Assessment of articulatory sub-systems of dysarthric speech using an isolated-style phoneme recognition system

P. Vijayalakshmi[1], *M. R. Reddy*[1], *and Douglas O'Shaughnessy*[2]

[1]Biomedical Engineering Division
Indian Institute of Technology, Madras, India.
[2]INRS-EMT, University of Quebec, Montreal, Canada.
(pvijayalakshmi, rsreddy)@iitm.ac.in, dougo@emt.inrs.ca

Abstract

In this work, the variation in the acoustic realizations of phonemes between normal and dysarthric speech is utilized for the assessment of articulatory sub-systems of dysarthric speech. Using a speech recognition system for the assessment of dysarthric speech is well established either by considering the continuous speech as a sequence of phonemes, or by considering isolated words. These systems will provide information about the intelligibility alone. However, the problems associated with the sub-systems of speech can be well captured when a dysarthric speaker is asked to speak continuously and the resultant speech is analyzed at phoneme-level in isolation. Considering this aspect, in our work, an isolated-style triphone-based phoneme recognition system is developed for analyzing continuous speech. Acoustic variations of the phonemes based on place of articulation alone provides information associated with the malfunctioning of articulatory sub-system and are correlated with the Frenchay dysarthric assessment (FDA) scores. The correlation error between our system and FDA scores is found to be only 9%.

Intex Terms: Speech disorder, speech recognition system, triphone, dysarthria

1. Introduction

Dysarthria is a family of neurogenic speech disorders that interferes with the basic processes of speech production such as phonation, articulation, and prosody. These interferences affect some or all of the sub-systems of speech such as laryngeal, velopharyngeal, and articulatory sub-systems. This results in unintelligible, slow, imprecise or uncoordinated speech, impeding effective and efficient communication. To improve the communication ability of the dysarthric speakers it is necessary to evaluate the malfunctioning of the speech sub-systems.

Current clinical assessment methods involve subjective judgment of dysarthric speech. Evaluation based on subjective criteria requires experienced or well trained listeners for the assessment and is frequently hampered by a lack of consensus among experts [1]. This necessitates an automatic scheme that can avoid human interface during the evaluation of dysarthric speech.

If an ASR system is trained with normal speakers' speech data, the system learns the correct utterance and if it is tested with dysarthric speech data, it may clearly indicate the speech impairment rating. Based on these aspects, researchers have focused on utilizing an automatic speech recognition (ASR) system for the assessment of dysarthric speech ([1], [2], [3], [4]). The techniques utilized in these studies, based on the overall performance of the speech recognition system, will be able to provide information about the intelligibility of a given dysarthric speech. However, during testing of such a system on continuous speech, during global Viterbi alignment, due to the variation in speech rate between normal and dysarthric speakers more insertions may occur. This may cause poor performance of the system. 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.

In isolated word-based speech recognition system ([2], [3]) or in a distance measure-based system ([1], [4]), the vocabulary should be carefully selected to derive information for the assessment of speech sub-systems. The recognition has to be performed in closed set and the set of words should have only few different phonemes. Apart from this, in an isolated word-based system the speakers will be asked to speak the words in isolation. In case of mild or moderate dysarthric speakers if they speak words in isolation, they may produce it more clearly than in continuous speech, which may not give a clear picture about their problems in the speech sub-systems.

Based on these aspects, to derive information about the problems associated with the sub-system of speech the dysarthric speakers should utter words continuously, at least some simple sentences. The malfunctioning of speech sub-systems will be better portrayed at the phonemic level. That is, the assessment has to be performed at the phonemic level in isolated style.

In the current work, an isolated-style phoneme recognition system is developed to overcome the issues in the assessment of dysarthric speech. Context-independent (CI) phonemes (monophones) have sufficient training data to generate robust models. However, CI phonemes occurring at different contexts are not similar. Instead, a triphone is a context-dependent (CD) phoneme which takes both the left and right phonetic contexts into consideration, thus avoiding the variability. Triphone modeling with more numbers of examples are more powerful than monophone models, as it models the co-articulatory effect. For the present study, contexts are derived from the Nemours database of dysarthric speech. Triphone models are trained for these contexts from the TIMIT speech corpus. As the segmented data is available, in the current study focus is shown to find out whether the isolated phonemes give information about the malfunctioning of articulatory sub-systems. This can be carried out by correlating the performance of

the speech recognition system based on place of articulation and the Frenchay dysarthric assessment (FDA) scores provided with the Nemours database.

The outline of the paper is follows: In Section 2 the experimental setup for the current study is described. Section 3 describes the assessment of articulatory sub-system of dysarthric speech based on an isolated style phoneme recognition system.

2. 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 [5]. With this corpus time-aligned phonetic transcriptions are given for all the dysarthric speakers' speech data. This database contains FDA scores for 9 dysarthric speakers. FDA is a well-established test for the diagnosis of dysarthria. The test is divided into 11 sections, namely, reflex, palate, lips, jaw, tongue, intelligibility, etc. Each dysarthric speaker is rated on a number of simple tasks. In FDA, a score of **'8'** represents normal function and **'0'** represents no function. The list of speakers and their intelligibility scores as in FDA are shown in Table 1. From the intelligibility scores in FDA, for our work the dysarthric speakers are grouped as shown in Table 1. (Hereafter the identity of the dysarthric speakers will be represented by the alphabets as in column 1 of Table 1).

Table 1: Dysarthric speakers and their corresponding word and sentence intelligibility scores as found in FDA provided with the Nemours database

	group	speakers	word	sentence
a	mild	bb	4	8
d		fb	-	-
f		ll	4	4
g		mh	8	4
b	severe	bk	0	0
j		sc	1	1
c		bv	0	2
e	moderate	jf	4	3
h		rk	4	1
i		rl	4	3
n	normal	jp	-	-

The phoneme-based ASR system is trained with normal speakers' speech data using both train and test data of the TIMIT speech corpus to capture the normal characteristics of the phonemes. For the present study, the contexts are derived from the Nemours database of dysarthric speech. Word-internal triphones alone are considered. There are 231 word-internal triphones derived from dysarthric speech data. Out of the 231 triphones only 128 had more than 10 examples in the TIMIT speech corpus. The triphone training procedure essentially involves the following steps:

- Generation of monophone models with a nominal number of states and a single mixture/state, and re-estimation of these models.
- Creation of triphone transcriptions from monophone transcriptions.
- Initial triphone training by cloning the single-mixture monophone models. Re-estimation of the cloned triphone models.
- Triphone clustering. For the present study triphones are clustered and acoustically similar states are tied using a tree-based clustering procedure.
- Splitting the single mixture Gaussian distributions by a divide-by-two algorithm. Re-estimation of these triphone models.

The CI models are generated with 3 state left-to-right continuous density hidden Markov models without skip states, with a single mixture/state. The phoneme model parameters are re-estimated using the Baum-Welch re-estimation method. Finally, the triphone models are split into 8 mixtures/state. The features used for the present study are 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.

3. Assessment of dysarthric speech

Dysarthria refers to speech problems that affect one or more subsystems of speech. This is manifested as an acoustic deviation from normal speech. If the correlation between this deviation and the response of an ASR system is evaluated, then the response of an ASR system can be used for the assessment of dysarthric speech. In our present work, the ASR system is trained with the TIMIT speech corpus and tested with dysarthric speakers' speech data with varying degrees of dysarthria. Training the ASR system with a normal speech corpus and testing with a dysarthric speech corpus may have the following problems: (a) different recording environments, (b) variations in the acoustic realization of the phonemes. As mentioned in the previous section, the channel variations are, to a certain extent, compensated by cepstral mean subtraction of the features extracted. For the present study, specific interest is shown to extract the variations in the acoustic realizations of the phonemes.

3.1. Isolated-style phoneme recognition system

In order to build an isolated-style phoneme-based ASR system, the test data should be segmented into phonemes *a priori*. The test corpus (Nemours) provides the time-aligned phonetic transcriptions along with the speech data. However, in the actual test environment, only a raw speech signal will be available without segment boundaries. To simulate a similar test environment, phonetic boundaries of the speech signal has to be derived. Given the phoneme models, the speech signal, and the corresponding phonetic transcription, phonetic boundaries can be automatically derived using a forced-Viterbi alignment procedure. Initially, the speech data of all the dysarthric speakers are segmented automatically using the models trained from the TIMIT speech corpus. The resultant boundaries are checked with the boundaries available with the corpus. Here, a boundary is considered as erroneous if the error between the actual boundary and the derived boundary is more than 10 ms. For the whole test corpus, the performance of this segmentation method is only 60%. Improving the accuracy of this segmentation approach is taken up for future study. As our current focus is to find out the significance of isolated

phonemes in the assessment of dysarthric speech, for the present work segment boundaries provided with the database are used and a separate inventory is created for each of the dysarthric speakers. For this isolated-style speech recognition system, since the interest is shown only in 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. With this system two tests are conducted: (i) Assessment of intelligibility based on the overall performance of the system and (ii) Assessment of articulatory sub-systems of dysarthric speech based on the performance of phonemes grouped based on place of articulation.

3.2. Assessment of Intelligibility

All the phonemes in the test utterances are recognized in isolation separately and the average performance over all the phonemes is computed for each of the dysarthric speakers. The performance (over all the phonemes) of the isolated-style speech recognition system for each of the dysarthric speakers' speech data along with the corresponding intelligibility (word + sentence) score (in FDA) provided with the corpus is shown in Fig. 1. With the present framework, invariably for all the dysarthric speakers the overall performance shows an improvement from a minimum of 1.3% to 5% over the previous study [6] with pure monophones. From Fig. 1, one can observe that the recognition performance is proportional to the intelligibility of dysarthric speech. Hence the ASR system trained with normal speech and tested with dysarthric speech can be used for the assessment of dysarthria. For the dysarthric speaker **'d'** FDA scores are not available since his dysarthria is mild [5]. The performance of the normal speaker **'n'** is considered as reference.

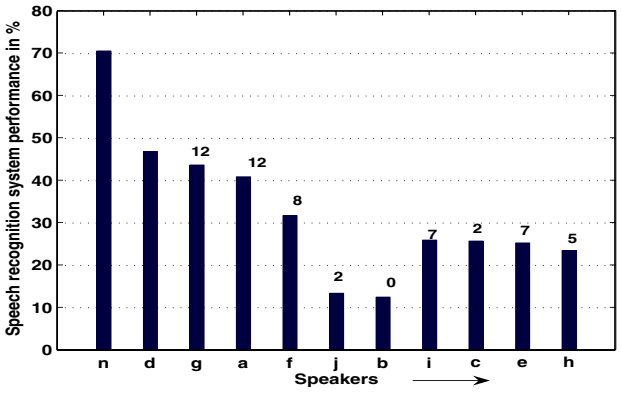

Figure 1: Overall performance of the isolated phoneme recognition system along with the (word + sentence) intelligibility scores of the FDA (shown on the top of the bars) showing the severity of the disease. The ASR system's recognition performance is arranged as mild followed by severe and moderate.

For the above experiment, for each speaker, the recognition performance is computed by taking the average performance of all the phonemes. As mentioned earlier averaged-performance correlates well with the intelligibility (except for the speaker **'c'**) scores in FDA. However, it does not provide any information about the inactive articulators. To derive further information regarding the problems with a specific articulator, the phonemes are grouped based on the place of articulation and the recognition-performance is analyzed as discussed below.

3.3. Assessment of articulatory sub-systems

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 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, **/k/** is said to be recognized correctly if it is recognized either as **/k/** or **/g/**, 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).

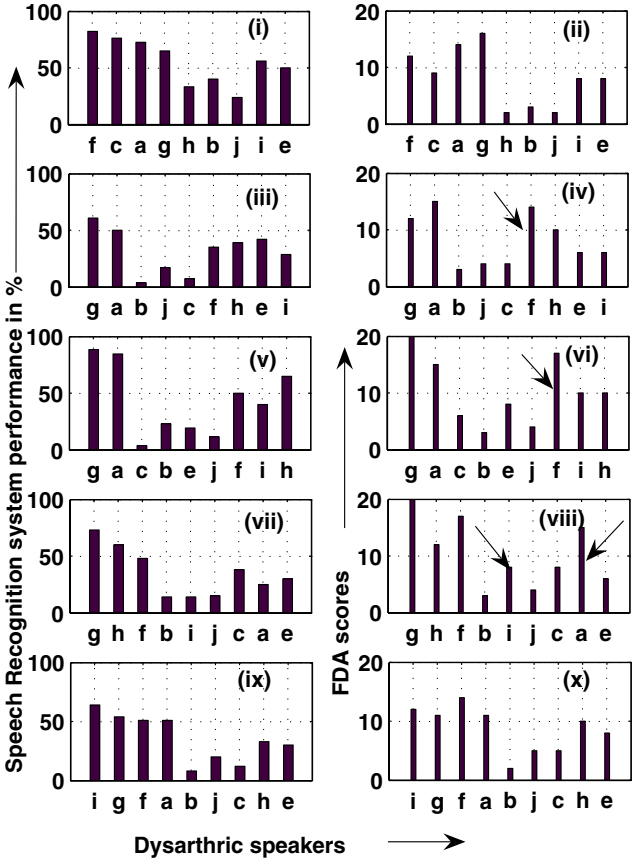

Figure 2: Comparison of performance of the speech recognition system based on place of articulation and the FDA scores: ASR performance of (i) Bilabial, (iii) Velar, (v) Palatal, (vii) Alveolar (ix) Dental and (ii), (iv), (vi), (viii) and (x) representing the corresponding FDA scores.

The performance of the isolated-style speech recognition system for the 5 groups are correlated with the sum of scores of entries from relevant sections in the FDA as shown in Fig. 2. For in-

stance, for the palatal sounds the sum of the FDA scores of palate in speech, tongue elevation, and tongue in speech and the performance of the palatal class obtained from the speech recognition system are compared. As speaker **'d'** does not have FDA scores the comparison for the assessment is made for the rest of the 9 dysarthric speakers. In Fig. 2 for clarity, the performances are arranged as mild followed by severe and moderate. The observations and interpretations from the correlation between the performance of the ASR system for the 5 classes of place of articulation and the FDA scores are summarized as follows:

- Speakers **'b'** and **'j'** have all their articulators affected severely (refer to Fig. 2), which implies that they require more attention in speech therapy for articulatory movements.
- Speaker **'h'** has the bilabial movement more severely affected (refer to Fig. 2(i)) than the other articulators and requires an improvement for labial movement rather than the other articulators.
- Similarly, speaker **'e'** has his palatal movement severely affected (refer to Fig. 2(v)) with all other articulators moderately functioning thus needs more attention for palatal movement during speech therapy.
- Speaker **'g'** has all his articulators functioning properly which is reflected in the performance as **mild**, and does not require serious attention towards his articulatory movement.
- For some of the speakers the performance of the ASR system does not seem to correlate with the FDA scores and these are denoted by arrows in the Fig. 2 (iv), (vi) and (viii). That is, out of 45 entries (9 speakers and 5 classes) 4 entries are found to be uncorrelated. That is, it shows a correlation error of only 9% and it is an improvement by 4% over the pure monophone-based assessment system described in [6].

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

4. Conclusions

In this paper, an efficient phoneme (triphone) recognition system for the assessment of dysarthric speech is described. It is suggested that if the dysarthric speaker is asked to speak continuously and the recognition is performed at the phonemic-level in isolated-style, then the problems associated with the articulatory sub-system of speech can be better captured. If the performance of the phoneme segmentation method is perfected, a fully automated system can be realized for the complete assessment of dysarthric speech.

5. References

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

[2] 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.

[3] 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.

[4] Gu, L., J. G. Harris, R. Shrivastav, C. Sapienza, "Disordered speech evaluation using objective quality measures," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 2005, vol. 1, pp. 321–324.

[5] 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.

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