RECOGNITION OF SPEECH UNDER STRESS AND IN NOISE

Periagaram K. Rajasekaran
George R. Doddington
and
Joseph W. Picone*
Texas Instruments Inc.
Central Research Laboratories
P.O. Box 226015, MS 238
Dallas, Texas 75266, USA

ABSTRACT

Speech recognizers trained in one condition but operating in a different condition degrade in performance. Typical of this situation is when the recognizer is trained under normal conditions but operated in a stressful and noisy environment as in military applications. This paper reports on recognition experiments conducted with a "simulated stress" data base using a baseline algorithm and its modifications. These algorithms perform acceptably well (1 % substitution rate) for a vocabulary of 105 words under normal conditions, but degrade by an order of magnitude under the "stress" conditions. The experiments also show that the speech production variation caused by noise exposure at the ear is far more deleterious than ambient acoustic noise with a noise cancelling microphone.

1. INTRODUCTION

In military environments such as a fighter cockpit, the speech recognizer has to operate reliably in high ambient noise and with the pilot under stress. However, the required training is usually performed under quiet and non-stressful conditions. Psychological and physiological stress on the pilot manifest themselves as variabilities in the acoustic signal produced. Noise in the cockpit affects the acoustic signal in two ways: as additive noise signal at the microphone, and more importantly by influencing speech production to overcome noise levels in speaker's own ears This latter effect is known as Lombard effect [1].

This effort was funded under contract no. N00039-85-C-0162 monitored by SPAWAR/DoD of USA.

* ATT Bell Laboratories, Naperville, Ill.60566

In this paper we describe some recognition experiments conducted on a "simulated stress" data base collected by Texas Instruments. The data base "simulates" stress by eliciting stress-like changes of the acoustic signal by asking the speaker to vary vocal effort levels, and also by exposing 95 dB pink noise at the subjects' ears to produce Lombard effect. Experiments were also conducted to study the effects of additive noise as opposed to Lombard effect. Section 2 describes the data base.

Section 3 describes briefly the algorithms investigated. Section 4 presents the various experiments conducted and Section 5 presents the conclusions drawn from the experiments.

2. "SIMULATED STRESS" DATA BASE

Psychological and physiological stress on a speaker manifest themselves as variabilities in the acoustic signal produced. Typical of the variabilities are the changes in the spectral slope, fundamental frequency, formant locations, level and duration of the acoustic events of the speech signal[2]. Stress-like degradations of the speech signal were elicited by asking the speaker to produce speech with vocal efforts/effects corresponding to Normal, Fast, Loud, Shout, and Soft conditions as well as with Noise Exposure (95 dB) in the ears. The vocabulary consisted of 105 words including monosyllabic, polysyllabic and confusable words such as "one", "destination", "advisory", "six", "sixty", "fix" etc. Training data consisted of 5 samples of each of the 105 words in a random order under normal conditions, and test data consisted of 2 samples of each word under each stress condition listed above. Data were collected from 5 adult male and 3 adult female speakers, and digitized at a sampling rate of 20 kHz using a 16-bit A/D converter. The data used in our experiments were downsampled to 8 kHz from 20 kHz by means of a

CH2243-4/86/0000-0733 $1.00

downsampling program. Figure 1 shows the wideband spectrograms of the word "zero" (8 kHz) under the six different conditions.

3. RECOGNITION ALGORITHMS

3.1 Baseline Method (PSC)

The baseline algorithm investigated is called the principal spectral components (PSC) method and is described in detail in [3]. Figure 2 shows the generic block diagram of the method.

The LPC parameter vector characterizing a frame of speech (test or reference) is transformed to spectral amplitudes (on a dB scale) normalized to the frame energy using a simulated filter bank. A critical-band filter bank [4] was used in the study. The filter bank amplitudes constitute a vector that may be characterized as normally distributed with mean vector depending on the word(hypothesis), and a covariance matrix. This covariance matrix may be estimated by pooling all available data for the entire vocabulary. Implicit in this process is the assumption that all frames are statistically independent and have the same covariance matrix. A reference template, then, consists of a sequence of hypothesis-dependent mean vectors of filter bank amplitudes, and its statistical variability is described by a single covariance matrix. The recognition problem is to compute, given the input characterization, the likelihoods corresponding to each word hypothesis, and choose that with the largest likelihood. This corresponds to maximum likelihood decision.

In general, the amplitudes of adjacent filters are highly correlated and provide potential for reduction of dimensionality of the feature vector. The filter bank amplitudes are rotated by the eigenvectors of the covariance matrix so that the resulting transformed features are statistically uncorrelated [5]. These features are ranked in decreasing order of statistical variance (eigenvalues), and the least significant features are discarded resulting in a dimensionality reduction. Finally each of these new features is scaled so that its variance is unity. The resulting features are called principal spectral components(PSC), and previous studies have established correlations with perceptual space for certain classes of sounds [6]. A Euclidean distance in this feature space is used as the metric to compare input and reference frames of speech data.

3.2 Enhanced Method (PFV)

The energy-time profile of a speech signal appears to be rich in information for human recognition. It is only reasonable to include the rms energy of a frame of speech signal as an additional feature to the filter bank spectral amplitudes of the the PSC method. The enhanced set can again be orthogonalized statistically as in PSC method, and the higher variance components chosen. The resulting vector is called the Principal Feature Vector (PFV). The Euclidean distance metric and the statistical optimality of maximum likelihood decision is maintained.

3.3 Parameters

The following parameters were used in the algorithms:

LPC analysis: Autocorrelation Method
Hamming Window
LPC order :10
Frame Period: 20 ms / 10 ms
Analysis window: 30 ms
Number of Filters: 14
PFV/PSC dimensionality: 10

4. EXPERIMENTS AND RESULTS

Training was done from training data collected under normal conditions. There were 5 tokens of each word, which were time aligned using dynamic time warping (DTW) algorithm and averaged. Each of the "stress" conditions was tested using the following methods:

PSC 20 ms frame period
PFV 20 ms frame period
PFV 10 ms frame period

In all the recognition tests, a decision was forced. That is, the ability to reject any input without classifying into one of 105 words was disabled. Table 1 shows the substitution rate obtained for each of the condition and the average rms energy level relative to NORMAL condition. Note that adding the rms energy as a feature reduces the error by as much as 30%, and a finer temporal description of 10 ms frame period provides additional 10% reduction. In all cases, the SHOUT condition performance is significantly worse than the other conditions, and the differences in performances for SHOUT due to the recognition algorithms is not statistically significant. Also there is no significant change in the performance under NORMAL condition, which is already excellent for the baseline system. A further breakdown of the error rates by speaker's sex did not reveal any

recognition preferences, and is not shown here.

A second set of recognition experiments were conducted to study the effect of the two noise factors: additive noise and Lombard effect. A subset of data (2 speakers) from the 8 speaker data were corrupted by digitally adding F-16 noise at different levels to produce various signal-to-noise(SNR) ratios. The substitution rates obtained at three different SNR's are compared with the NOISE EXPOSURE condition of 95 dB pink noise in Table 2. It is seen that the changes in speech production attributable to noise degrade far more than the effects of additive noise. The additive noise results appear to be in general agreement with the results in [3]. Table 3 shows the measured SNR for various noise level in dB SPL for the data base in [3]. It is seen that the SNR even at 112 dB SPL is on the order of 20 dB, which is not too unfavorable to recognition algorithms.

5. CONCLUSION

The performance of a baseline recognition scheme (PSC) was determined with a data base showing significant acoustic variabilities. Enhancing the feature vector with energy measurements (PFV) and describing the signal with finer temporal quantization (10 ms frame period) improved the baseline performance. Yet the error rate under stress conditions is roughly ten times worse. worse than under normal conditions. Experience with speaker independent recognition [7] of isolated digits shows that comprehending the spectral dynamics and defining features that are specific to the hypothesized acoustic event can be very beneficial in accommodating acoustic variability of speech signals. Such an approach appears worth investigating for handling stress conditions.

Experiments with additive noise and Lombard effect have shown the relative influence of noise on recognition. In particular, the psychological effect of noise on speech production degraded recognition performance far more than the additive effect of combining noise and speech signals.

6. REFERENCES

[1] Lombard, E., Le Signe de l'Elevation de la Voix, Ann. Maladiers Oreille, Larynx, Nez, Pharynx Vol. 37, 1911, pp. 101-119.

[2] Pisoni, D., R.H. Bernacki, H.C. Nusbaum, and M. Yuchtman, Some Acoustic Phonetic Correlates of Speech Produced in Noise, Proceedings of ICASSP 1985, pp. 1581 - 1584.

[3] Rajasekaran, P.K. and G.R. Doddington, Speech Recognition in the F-16 Cockpit using Principal Spectral Components, Proceedings of ICASSP 1985, pp. 882 - 885.

[4] Zwicker, E. and E. Terhardt, Analytical Expressions for Critical- band Rate and Critical Bandwidths as a Function of Frequency, J. Acoust. Soc. Am. 68(5), pp. 1523-1525, Nov. 1980.

[5] Pols, L.C.W., Real-Time Recognition of Spoken Words, IEEE Trans. comput., Vol. C-20, pp. 972-978, Sept. 1971.

[6] Pols, L.C.W.,L.J.Th.v.d. Kamp, and R. Plomp, Perceptual and Physical Space of Vowel Sounds, J. Acoust. Soc. Am., Vol. 46, pp. 458-467, Aug. 1969.

[7] Bocchieri, E.L. and G.R. Doddington, Frame-Specific Statistical Features for Speaker-Independent Speech Recognition, Trans. ASSP, 1986 (To appear).

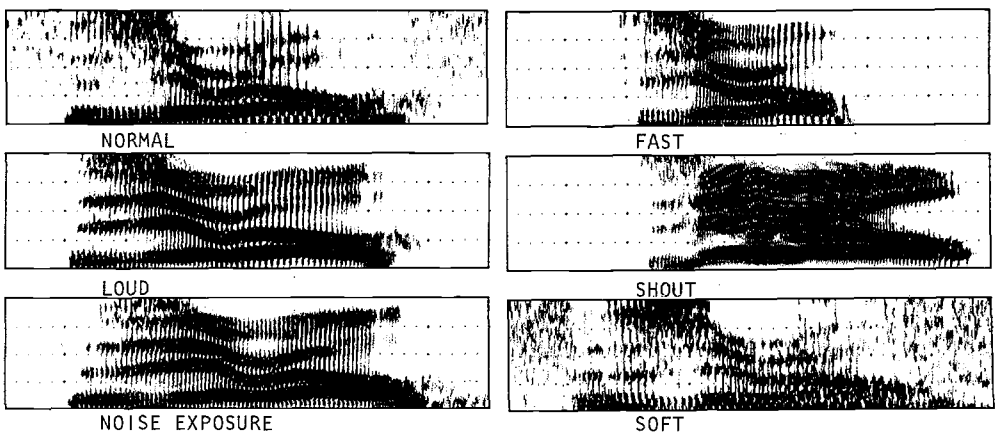

Figure 1: Wideband Spectrograms for the word "Zero" spoken by an adult male under different "stress" conditions. Y-axis scale is 0 - 4 kHz, and X-axis is time.

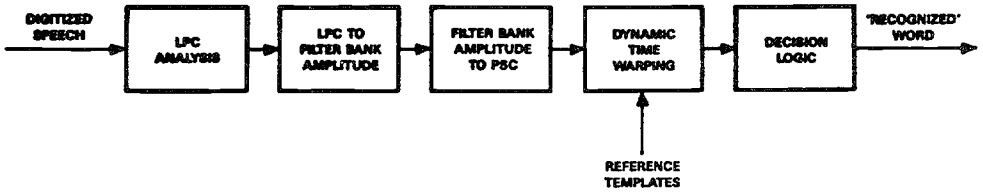

Figure 2: Block Diagram of the Principal Spectral Components Recognition Scheme

TABLE 1
RESULTS FOR THE VARIOUS ALGORITHMS AND THEIR VARIATIONS
DATA BASE: 105 WORDS, 8 SPKRS
SUBSTITUTION RATE (%)

EXPERIMENT	NORM	FAST	LOUD	NOISE	SOFT	SHOUT
PSC METRIC 20 MS FRAME PERIOD	1.1	10.2	24.4	13.8	11.9	78.4
PFV METRIC 20 MS FRAME PERIOD	1.0	7.9	19.3	9.4	4.9	74.5
PFV METRIC 10 MS FRAME PERIOD	0.9	6.0	17.7	7.3	4.3	74.3
AVG. LEVEL RELATIVE TO NORMAL (DB)	0	4	13	9	-14	25

TABLE 2
SUBSTITUTION RATES(%) FOR ADDITIVE NOISE VS. LOMBARD EFFECT

NORM	SNR=30 dB	SNR=20 dB	SNR=10 dB	LOMBARD EFFECT (95 dB Pink Noise)
1.1	1.1	2.8	11.6	13.8

TABLE 3
MEASURED VALUES OF SIGNAL TO NOISE RATIO FOR SEVERE NOISE ENVIRONMENTS
AFTI/F-16 NOISE, M101 MICROPHONE WITH OXYGEN MASK AND REGULATOR

SUBJECT	AMBIENT NOISE LEVEL (dB SPL) ENROLL	TEST		
	85	97	106	112
BK	35	34	26	22
CH	37	33	26	21
DW	38	32	25	20
HH	24	23	17	16
KB	33	29	26	20