Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Assessment of dysarthric speech and an analysis on velopharyngeal incompetence

P. Vijayalakshmi and M. R. Reddy
Biomedical Engineering Division
Indian Institute of Technology, Madras, India.
Email: (pvijayalakshmi, rsreddy)@iitm.ac.in

Abstract— **In the present work the role of a speech recognition system in the assessment of dysarthric speech is studied. Initially a continuous speech recognition system is developed for the assessment of dysarthric speech. The performance of the continuous speech recognition system on the assessment of dysarthric speech is found to be dis-satisfactory due to greater number of insertions. Analysis conducted on duration of phoneme and speech rate confirms the fact that the more number of insertions in the recognizer output is due to lower speech rate of dysarthric speakers. To overcome the problem with the continuous speech recognition system an isolated-style speech recognition system is developed. The performance of this system on the assessment is compared with the Frenchay dysarthric assessment (FDA) scores provided with the Nemours speech corpus. From the performance of the isolated-style speech recognition system it is observed that apart from the articulatory problems, some of the speakers are affected with velopharyngeal incompetence also and is analyzed with group delay function based acoustic measure for the detection of hypernasality on dysarthric speech.**

I. Introduction

Dysarthrias are a family of neurogenic speech disorder that affect some or all of the sub-systems of speech such as laryngeal, velopharyngeal, and articulatory sub-systems. Dysarthric speech, depending on the extent and location of damage in the nervous system, may vary in intelligibility, rate of production, etc. Based on these aspects, researchers have focused on utilizing an automatic speech recognition (ASR) system for the assessment of dysarthric speech.

Sy and Horowitz [1] have developed a statistical causal model to provide a correlation between naive listeners' subjective judgments and the response of a dynamic time warping (DTW) based speech recognition system for the assessment of dysarthric speech. Pidal et al. [2] have compared the performance of a discrete hidden Markov models (DHMM) based speech recognition system and the listeners' judgment for the assessment of dysarthric speech. They [2] conclude that due to the low overall accuracy of the system it cannot be used for the assessment. Carmichael and Green [3] have proposed an intelligibility metric for dysarthria assessment based on probabilistic likelihood scores derived from the forced alignment of dysarthric speech based on whole-word HMMs.

The DTW, a template-based approach, neglects the second-order statistics which accounts for the variations at the frame level. To model the observations which are continuous vectors, it would be advantageous to use HMMs with continuous observation densities instead of discrete probability density within each state as in DHMM [4]. Hence continuous density hidden Markov models (CDHMM) will be a better choice to overcome the above mentioned problems. In this paper, the role of a CDHMM-based speech recognition system in the assessment of dysarthric speech is analyzed instead of a DTW [1] or a DHMM [2] based system.

In our work, initially a continuous speech recognition system is developed for the assessment of dysarthric speech. The performance of the continuous speech recognition system on the assessment of dysarthria is found to be dis-satisfactory due to a greater number of insertions. The greater number of insertions may be due to the variation in the rate of speech production of the dysarthric speakers. This is verified with duration and speech rate analysis of the dysarthric speakers, at the phonemic-level. An isolated-style phoneme (monophone-based) recognition system is developed to overcome the problem due to insertions in continuous speech recognition system. The speech recognition system performance is correlated with Frenchay dysarthric assessment (FDA) scores available with the Nemours database. Apart from providing information about the inactive articulators, the performance also gives a clue on problems associated with velopharyngeal sub-system of some of the dysarthric speakers. Previously we proposed an acoustic measure, based on a group delay function [5], [6] for the detection of hypernasality for the speakers with cleft palate. The same technique is applied here for the assessment of velopharyngeal dysfunction of dysarthric speakers.

The rest of the paper is organized as follows: The following section briefs the speech recognition system configuration used for the present study. In Section III the problems encountered due to continuous speech recognition system, the duration analysis on dysarthric speech and the assessment of dysarthric speech for the articulatory and velopharyngeal sub-systems are presented.

II. Experimental setup

The database used for the assessment of dysarthria consists of 10 dysarthric speakers' and one normal speaker's speech data from the Nemours database of dysarthric speech [7]. With this corpus time-aligned phonetic transcriptions are

1-4244-0033-3/06/$20.00 ©2006 IEEE.

given for all the dysarthric speakers' speech data. This database contains FDA scores for 9 dysarthric speakers. The phoneme-based ASR system is trained with normal speakers' speech data using both train and test data of TIMIT speech corpus to capture the normal characteristics of the phonemes. 38 left to right monophone continuous density hidden Markov models are trained with 3 effective states and 50 mixtures/state. The feature used for the present study is 13 dimensional Mel frequency cepstral coefficients (MFCC) + 13 dimensional delta coefficients + 13 dimensional acceleration coefficients. Features are cepstral mean subtracted to compensate for the different recording environments of the two speech corpora (TIMIT and Nemours). Features are extracted with 20 ms frame-size and 10 ms frame-shift.

III. Assessment of dysarthric speech

The problems associated with dysarthria are manifested in the form of acoustic variations from the normal speech. If the correlation between these variations and the performance of a speech recognition system is captured then the speech recognition system can be used for the assessment of dysarthric speech. For the present study, as the speech recognition system is trained with normal speech and tested with dysarthric speech, there might be variations in the acoustic realizations of the phonemes. Our current interest is to extract these variations in the acoustic realizations of the phonemes for the assessment of dysarthric speech.

A. Continuous speech recognition system

Initially a continuous speech recognition system is developed for the assessment of dysarthric speech. When the continuous utterances of dysarthric speech are used for testing, the following problems are encountered: (a) Invariably for all the dysarthric speakers, the accuracy[1] of this system is found to be negative. This is due to a greater number of insertions (thrice the number of phonemes, for the speaker **'bk'**). (b) Due to more insertions, the performance with respect to the number of phonemes correctly recognized, seems to be reasonably good. However, the response of the continuous speech recognition system does not correlate with the deviation in dysarthric speech. During training, since the normal speakers' speech data (TIMIT) is used, the durational information available in the self-transition probabilities[2] of the hidden Markov models capture only the duration of normal utterances. As the dysarthric speech is uncoordinated, the duration of each phoneme produced by the dysarthric speakers may be elongated depending upon the extent and the location of the damage of the nervous system. To verify these facts a duration analysis is conducted on normal and dysarthric speech.

[1]
$$Accuracy = 100 - \frac{deletion + substitution + insertion}{Total\ number\ of\ phonemes}\%$$

[2]
$$duration \quad d = \frac{1}{1 - a_{ii}}$$

B. Durational analysis

From the time-aligned phonetic transcription available with the Nemours and the TIMIT database the duration of each phoneme for each of the speaker is computed. The means and variances of the duration of dysarthric speech are normalized with respect to the normal speech. From these calculations it is observed that the normalized mean duration of dysarthric speakers, over all the phonemes, is always greater than ($\approx$ twice) that of the normal speakers. That is, the phoneme duration of dysarthric speakers are elongated than the duration of the normal speakers. The normalized variances of duration of phonemes are found to be varying from a minimum of 3 times to a maximum of 20 times greater than that of the normal speech. An interesting observation from the statistical analysis of duration of dysarthric speech is that, it provides information about the intelligibility of the dysarthric speakers. When the intelligibility scores from the FDA are correlated with the variances of the dysarthric speakers it is found that for 7 out of 10 dysarthric speakers the FDA scores correlates well with the statistical information and contradicts for the rest of the speakers. The variances and intelligibility scores from FDA are shown in Table I. For the present work, based on the intelligibility scores, the dysarthric speakers are rated as shown in column 1 of Table I. In Table I the speakers, for whom the variances are uncorrelated with the intelligibility scores are denoted by the symbol *.

TABLE I

Dysarthric speakers and their corresponding word and sentence intelligibility scores as found in FDA provided with the Nemours database and the normalized variances of the duration of phonemes

group	speakers	word	sentence	variance
mild	mh	8	4	2.7
	bb	4	8	3.6
	fb	-	-	4.1
	ll	4	4	7.7*
severe	bk	0	0	20.7
	sc	1	1	11.5
	bv	0	2	7.8*
moderate	jf	4	3	17.7*
	rk	4	1	5.3
	rl	4	3	8.3

Apart from duration, the speech rate (number of phonemes per second) for each dysarthric speaker and the normal speakers (TIMIT) are calculated. The speech rate of each dysarthric speaker is normalized with respect to the mean speech rate of the normal speakers. From the analysis, it is observed that the speech rate of dysarthric speakers varies from a minimum of 1.5 times to a maximum of 2.8 times less than that of the normal speech. i.e., the duration of phonemes produced by dysarthric speakers are much longer than the duration of the normal speech as observed in the previous analysis. Due to this durational variations between the normal (train) and dysarthric speech (test) data, there are a greater number of insertions in the recognized output

which causes a negative accuracy as discussed above. The relative levels of insertion errors can be controlled by tuning the insertion penalty during Viterbi decoding. As the speech rate of each dysarthric speaker is differing from one another, it is reflected as a greater variation in the amount of insertions among them. Tuning the insertion penalty for each and every dysarthric speaker separately is impractical.

To avoid such insertions due to global Viterbi alignment in continuous speech recognition system, an isolated-style speech recognition system is developed and is discussed in the following subsection. In isolated-style speech recognition system, since the phonemes are considered in isolation, only a state-level alignment has to be done.

C. Isolated-style phoneme recognition system

For this isolated-style speech recognition system, since the interest is shown only on finding out the acoustic-similarity of a test phoneme of a dysarthric speaker with the normal speakers' phoneme model, the decision-metric used is only the acoustic-likelihood of the phonemes for the given models. The overall performance of the system correlates well with the intelligibility which is not the focus of the current study. In the present work, focus is given only to the assessment of articulatory sub-system of dysarthric speech based on the performance of the phonemes corresponding to the place of articulation. Based on the place of articulation, the phonemes are classified into 5 groups, namely, (1) velar (**/k/ & /g/**), (2) palatal (**/ch/, /jh/ & /sh/**), (3) alveolar (**/t/ & /d/**), (4) dental (**/th/ & /dh/**) and (5) bilabial (**/p/, /b/ & /m/**). The rest of the phonemes are not considered for this task. For the assessment of articulatory sub-system, the performance of the isolated-style speech recognition system for phonemes corresponding to 5 classes of place of articulation are correlated [8] with the sum of scores of the relevant sections in FDA available with the Nemours database as shown in Figs. 1 and 2. For instance, for the bilabial class of place of articulation, sum of FDA scores of lips seal and lips in speech are correlated with the performance of bilabial class obtained from speech recognition system. For this analysis, a recognized-phoneme is considered to be correct even if it is confused with anyone of the phonemes belonging to the same group. For example, **/b/** is said to be recognized correctly if it is recognized either as **/b/ /p/** or **/m/**, as our focus is to find whether the speaker is able to articulate phonemes originating from that particular place of articulation rather than manner of articulation (voiced/unvoiced). The speaker 'fb' does not have FDA scores as his dysarthria is mild [7]. Therefore the comparison for the assessment is made only for the rest of the 9 speakers.

D. Performance Analysis

From the performance it is considered that, lower the performance the less active is the corresponding articulator. From the performance it is observed that,

- the speakers **'bk'** has all the articulators severely affected (refer Figs. 1 and 2) indicating that the speaker requires more attention towards all the articulators during speech therapy.
- The speaker **'mh'** has all the articulators functioning properly (refer Figs. 1 and 2).
- The speaker **'rk'** has all the articulators functioning moderately except bilabial which is found to be severe (refer Fig. 1(a)), indicating that the speaker requires much more attention for his lip movement than the others during speech therapy.
- For some of the speakers the performance of the speech recognition system does not seem to correlate with the FDA scores and are denoted by the arrows in Figs. 1 and 2.

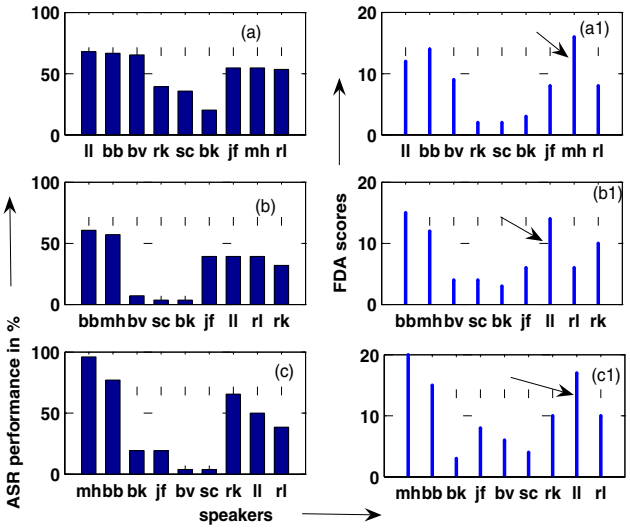

Fig. 1. Comparison of performance of the speech recognition system based on place of articulation and the FDA scores: ASR performance of (a) Bilabial, (b) Velar, (c) Palatal, (a1), (b1) and (c1) represent the corresponding FDA scores.

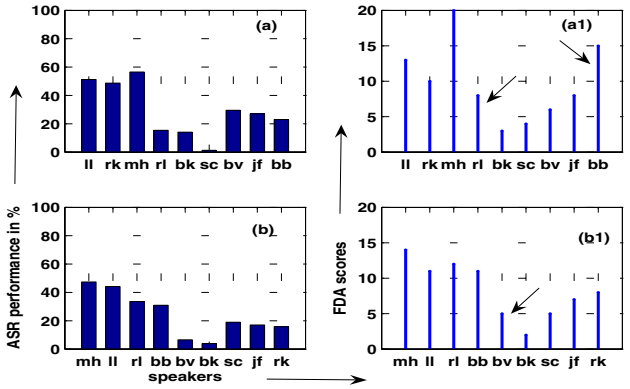

Fig. 2. Comparison of performance of the speech recognition system based on place of articulation and the FDA scores: ASR performance of: (a)Alveolar, (b) Dental and (a1) and (b1) represent the corresponding FDA scores.

From these observations, it is evident that the analysis on the place of articulation clearly indicates where exactly the speaker misarticulates, that in turn indicates the correspond-

ing articulator that is inactive or less active and requires more attention during speech therapy.

From the performance of speech recognition system for the bilabial class it is observed that for the speakers **'rl' 'rk' 'jf'** and **'bv'** most of the time the consonant **/b/** and **/p/** are recognized as their nasalized counterpart **/m/**. As far as the assessment of articulatory sub-system is concerned, as we are interested only in the origin of place of articulation, this misarticulation is considered to be correct. However, this gives a clue that the speaker might have velopharyngeal dysfunction. The velopharyngeal dysfunction in dysarthric speakers are due to the physiological defects and is called velopharyngeal incompetence. Subjects with velopharyngeal incompetence have anatomically intact soft palate, however due to neuromuscular weakness they are unable to achieve valving mechanism of the soft palate during production of oral sounds. Due to velopharyngeal incompetence, during the production of oral sounds additional nasal resonances are introduced into the oral resonances resulting in hypernasal speech.

E. Velopharyngeal incompetence in dysarthria

The dysarthric speakers analyzed in this study are affected either with cerebral palsy or head trauma with spastic dysarthria. For speakers with spastic dysarthria, though hypernasality typically occurs, it is usually not severe. For the analysis on hypernasality in dysarthric speakers, the vowel **/a/** following the bilabial consonants **/b/** and **/p/** are extracted manually. 25 examples for each of the dysarthric speaker is extracted for the analysis. To detect hypernasality, the group delay based acoustic measure proposed by the authors [5], [6] is utilized. For clarity some of the basic concepts used in [5], [6] are repeated here also. The group delay based acoustic measure (GDAM) is given by,

$$\begin{aligned} \text{GDAM} \;&>\; 1 \quad \text{hypernasal speech} \\ &<\; 1 \quad \text{normal speech.} \end{aligned} \tag{1}$$

The analysis on hypernasal speech uttered by speakers with cleft palate by the authors [5], [6] revealed the fact that an additional nasal frequency gets introduced into the oral resonance in the low frequency region around 250Hz below the first formant frequency due to velopharyngeal dysfunction. To analyze hypernasality in dysarthric speech, the speech signal, uttered by dysarthric speakers, is low-pass filtered around 800 Hz (F1 of /a/) and the formant frequencies are extracted using group delay function. The first two strongest peaks in the low frequency region are extracted with their corresponding group delay functions and their frequency locations.

For a hypernasal speech the F1 (first peak) is found to be the additional nasal frequency and F2 (second peak) is found to be the first formant of the corresponding vowel (refer Fig. 3(a)). In case of normal speech F1 is due to pitch harmonics and F2 is the first formant of the vowel (refer Fig. 3(b)). Hence, the ratio of group delay of F1 to F2 for hypernasal speech will always be greater than 1 and for

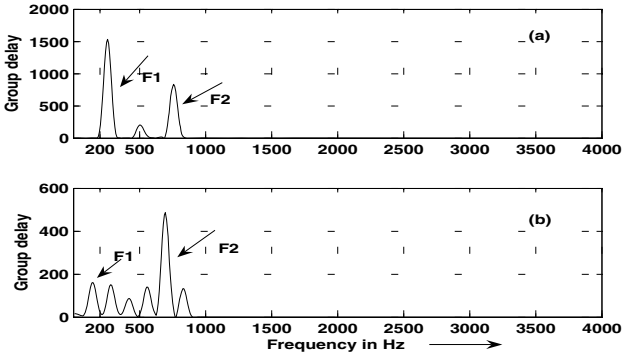

Fig. 3. Group delay spectrum of (a) hypernasal speech (b) normal speech

normal speech since the group delay value of pitch harmonics is less than that of F2, the ratio is found to be always less than 1. The group delay ratio of F1 to F2 is computed for all the dysarthric speakers and normal speaker **'jp'** from Nemours database. It is observed that only the speakers **'rl', 'rk' 'jf'** and **'bv'** are found to be hypernasal. Especially for the speaker **'rl'** 100% of the utterances are found to be hypernasal. For the other 3 speakers only around 60 to 75% of the utterances showed hypernasality. This variation is as observed in the recognition performance of **/b/** and **/p/**.

IV. Conclusions

In this study, the problems encountered due to continuous speech recognition system in the assessment of dysarthric speech are analyzed and verified by duration analysis. To overcome the problems, an isolated-style speech recognition system is suggested. The performance of speech recognition system provided clue regarding velopharyngeal dysfunction of some of the dysarthric speakers apart from the assessment of articulatory sub-system. Using the group delay function based acoustic measure the hypernasality in dysarthric speakers is also verified.

References

[1] Sy, B. K., and D. M. Horowitz, "A statistical causal model for the assessment of dysarthric speech and the utility of computer based speech recognition," *IEEE Trans. on Biomedical Engineering*, vol. 40, no. 12, pp. 1282–1298, Dec. 1993.

[2] Pidal, X. M., J. B. Polikoff, and H. T. Bunnell, "An HMM-based phoneme recognizer applied to assessment of dysarthric speech," in *Eurospeech*, 1997, pp. 1831–1834.

[3] Carmichael, J., and P. Green, "Revisiting dysarthria assessment intelligibility metrics," in *Proceedings of Int. Conf. Spoken Language Processing*, Oct. 2004, pp. 485–488.

[4] Rabiner, L. R., and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, New Jersey, 1993.

[5] Vijayalakshmi, P., and M. R. Reddy, "The analysis of band-limited hypernasal speech using group delay based formant extraction technique," in *Eurospeech*, Lisbon, Portugal, Sept. 2005, pp. 665–668.

[6] Vijayalakshmi, P., and M. R. Reddy, "Detection of hypernasality using statistical pattern classifi ers," in *Eurospeech*, Lisbon, Portugal, Sept. 2005, pp. 701–704.

[7] Pidal, X. M., J. B. Polikoff, S. M. Peters, J. E. Leonzio, and H. T. Bunnell, "The Nemours database of dysarthric speech," in *Proceedings of Int. Conf. Spoken Language Processing*, 1996, pp. 1962–1965.

[8] Vijayalakshmi, P., and Douglas O'Shaughnessy, "Assessment of dysarthric speech using a CDHMM-based phone recognition system," in *IFMBE Proceedings ICBME, in CD*, Singapore, Dec. 2005, vol. 12.