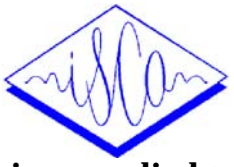

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

An HMM-based phoneme recognizer applied to assessment of dysarthric speech

Xavier Menéndez-Pidal†, Polikoff, J.B., and Bunnell, H.T.

Applied Science & Engineering Laboratories,
duPont Hospital for Children, P.O. Box 269, Wilmington, DE 19899, USA

†SONY Electronics Inc., 3300 Zanker Rd, MS SJ-2D4, San Jose, CA 95134, USA

E-mail: bunnell@.asel.udel.edu or xavier@lsi.sel.sony.com

Abstract: This paper describes work on the development of an HMM-based system for automatic speech assessment, particularly of dysarthric speech. As a first step, we compare recognizer performance on a closed-set, forced choice identification test of dysarthric speech with performance on the same test by untrained listeners. Results indicate that HMM recognition accuracy averaged over all utterances of a dysarthric talker is well-correlated with measures of overall talker intelligibility. However, on an utterance-by-utterance basis, the pattern of errors obtained from the human subjects and the machine, while significantly correlated, accounts for, at best, only about 25 percent of the variance. Potential methods for improving this performance are considered.

10.21437/Eurospeech.1997-319

1. INTRODUCTION

Recent studies have examined the use of automatic speech recognition systems (ASR) by dysarthric speakers with mixed success [1,2]. One frequently cited problem is the amount of variability present in dysarthric speech which is sufficiently large that few items can be unambiguously recognized. However, even if recognition of dysarthric speech is problematic, it is possible that recognition technology could be applied to assessment of dysarthric speech. In an assessment setting, the speech samples intended by the dysarthric talker are known exactly to the assessment system. Consequently, its task is not one of recognition, but of fitting specific models to known (but potentially error-laden) utterances. In this application, metrics which report goodness of fit to selected alternative models might provide useful diagnostic information and/or clinically relevant objective measures to assist speech pathologists in therapy sessions, particularly to assess progress during therapy.

The efficacy of this approach may depend upon how well the output of an ASR device correlates with that of a human listener. Consequently, in the present pilot work a comparison between human and machine is presented. Specifically, we compare the performance of an HMM-based phonemic recognition system with that of human listeners in a forced choice, closed-set, identification task. The HMM system was trained as a speaker independent phonemic recognition system for normal speech. Two different analyses were performed in order to determine if a relationship between human and ASR systems exists in the context of speech evaluation. The first analysis attempts to examine if the speech recognition accuracy obtained by the ASR system and the results from perceptual tests are well-correlated. In the second experiment, we try to identify whether an ASR system provides a similar error pattern as a human perceptual response.

2. DATABASE OVERVIEW

The experiments were performed using the Nemours Database of Dysarthric Speech [3]. In this database, the speech from 10 American talkers with different degrees of Dysarthria was recorded and analyzed. Each talker produced 74 nonsense sentences of the form "the N1 is Ving the N2", where N1 and N2 are monosyllabic nouns and Ving are bisyllabic verbs selected randomly from a target list. Each word in the target list (e.g., boat) was associated with a number of minimally different foils (e.g., vote, moat, goat) so that specific types of phonetic contrasts such as place, manner and voicing contrasts could be analyzed in a forced choice listening test. The words for the first 37 sentences were randomly chosen from the target list, and the second 37 sentences were constructed by swapping the first and second nouns in each of the first 37 sentences.

The recording sessions were conducted in a wheelchair accessible sound-attenuated booth using a table-mounted Electrovoice RE55 dynamic omni-directional microphone connected to a Sony digital audio tape recorder, model PCM-2500 situated outside the recording booth. The talker was seated, typically in a wheelchair, next to the experimenter or speech pathologist, and approximately 12 inches from the microphone. The recording sessions began with a brief examination by a speech pathologist. Following the assessment and after a short break, the experimenter entered the recording booth to lead the talker through a series of recordings which included the set of 74 semantically anomalous sentences described above. For the sentence material, each sentence was read first by the experimenter and then repeated by the talker. This assisted all talkers in pronunciation of words and was essential for some subjects with limited eyesight or literacy. The recorded sentences of both the dysarthric talker and the experimenter were later

digitized and the six words in each sentence were marked using a waveform/spectrogram display and editing program [4].

3. PERPCEPTUAL TESTS

A minimum of five normal hearing listeners were recruited for listening tests with each of the dysarthric speakers. Listeners were seated in a sound dampened room facing a touch screen terminal and heard sentences presented binaurally over TDH-49 headphones at an average level of 72 dB SPL.

At the start of each trial, the terminal screen was cleared and a new sentence frame appeared with the three target word locations in each sentence containing a list of possible response words from which listeners attempted to select the words that they thought the talker was attempting to say. For instance, a sentence might appear as follows:

	FIN		SIPPING		BATH
The	THIN	is	SINNING	the	BADGE
	SIN		SITTING		BATCH
	BIN		SIPPING		BASH
	PIN				BASS
	INN				

Thus, each target word was associated with several similar sounding foils and the listener had to pick the correct alternative from the list (depending on the target word, anywhere from four to six alternatives were available). Subjects selected a response alternative by touching that alternative on the screen of the CRT. To minimize possible list position effects in the response data, the order of response alternatives for the target words was randomly selected each time a sentence was presented.

On average, 5 listeners not familiar with the speaker were assigned to each talker over 12 sessions. In each session, the listener heard the complete set of 74 sentences once in original (as recorded) format, and once in a time-altered format (not relevant to the present study). There were two utterances in original format for each word per session, one from the first 37 sentences and one from the second 37 sentences. Thus, there were 120 identification responses collected per word.

4. ASR DESIGN

In this experiment a Discrete HMM trained in the TIMIT training dataset was built to capture and reflect normal phonetic characteristics of American English. Sixty-one context independent phoneme models were trained using the phonetic labels provided with the TIMIT database. With the DHMM the recognition experiment in the Nemours database was set up similar to a standard isolated word recognition task. To identify the three stimulus words N1, Ving and N2, hand marked word boundary labels were used. Also, the same lexicon (target list) of perplexity 5, was evaluated using the Viterbi decoder algorithm.

4.1. Labeling accuracy in the TIMIT data base

The HMM configuration which provided the best phoneme labeling accuracy in the TIMIT set was obtained when using, for each phoneme model, a number of states proportional to the average phoneme length and the states could be skipped. On average, 7 states were used for each phoneme. With a 30 ms. margin, the system obtained 97% accuracy of phoneme boundary locations over all test files of the TIMIT database. When the margin of error was only 10 msec (i.e., one analysis frame), the accuracy dropped only to 84.5%. These results have outperformed the accuracy obtained by the "Aligner" software developed by Entropic Inc. [5].

4.2. Recognition accuracy in the Nemours data base

At a feature level, the best scores with the Nemours database were obtained using Rasta Mel-cepstrum features. The use of the Rasta bandpass filter compensated for the different recording procedures adopted between the two data sets (TIMIT and Nemours) and improved the recognition accuracy in the Nemours corpus, see Table I. The use of time dynamic features (delta Rasta Melcep) also improved the accuracy. On the other hand, the use of a third VQ (delta2 features) did not provide significant improvement in the Nemours corpus, see Table I.

Table 1: Evaluation of Word Errors in the Nemours database using different front Ends

System	Word Accuracy
Rasta Melcep (1 VQ)	36.3%
Rasta Melcep (2 VQ)	42.0%
Rasta Melcep (3 VQ)	43.4%
Melcep (3 VQ)	37.8%

5. Comparision of Human & Machine Behavior

Table 2 shows the word accuracy obtained during the human perceptual experiments and with the DHMM system for the 10 speakers in the Nemours Corpus. The HMM system provided poorer performance than the human listeners. On the other hand, the linear correlation coefficient obtained between human and machine accuracy was very high (0.94) indicating that the HMM seemed to fail or succeed in a pattern similar to that of the human listeners. This result suggests that standard HMM techniques can be adopted as speech evaluation tools.

Also in Table 2, the linear correlation index obtained for the two error patterns for each speaker is presented. This index was estimated comparing the ranking recognition order obtained by the HMM and by the human listeners for all the words in the data base (74 sentences * 3 stimulus words * 5 target words). While the correlations are not very high, because the correlation coefficients were estimated over a large data sample of 1.100 points

(11.000 points for all the speakers), all represent significant levels of correlation. Nevertheless, the actual system seems not to be a good tool for detection and analysis of specific errors, mainly due to the low overall accuracy of the system. For this purpose, a higher and more accurate technology based on Continuous HMM combined with context dependent phonemes or Neural front-end should be tested.

Table 2: Human v Machine accuracy, and Linear Correlation of the ranking recognition test

SPK	Human	Hmm	Correl.
bb	89.7	52.7	0.48
bv	56.8	31.8	0.32
bk	57.9	33.3	0.33
fb	93.5	57.2	0.53
jf	78.5	44.1	0.42
ll	82.4	41.4	0.44
mh	91.5	56.8	0.50
rl	72.5	41.0	0.41
rk	68.2	44.5	0.40
sc	51.8	30.6	0.32
Ave.	72.3	43.4	0.42

Comparing the different phonetic contrasts, the largest differences were found between the Vowels and Consonants. Consonant recognition was substantially lower for both human and HMM recognizers (see Table 3.).

Table 3: Consonant and Vocalic accuracy

Sound	Human	HMM
Cons.	71%	42%
Vocalic	85%	52%

ACKNOWLEDGEMENTS

Funding for this research has been provided by Grant Number H133E30010, from the National Institute of Disability and Rehabilitation Research, U.S. Dept. of Education, with additional support by the Nemours Foundation, and partially by the Spanish Ministry of Education and Science through a post-doc scholarship grant number FPI-PF94-51.368.925.

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] Carlson, G.S., Bernstein, J., "A voice-input communication AID", SRI-Tech-Rep, project-1252, April-88.

[3] Menéndez-Pidal, X., Polikoff, J.B., Peters, S.M., Leonzio, J.E., Bunnell, H.T. (Oct. 96). "The Nemours Database of Dysarthric Speech", *Proceedings of ICSLP-96*, Philadelphia, USA

[4] Bunnell, H. T., and Mohammed, O. (1992). "EDWave - A PC-based Program for Interactive Graphical Display, Measurement and Editing of Speech and Other Signals." Software presented at the 66th Annual Meeting of the Linguistic Society of America.

[5] Aligner Entropic public documentation. Sept.-96.