for the HMM. However computing the model’s other parameters is difficult, since the states are not directly observable and transitions are probabilistic. One method used to tackle this is the Baum Welch algorithm [12–13], which trains the HMM. This algorithm finds the HMM parameters *T*, *B*, and *P*0, with the maximum likelihood of generating the given symbol sequence. To appropriately use the HMM for a speech-recognition application, one must only decode this information in the HMM. Although the states cannot be directly observed, the most likely sequence of states for a given sequence of observed outputs could be computed with the Viterbi algorithm [12–14]. The Viterbi algorithm calculates the likelihood for each HMM in the word model. In the isolated word case, where each unique word is associated with a unique HMM, the index of the HMM that produced the highest likelihood corresponds to the recognized word.

Thus, we set up a 10-state ergodic model with a slight left-to-right character, wherein the initial state is fixed. An illustration of one such trained HMM used in the recognition system is shown in **Figure 5**. The HMM system comprised 25 trained individual HMMs, one for each word/digit in the vocabulary. The index of the HMM that produced the highest likelihood corresponded to the recognized word. The performance of the models was determined by the RR, which is defined as the ratio of the correctly recognized utterances to the total number of utterances used in testing the system. To verify the functionality of the system, we used a set of normal speech data, i.e., ND and NW, to first train the HMM setup. The

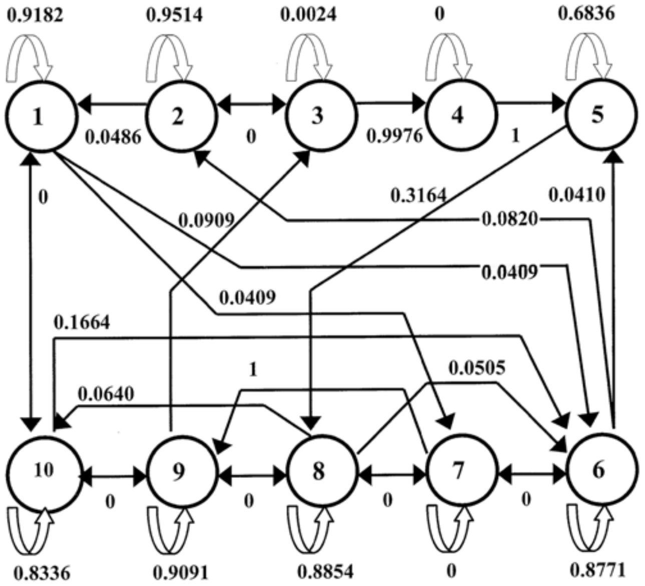

Figure 5.
Ten-state ergodic model for word “yes” obtained from set L. Transition probabilities from one state to another in trained model are depicted.

RR of the system to normal test data of both the high and lower sampled rates was verified to be 100 percent.

RESULTS

The MFCC feature vectors of the dysarthric speech served as input (training/testing set) to the established system. The results of this investigation are summarized in **Tables 2** and **3**. **Table 2** provides the mean values of the RR for sets L and H. **Table 3** provides a comparison of the difference between the means of the two conditions (digit data and word data) with the use of a paired sample *t*-test. From these two tables, one can see that the Mel-cepstral HMM model provided a slightly higher mean RR when using set L than when using set H. The two conditions have a sufficiently high correlation. Additionally, the paired *t*-test results indicate that the marginal difference in the means of the two conditions is statistically significant.

DISCUSSION

From the results of our investigation, one can assume that no significant useful information/cues are available to the system above 5.5 kHz in the data set that would enhance its ability to discriminate dysarthric speech. In some cases, set H proved slightly detrimental to the recognition task. This difference may be due to higher variability in both lower and higher frequency components in some utterances, e.g., “seven,” “start,” “front,” etc., which may be due to the inability of the subjects to articulate those terms consistently. The Mel-cepstral HMM model depends on some level of consistency in the data during its training phase to learn the underlying characteristics of the word being modeled. This permits appropriate recognition results during the testing phase. Even though variability exists in the lower frequency region, the presence of variability in the high-frequency region might have an additive detrimental effect on the recognition process. In this respect, the set L model might be better able to cope with the data by eliminating some of this high-frequency

Table 2.
Mean values of the recognition rate (RR) for sets L and H.

Dysarthric Data	Set L (11 kHz) (%)	Set H (44 kHz) (%)
D1D	92.50	87.50
D2D	97.50	95.00
D3D	95.00	95.00
Average DXD RR	95.00	92.50
D1W	86.70	83.30
D2W	91.70	92.00
D3W	90.00	90.00
Average DXW RR	89.47	88.33
Overall RR (for 25 utterances)	91.67	90.00

DXD = digit data for dysarthric subject X
DXW = word data for dysarthric subject X

Table 3.
Statistical analysis—paired sample *t*-test for sets L and H (N = 25).

Set	Mean ± SD	SEM	Correlation (L & H Pair)	
L	91.67 ± 5.89	1.18	0.9	
H	90.00 ± 6.80	1.36		
			95% CI of Difference	
L and H Paired Differences	1.67 ± 2.95	0.59	0.45 (lower)	2.88 (upper)

Note: Significance paired = <0.01 and significance 2-tailed = <0.01
SD = standard deviation CI = confidence interval
SEM = standard error of mean df = degrees of freedom

variability. Our study had limitations in the vocabulary size, nature of disability, and amount of data used. Hence, we may cautiously interpret our results to indicate that in a subset of dysarthria, the high-frequency components may not be contributing any new information to a Mel-cepstral HMM recognizer. Thus, a lower sampling rate maybe adequate for this subset. This has significant positive cost-benefit implications in a control application.

Additionally, we noted that the mean recognition of digit data was marginally higher than the mean for word data for both sets H and L. This conforms to previous research by Deller et al., who showed that mean word recognition was somewhat lower than mean digit recognition, due to phonemes that require extreme articulatory positions [8]. These are found more often in words owing to the greater number of phoneme combinations possible. In our study, even though words like “yes,” “my,” “no,” “go,” etc., are less complex or have similar complexity to the digit set, the presence of other words of greater complexity reduces the overall RR compared to the digit set. Moreover, we had a greater number of words (15) than digits (10), which also has an effect on mean RR.

CONCLUSION

This study investigated the usefulness of high-frequency spectral components toward recognition of dysarthric speech. Here, we used a Mel-cepstral-based HMM to develop a small vocabulary recognition system intended for a practical control application. We noted that no significant useful discriminant information was available to the system above 5.5 kHz in the data set. The result of this preliminary study provides clues to the direction that may be taken when dealing with such dysarthric speech. Here, the current set of data represents only a subset of dysarthria. Hence, investigation using similar techniques in data sets including data of severely dysarthric subjects with utterances having more vowel and consonants combinations might provide greater definition and validity. Data from three cerebral palsy individuals were used to test the model. However, the results may also be equally applicable to dysarthric individuals other than those with cerebral palsy, such as individuals with neurogenic communication disorders. Further development of this study could include investigation into a neural network/HMM hybrid structure. Such structures have been reported to provide equivalent or better performance than pure HMM structures in normal speech. They may also be more conducive to hardware implementation for control applications. The equivalent performance obtained at lower-band limiting/sampling rate for the current dysarthric speech set has a favorable cost-benefit implication, particularly in such a structure. This study is thus part of an effort to develop an artificially intelligent communication/control tool, either in stand-alone mode or in conjunction with other methods such as eye-tracking, etc., for speech- and motor-impaired individuals.

REFERENCES

1. Menendez-Padial X, Polikoff JB, Peters SM, Leonzio JE, Bunnell HT. The Nemours database of dysarthric speech. Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP), 1996 October 3–6, Philadelphia (PA). Wilmington (DE): Applied Science and Engineering Laboratories; 1996. p. 1962–65.
2. Gold CJ. Cerebral palsy—John Coopersmith Gold. Berkeley Heights (NJ): Enslow Publishers, Inc.; 2001.
3. Noyes JM, Frankish CR. Speech recognition technology for individuals with disabilities. Augmentative Altern Commun. 1992;8:297–303.
4. Sy BK, Horowitz DM. A statistical causal model for the assessment of dysarthric speech and the utility of computer based speech recognition. IEEE Trans Biomed Eng. 1993; 40(12):1282–98.
5. Patel R. Identifying information bearing prosodic parameters in severely dysarthric speech [dissertation]. University of Toronto (Canada); 2000.
6. Goodenough C, Rosen M. Towards a method for computer interface design using speech recognition. Proceedings of the 14th Annual Rehabilitation Engineering and Assistive Technology Society of North America (RESNA) Conference, 1991, Kansas City (MO). Arlington (VA): RESNA Press; 1991. p. 328–29.
7. Jayaram G, Abdelhamied K. Experiments in dysarthric speech recognition using artificial neural networks. J Rehabil Res Dev. 1995;32(2):162–69.
8. Deller JR Jr, Hsu D, Ferrier LJ. On the use of hidden Markov modeling for recognition of dysarthric speech. Comput Methods Programs Biomed. 1991;35(2):125–39.
9. Chen F, Kostov A. Optimization of dysarthric speech recognition. Proceedings of the 19th Annual International Conference of the Institute of Electrical and Electronics Engineers, Inc. (IEEE), 1997 October 30–November 2, Chicago (IL). Piscataway (NJ): IEEE; 1997. p. 1436–39.
10. Kain A, Niu X, Hosom J, Miao Q, Santen J. Formant resynthesis of dysarthric speech. Center for Spoken Language

Understanding, Oregon Graduate Institute School of Science and Engineering, Oregon Health and Science University, Portland (OR), 2004 [cited May 2005]. Available from: http://www.cslu.ogi.edu/~kain/tts2004/kain.pdf/

11. Mak B, Tam YC, Li Q. Discriminative auditory-based features for robust speech recognition. IEEE Trans Speech Audio Proc. 2004;12(1):27–36.
12. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE. 1989;77(2):257–86.
13. Jelinek F. Statistical methods for speech recognition. Language, speech and communication: a Bradford book. Cambridge (MA): The MIT Press; 1998. p. 15–54, p. 79–90, p. 147–163.
14. Kavcic A, Jose Moura MF. The Viterbi algorithm and Markovian noise memory. IEEE Trans Inf Theory. 2000; 46(1):291–301.

Submitted for publication June 16, 2004. Accepted in revised form October 22, 2004.