The Nemours Database of Dysarthric Speech

Xavier Menéndez-Pidal, James B. Polikoff, Shirley M. Peters, Jennie E. Leonzio, H. T.Bunnell
{menendez, polikoff, peters, leonzio, bunnell}@asel.udel.edu

Applied Science & Engineering Laboratories (ASEL),
A.I. duPont Institute, P.O. Box 269, Wilmington, DE 19899, USA

ABSTRACT

The Nemours database is a collection of 814 short nonsense sentences; 74 sentences spoken by each of 11 male speakers with varying degrees of dysarthria. Additionally, the database contains two connected-speech paragraphs produced by each of the 11 speakers. The database was designed to test the intelligibility of dysarthric speech before and after enhancement by various signal processing methods, and is available on CD-ROM. It can also be used to investigate general characteristics of dysarthric speech such as production error patterns. The entire database has been marked at the word level and sentences for 10 of the 11 talkers have been marked at the phoneme level as well. This paper describes the database structure and techniques adopted to improve the performance of a Discrete Hidden Markov Model (DHMM) labeler used to assign initial phoneme labels to the elements of the database. These techniques may be useful in the design of automatic recognition systems for persons with speech disorders, especially when limited amounts of training data are available.

1. INTRODUCTION

Dysarthrias are a family of neurologically based speech disorders. Speech produced by dysarthric talkers can be difficult to nearly impossible for persons unfamiliar with the speaker to understand. Although speech therapy may help dysarthric talkers improve their speech intelligibility, therapy cannot be expected to restore "normal" speech quality. For such talkers, a speech prosthetic device which accepted the dysarthric talkers' speech as input and produced more intelligible speech output would be very desirable.

Several investigators have attempted to develop speech prostheses of this sort by coupling off-the-shelf or laboratory speech recognizers with speech synthesizers. The speech recognizer would be trained to the characteristics of a single talker and its output synthesized to create speech that is easier to understand than the dysarthric talker's natural speech. However, such systems have not been particularly successful. In large part this is due to the fact that dysarthric speech is quite variable making it difficult to recognize more than a small number of distinct words from a given talker.

There are other drawbacks to the recognition driven synthesis approach as well. First, template-based word recognizers force the use of a (possibly small) finite vocabulary thus constraining the talker to stay within the bounds of the vocabulary in conversation. Secondly, the synthetic "voice" is perceived as less desirable than natural speech by many talkers.

As an alternative to recognition driven synthesis, we have been examining the possibility of using recognition technology to screen dysarthric speech for patterns associated with articulation errors common to a particular talker and, where possible, use signal processing to "repair" the talker's acoustic speech signal. This approach has the advantage of not being bounded by a finite recognition vocabulary (in the lexical domain) and of retaining the talker's natural voice.

Of course, the approach we propose will only be practical if the articulation errors of a talker form a finite and recognizable set of acoustic patterns. Partly to determine if this appeared to be the case for a set of dysarthric talkers, and partly to test some preliminary methods of speech enhancement for dysarthric speech, the speech data for the Nemours database was collected.

2. DATABASE STRUCTURE

Each nonsense sentence in the database is of the form "The X is Ying the Z." Specific sentences were generated by randomly selecting X and Z ($X \neq Z$) without replacement from a set of 74 monosyllabic nouns and selecting Y without replacement from a set of 37 disyllabic verbs. This process produced 37 sentences from which another 37 sentences were generated by swapping the X and Z tokens in the original set. Thus, over the complete set of 74 sentences, each noun and verb was produced twice by each talker.

The target words (X, Y and Z) were chosen based on constraints similar to those used by, for example, Kent, *et al*. (1990) to provide closed-set phonetic contrasts (e.g., place, manner, and voicing contrasts) within an associated set of four to six words. Thus, all of the target words within a set differ in a single phoneme so that they may be used as alternatives in a closed-response perceptual test for intelligibility. Each talker also recorded two paragraphs of connected speech: the "Grandfather" passage and the "Rainbow" passage. Both are commonly used speech passages. One non-dysarthric talker recorded the entire speech corpus as a control.

3. RECORDING PROCEDURE

A recording session consisted of three segments: an initial assessment session conducted by a speech pathologist, recording of the 74 nonsense sentences, and recording of the two speech passages. All parts of the session took place in a small sound dampened room with the talker seated (typically in his wheel chair) next to the speech pathologist or experimenter and in front of a table mounted microphone (Electro-Voice RE55) connected to a digital audio tape recorder (Sony PCM-2500). The session began with administration of the *Frenchay Dysarthria Assessment* (Enderby,

1983) by a speech pathologist. Following this assessment and a short break, the experimenter entered the room to assist in the speech recordings. First the nonsense sentences and then the speech passages were recorded with additional breaks as needed by the talker. Nonsense sentences were written in large print on a sheet placed in front of the talker and each sentence was read first by the experimenter and then repeated by the subject. This assisted all talkers in pronunciation of words and was essential for some subjects with limited eyesight or literacy. On average the entire recording session was completed in two and one half to three hours, including time for breaks.

Subsequent to the recording sessions, all speech materials were digitized from the audio playback of the DAT recording using a 16 kHz sampling rate at 16-bit sample resolution with appropriate low pass filtering. Each nonsense sentence was saved in a separate file in standard RIFF format. For convenience, the longer speech passages were also broken down into sentences and each sentence saved as a separate waveform file. However, for these materials care was taken to ensure that inter-sentence timing was preserved to allow sentences to be concatenated to restore the original paragraph length recording exactly.

As part of a series of experiments, perception data was collected on each sentence. In these experiments a minimum of five listeners identified the target words in the nonsense sentence in a closed response set task. Over a series of data collection sessions, individual listeners heard each sentence a total of 12 times and the distribution of the listeners' responses over the response alternatives associated with each word forms the basic perceptual data in the database. Average percentage correct identifications for each talker are given in Table 1 along with other speaker characteristics derived from the Frenchay speech assessment. Note that a speech assessment was not conducted for talker FB whose dysarthria was extremely mild, and perceptual data were not collected on talker KS whose dysarthria was severe.

Talker		BK	SC	BV	FB	RK	BB	KS	RL	MH	LL	JF
Tongue In Speech	Assessment Scores	1	2	2	_	2	7	1	4	6	7	4
Laryngeal In Speech		2	1	2	_	1	8	0	3	7	6	3
Respiratory In Speech		1	1	5	_	3	8	4	4	8	6	4
Intelligibility Conversation		3	3	3	_	2	8	1	4	6	6	4
Classification		AQ	SAQ	SQ	SA	SbQ	SQ	HTQ	SA	SA	SA	SA
Initial Cons.	% correct	71.3	50.3	53.7	94.8	69.0	91.7	-	79.2	93.6	91.8	77.7
Final Cons.		43.2	46.0	60.4	95.3	76.3	86.3	-	68.2	89.2	74.8	81.1
Inter V Cons.		50.5	49.9	53.7	86.6	50.7	86.9	-	58.2	90.4	77.7	71.6
Overall		58.2	51.5	57.5	92.9	68.6	89.7	-	73.3	92.1	84.4	78.5

Table 1: Average Percentage Correct Identifications With Frenchay Dysarthria Assessment Scores

4. LABELING PROCEDURE

To aid in the comparison of acoustic and perceptual data within the database, and to provide tags necessary for training various recognizers, the database has been labeled at both the word and phoneme levels (with the exception of talker KS whose speech did not allow phonemic labeling). Word-level labels were assigned manually, however, the phoneme-level labels were assigned using a DHMM labeler followed by manual inspection and correction.

4.1. Transcription

The automatic labeling procedure aligned a broad phonemic transcription of the talkers production which included syllabic false-starts and dysfluencies, but did not attempt to encode any non-speech events (e.g., grunts, breath sounds, etc.) in the sound stream.

The transcription of the database used a set of 39 segment labels derived from the ARPAbet symbol set. Additionally, during the manual examination and correction stage, disfluent segments were identified and flagged by the addition of a character to the phoneme symbol.

4.2. Automatic Labeler

Labeling of the entire database progressed in two stages. In the initial stage, the nonsense sentences (for which individual words had been manually labeled) were assigned phonemic labels using the DHMM labeler. For these sentences, initial phoneme labels were assigned by assuming that all segments within a word were of equal duration. This initial labeling was used to train the DHMM labeler which, at the same time, adjusted the label boundaries to improve the fit of the phoneme models to the data. In the second stage, labels were assigned to the paragraph materials, and at the same time, optimized over both the nonsense sentence and paragraph speech data.

Discrete Hidden Markov Model (DHMM) technology was chosen because it is well suited to recognition of a limited training size database. The system was adapted from a DHMM labeler developed by the Speech Group (GTH) of the Madrilenian Telecommunication School (ETSIT-UPM) (Ferreiros, *et al*. 1993). The main system is composed of a spectral analysis front-end, a vector quantizer system, and a DHMM Viterbi algorithm used to align the phonetic labels and to train the 39 phoneme models. Several variations in these components were explored to find combinations which provided the most accurate labeling performance. These variations are described in the following sections and summarized along with performance results in Table 2.

Spectral Analysis. In the pre-processing stage, 13 mel-cepstral coefficients and the log energy of the acoustic frame were estimated every 10 ms. Additionally, the contribution of dynamic information was assessed by using the first and second time-derivatives (designated ΔCep and Δ^2Cep respectively in Table 2) of the 14 static coefficients in some analyses.

Vector Quantizer. For each speaker, 3 codebooks were developed using a hard Vector Quantization (VQ) process for the cepstral, ΔCep and Δ^2Cep coefficients respectively. A standard Mahalanobis distance measure was used for computing distortion and 128 centroids per quantizer appeared to provide enough accuracy and generality in the DHMM system.

An alternative training procedure derived from the classical LBG algorithm (Linde, *et al.* 1980) was developed to improve the generalization capabilities of the quantizer. One of the weaknesses of the LBG procedure is that it tends to produce centroids with very irregular numbers of observations per centroid. This is due to the unorganized splitting criteria adopted in the training procedure. In the LBG method, after optimizing each centroid using the K-means algorithm the size of the codebook is doubled by splitting each centroid, adding a new centroid in the vicinity of the previous one. In the present adaptation of the LBG algorithm, a guided splitting mechanism has been introduced which is sensitive to the distribution of observations among the centroids. In this procedure, the centroids are split proportionally to the number of observations assigned to each centroid. If a centroid has fewer observations than the mean number of observations per centroid it is not split. Further, if the number of observations in a centroid is four times higher than the average, the centroid is split by adding four new centroids in its vicinity.

Figure 1 illustrates the evolution of the global distortion of the VQ, using the LBG and the proportional splitting procedure. The global distortion is the sum of the distances of each observation to its nearest centroid divided by the number of observations (see Eq.1).

$$D_G = \frac{1}{N_{Obs}} \sum_{i}^{N_{Obs}} \underset{j}{Min}\ (dist_{ij}) \qquad (1)$$

Both algorithms appear to provide very similar global resolution and the global distortion of the VQ has very similar behavior (see Figure 1). In Figure 1, the evolution of the average of the mean distortion of all the centroids is also illustrated. This distortion measure (see Eq. 2) evaluates the average local resolution of all centroids.

$$\bar{D} = \frac{1}{N_{Cent}} \left[\sum_{j}^{N_{Cent}} \frac{1}{Num_{C_j}} \left(\sum_{i \in C_j}^{Num_{C_j}} dist_{ij} \right) \right] \qquad (2)$$

Using the proportional splitting procedure, the local average resolution of the centroids was improved by approximately 10 to 20% as is shown in Figure 1. Also, a much more regular VQ structure was obtained, reducing the standard deviation of the number of observation per centroids by about 60%.

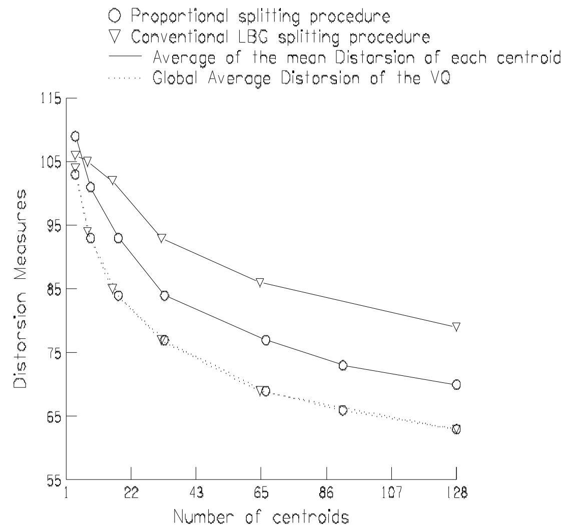

Figure 1: Distortion measures using different splitting criteria

DHMM topology. Several HMM topologies varying in the number of states per phoneme model were tested (see Figure 2 for an illustration of a 6-state model), resulting in the choice of a topology with a relatively large number of states. This may be related to the structure of the dysarthric speech which tends to have long, poorly differentiated phonetic units. The use of a long HMM topology imposes some time constraints and forces the system to locate long phonetic zones reducing the probability of producing a wrong alignment. Experiments reported by Deller, *et al*. (1991) lead to similar conclusions for isolated word recognition for dysarthric talkers. As Table 2 shows, an 8-state model marginally outperformed a comparable 6-state model while the 6 state model substantially outperformed a comparable 3-state model.

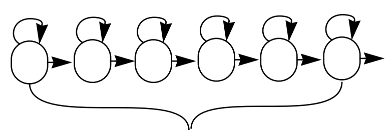

Figure 2: HMM having 6 states per phoneme

As the Nemours database contains continuous speech, the labeler must accommodate the possible--but not necessary--presence of pauses or silence between words. To minimize problems related to word boundaries, strategies similar to those used for continuous speech labeling of normal speech have been adopted. A generic silence model with only 1 state and a double forward connection was used in the last state of each word. This allowed the labeler to jump the silence model if no silence was detected or insert as many silence frames as needed if silence was found between two words (see Figure 3).

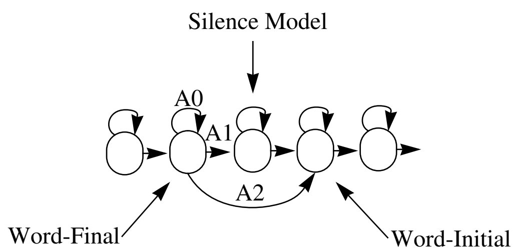

Figure 3: Silence and Word-HMM concatenation topology

When the probability of jumping the silence model (A2) was directly estimated from a context independent phoneme model, the value obtained was very low causing incorrect insertion of silence. This problem was apparently caused by the relatively infrequent occurrence of any particular phoneme in word-final position. To compensate, a new parameter (Pjump, probability of jumping a silence) was estimated. Pjump was simply estimated as the number of word junctures without silence divided by the total number of word junctures ($N_{Juncture}$) (see Eq 3.)

$$Pjump = \frac{N_{Juncture} - N_{Silent}}{N_{Juncture}} \qquad (3)$$

Using Pjump the forward probability at the end of a word could be decomposed to estimate more accurate A1' and A2' (see Eq.4) reducing the insertion of spurious silence.

$$\begin{aligned} A2' &= (A1 + A2)Pjump \\ A1' &= (A1 + A2)(1 - Pjump) \end{aligned} \qquad (4)$$

The false silence insertion in the labeler was reduced using the double forward connection at the end of each word from 11% insertion errors to 2% insertion errors when including the Pjump probability in the system.

Smoothing technique. The HMM training was performed in two phases. First, the models were initialized as isochronous segments within the hand-labeled word boundaries of the nonsense sentences. The phoneme HMMs were then trained on these data by successively improving the fit of the initial models over five Viterbi itterations of the system. The labeling procedure was then completed using the HMMs trained on the nonsense sentences with five additional Viterbi iterations over the entire database to adjust the initial HMM models and assign labels to the paragraph data. Since there were only about 1200 phonemes, a progressive smoothing technique was adopted to help the Viterbi algorithm better locate the phoneme boundaries. For each VQ, the minimum probability of emission of the HMMs was used as a threshold probability to effectively smooth the HMM models instead of using a fixed threshold. To complete the lack of data existing in the initial training set the threshold was raised by a factor of two in the initial Viterbi iteration. The threshold was then lowered by a factor of 0.75 in each successive Viterbi iteration. As Table 2 illustrates, this progressive smoothing approach performed better than fixed smoothing for comparable 6-state models.

States per phoneme	Codesets	Smoothing	%Correct phoneme Labeling
6	Cep	Progressive	88.5%
6	Cep+ΔCep	Progressive	92.2%
6	Cep+ΔCep+Δ^2Cep	Progressive	**93.3%**
3	Cep+ΔCep+Δ^2Cep	Progressive	86.0%
6	Cep+ΔCep+Δ^2Cep	Progressive	**93.3%**
8	Cep+ΔCep+Δ^2Cep	Progressive	**93.8%**
6	Cep+ΔCep+Δ^2Cep	Fixed	91.8%

Table 2: Percentage correct label assignments for two speakers in the paragraph data (~1500 phonemes) for tested model topology, different VQ codesets and different smoothing techniques.

5. REFERENCES

1. Deller, J.R., Hsu, D., Ferrier, L.J. “On the Use of Hidden Markov Modeling for Recognition of Dysarthric Speech,” *Computer Methods and Programs in Biomedicine,* vol 35, no. 2, 1991.

2. Enderby, Pamela M. “Frenchay Dysarthria Assessment,” College Hill Press*,* 1983.

3. Ferreiros, J., de Cordoba, R., Pardo, J.M. “Continuous Speech HMM Training System: Applications on Speech Recognition and Phonetic Labels Alignment,” *Proc. of NATO-ASI*, 1993.

4. Kent, R.D., Weismer, G. Kent, J.F., and Rosenbek, J.C., “Toward Phonetic Intelligibility Testing in Dysarthria,” *Journal of Speech and Hearing Disorders*, 54, 482-499, 1989.

5. Linde, Y., Buzo, A., and Gray, R.M. “An Algorithm for Vector Quantizer Design,” *IEEE Trans. Comm.,* 28, 1, pp. 84-95, 1980.