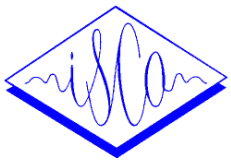

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

FEATURES FOR NOISE-ROBUST SPEAKER-INDEPENDENT WORD RECOGNITION

Brian A. Hanson and Ted H. Applebaum

Speech Technology Laboratory
Division of Panasonic Technologies, Inc.
3888 State Street, Santa Barbara, CA 93105, USA

ABSTRACT

Effects such as additive noise and noise-induced changes in vocal effort (Lombard effect) can cause significant loss of performance for recognizers trained on normal (non-noisy, non-Lombard) speech. In earlier work, improvements to recognition rate over a "standard" speech representation consisting of cepstral coefficients and their first time-derivative (calculated over a 50 msec interval) were demonstrated on the English digits vocabulary by lengthening the interval over which the first derivative is calculated and incorporating a second derivative feature. The current paper extends this work by considering recognition of a much more confusable vocabulary. The recognition results are analyzed for each proposed change in the speech representation, examined by confusable subsets of the vocabulary and contrasted with previous results. Most of the earlier findings for the digits vocabulary were confirmed for the confusable vocabulary. Additionally, it was found that adding a third derivative feature further enhances performance.

1. INTRODUCTION

As automatic speech recognition devices move toward application, emphasis will increasingly be placed on robust performance. Speech recognizers must be tolerant of changes in the environment, channel or talker. Interactions such as the Lombard effect, where an increased noise level causes the talker to modify his or her speech production, must be addressed. Much of the responsibility for robust performance must be borne by the higher levels of processing. However it is essential that the low level speech representation be relatively insensitive to irrelevant sources of variance in the signal.

The goal of our work is to find a speech representation for use in automatic speech recognition which is robust to additive noise and noise-induced changes in the talkers' speech production (Lombard effect). The desired speech representation should not be specialized to a particular noise condition, and small changes to any of the parameters of the representation should not cause serious loss of recognition rate. Recent work has shown that the first few derivatives of the cepstral coefficients are good candidates for such a speech representation. We will discuss the implementation of time-derivative features and review the recent evidence that these features are particularly robust to noise and Lombard effects.

1.1 Time-derivative Feature Implementations

The numerical time-derivatives of cepstral coefficients are calculated over a finite time interval or "window". The derivatives have been implemented via regression (e.g. [1-5]) or simple differences (e.g. [6-9]).

The difference implementation of the time-derivative features involves the least number of speech frames needed to approximate the derivative. The "dynamic" feature (D) is the first difference of two speech frames. The "acceleration" feature (A) is the second difference, involving at most four frames of speech [9].

The regression implementation (or "regression feature") is a smoother representation than the difference implementation as it uses all of the speech frames in the window. We will refer to the regression feature of order "r" and window length "w" as $R_r(w)$ or simply R_r. Regression feature R_0 is the average of the cepstral coefficients in the window. Regression features R_1, R_2 and R_3 are numerical first, second and third time-derivatives of the cepstral coefficients. The "static" feature (S) is the cepstral coefficient vector for a single speech frame and is equivalent to R_0, where w is set equal to the frame step size. See [10] for a description of the calculation of the regression features.

1.2 Previous Work

First derivative features (D and R_1) have been widely used in speech recognition, with window size generally in the range of 40 to 100 msec [11]. Second derivative features (A and R_2) have been less widely used. Furui [3] used static, first and second order regression features to recognize normal (non-noisy, non-Lombard) speech, but found that the use of second order regression feature gave no significant improvement of recognition rate beyond what was achieved by the combination of static and first order regression features. Ney [12] used static, first and second derivative features in the recognition of normal speech and found improvements over simply using static and first derivative features.

In two recent papers [9,10] we have applied time-derivative features to the recognition of normal, noisy, and Lombard speech by a recognizer trained on normal speech. These papers deal with the discrete and regression implementations of time-derivative features respectively. Both papers are based on the same recognition system as the present work, and deal with speaker-independent recognition of the relatively nonconfusable digits vocabulary.

1.2.1. Discrete Implementation

In [9] time-derivative features were applied to recognition of noisy and Lombard speech by a recognizer trained on normal speech. In that study initial experiments on the recognition of Lombard speech were used to select the analysis method, analysis order and cepstral weighting for use in subsequent experiments. The study contrasted two analysis methods: Linear Prediction and Perceptually based Linear Prediction (PLP) [13], at a wide range of analysis orders (5-15). Comparisons were made for unweighted and index weighted (RPS) cepstral coefficients [14]. The combination of 12th order PLP analysis and RPS cepstral weighting was selected on the

basis of best recognition rates for digits. Subsequent experiments compared the recognition performance of various combinations of features and determined the best window lengths for these features.

Figure 1 (based partly on Fig. 4 from [9]) summarizes the recognition experiments performed on Lombard, noisy-normal and noisy-Lombard test speech. "Baseline" denotes the combination of static and dynamic features with a 50 msec window for the dynamic feature. The 50 msec dynamic feature window length is in common use (e.g. in [8]). The highest recognition rates were achieved with window lengths of 130 and 270 msec for the D and A features respectively, as determined by successive one-dimensional searches. The notable results are that the combination of S, D and A features was effective for recognition of Lombard, noisy-normal and noisy-Lombard test speech, that the window lengths giving best overall recognition were longer than expected and that, in the case of noisy-Lombard test speech, recognition rate was improved by omitting the static feature.

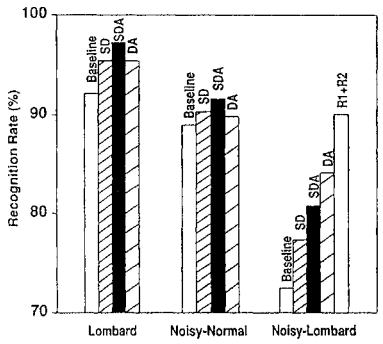

Figure 1 Recognition Rate for Digits Vocabulary. Noisy-Normal (middle graph) and Noisy-Lombard (right graph) test speech data are at 18 dB signal-to-noise ratio. Baseline represents the combination of static and dynamic features with a 50 msec window for the dynamic feature. The regression feature set is R_1(250 msec)+R_2(230 msec).

1.2.2. Regression Implementation

The work in [9] was extended to a regression implementation in [10]. The regression features R_0 through R_3 were studied, both as single features and in combination. As single features, recognition rates for R_1 and R_2 were similar to those achieved by the corresponding discrete features D and A. In combination however the regression feature set R_1+R_2 achieved about 5% higher recognition rate than did the difference-based feature combination D+A, as indicated in Fig. 1.

Window lengths were selected by successive one-dimensional searches. The combination of features R_1(250) and R_2(230), where window length is expressed in milliseconds, gave the best recognition rate for noisy-Lombard test speech (90.1%). For normal test speech this combination of features gave a 99.8% recognition rate which was as good as, or better than, that achieved by any other combination of features. Incorporating the R_3 feature did not significantly improve these recognition rates for any R_3 window length. Incorporating the R_0 feature reduced the recognition rate for noisy-Lombard speech by 3% and the normal speech recognition rate by 0.2%.

As found earlier for the difference implementation, the best window lengths for regression features R_1 and R_2 were quite long (greater than 200 msec). The paper [10] concludes with the conjecture that these long windows will not perform well for a more confusable vocabulary, where finer distinctions must be made.

2. EXPERIMENTS WITH CONFUSABLE VOCABULARY

In the present paper we present recognition results for a confusable vocabulary. As in the previous papers [9,10] the recognition system is trained on normal speech and tested on either normal or noisy-Lombard speech. The recognition system and database are discussed below.

2.1 Recognition System

This work used a discrete density hidden Markov model recognition system which combines multiple features by multiplying output probabilities at the frame level [7]. Note that the relative normalizations of the features have no effect on recognition results of this system. Therefore no ad-hoc weighting was required to combine the features. More details are given in [9] and [10], where the same recognition system was employed.

2.2 Database

The vocabulary consisted of confusable subsets of the English alpha-digits and short function words (Table 1). The speech data consisted of two normal and two Lombard repetitions by 12 male and 12 female talkers. While recording the Lombard part of this database the talkers listened through calibrated headphones to 85 dB SPL white Gaussian noise. White Gaussian noise at 18 dB SNR was later added to the Lombard speech data to produce "noisy-Lombard" data. The data were collected at the same time, and from the same talkers, as the testing digit data used in [9] and [10].

The talkers were divided equally into two gender-balanced groups. When testing with one group of talkers, VQ-codebooks and hidden Markov models were trained with normal speech data from the other group. Each recognition rate reported below is the average of the results obtained from each of the two testing groups.

SET	WORDS
A	a j k
E	b c d e g p t v z three
N	m n
O	go no oh
S	f s x

Table 1 Vocabulary Categorized by Confusable Subset

3. EXPERIMENTAL RESULTS

Initial experiments were performed to select a speech representation which gives relatively good recognition performance for noisy-Lombard test speech without losing recognition performance for normal test speech. Recognition performance was then re-examined for a series of cumulative changes to the speech representation, proceeding from the R_0 feature alone, to the selected representation. The effects of these changes are presented for overall recognition rate and for the recognition rate of each of the confusable subsets of the vocabulary.

3.1 Selection of Speech Representation

The effects of the R_1 and R_2 window lengths are evident in Fig. 2. Recognition rate *improvement* over the static speech representation ("$R_0(10)$") is plotted versus R_2 window length for combinations of features R_0, R_1 and R_2. The R_0 window length is fixed at 10 msec. R_1 window lengths of 50, 130 and 210 msec are shown. Recognition of noisy-Lombard test speech (left side of Fig. 2) is sensitive to both R_1 and R_2 window lengths. Recognition rate improves with increasing R_2 window length up to about 250 msec, and with increasing R_1 window length through 210 msec. Subsequent tests showed a recognition rate decrease at R_1 window length 290 msec. Recognition of normal test speech (right side of Fig. 2) shows much less sensitivity to window length, especially for the R_1 feature. The best recognition results are obtained for R_2 window lengths in the range 130 to 210 msec; however little change in recognition rate occurs even for R_2 window lengths as large as 290 msec.

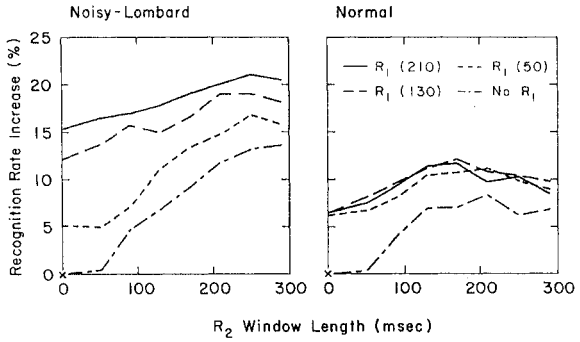

Figure 2 Recognition Rate Increase vs. R_2 Window Length for Confusable Vocabulary. R_0 window fixed at 10 msec. R_1 and R_2 window lengths as indicated. Recognition Rate Increase is with respect to the static feature recognition rate (indicated by "X").

The reduced speech representations R_0+R_1 and R_0+R_2 may also be seen in Fig. 2. Results for the feature set R_0+R_1 are shown as the y-axis intercept of each curve in Fig. 2. Adding R_2 to the commonly used R_0+R_1 feature set improves both the noisy-Lombard and normal speech recognition rates. The feature set R_0+R_2 is indicated by the "No R_1" curves in Fig. 2. Neither of these reduced representations performs as well as the combination of all three features.

Additional results, not shown in the figure, demonstrated the effects of dropping feature R_0 and adding feature R_3. The experiments reported in Fig. 2 were replicated without the R_0 feature. As found in [10], removing R_0 improved recognition rate for noisy-Lombard test speech. However for normal speech the feature set R_1+R_2 was found to be much more sensitive to window length than was $R_0+R_1+R_2$, with some loss of recognition rate in all cases. Further experiments investigated adding feature R_3 to the $R_0+R_1+R_2$ representation. Incorporating the R_3 feature was found to improve recognition rate for noisy-Lombard and normal test speech over a range of R_3 window lengths. The best results were obtained for R_3 window lengths close to 290 msec.

In the following section the speech representation $R_1(210)+R_2(250)+R_3(290)$ will be closely examined. This feature set was adopted as a compromise between the competing objectives of improving recognition rate for the difficult noisy-Lombard condition, and maintaining recognition rate for normal test speech.

3.2 Effects of Individual Features

In this section recognition rate is examined for a series of speech representations which progresses from the static feature alone to the final representation selected at the end of the previous section. Consecutive members of the series differ in one feature only. The six feature sets in the series are listed in Table 2.

#	FEATURE SET	CHANGE
0	$R_0(10)$ "static"	No Change
1	$R_0(10)+R_1(50)$ "baseline"	Add R_1
2	$R_0(10)+R_1(210)$	Lengthen R_1
3	$R_0(10)+R_1(210)+R_2(250)$	Add R_2
4	$R_0(10)+R_1(210)+R_2(250)+R_3(290)$	Add R_3
5	$R_1(210)+R_2(250)+R_3(290)$	Drop R_0

Table 2 Series of Feature Sets

3.2.1 *Relative improvements for average recognition rate*

Figure 3 shows recognition rate taken over the entire vocabulary for the six representations listed in Table 2. Comparing neighboring recognition rates (adjacent "bars") isolates the effect of an individual change in the speech representation. The first change ("Add R_1") results in what we consider to be a "standard" (or "baseline") speech representation. This change improves recognition rate by 5.3% for noisy-Lombard and 6.3% for normal test speech. This increment sets a scale of comparison for the following novel changes.

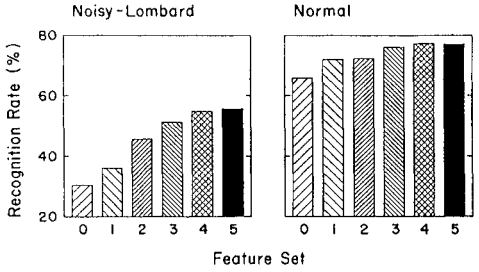

Figure 3 Recognition Rate for the Confusable Vocabulary. Noisy-Lombard (left graph) and Normal (right graph) test speech. Feature sets are defined in Table 2.

The four changes from baseline (R_0+R_1) to final ($R_1+R_2+R_3$) representation resulted in a net gain of recognition rate of 20% for noisy-Lombard and 5% for normal test speech. Lengthening the R_1 window provides the largest increase in recognition rate for noisy-Lombard speech. Although lengthening the R_1 window had little effect on normal speech recognition, it increased the noisy-Lombard speech recognition rate by more than 10%. Lesser gains are found from adding the R_2 and R_3 features. In these cases gains are found for both noisy-Lombard and normal test speech. Removing the R_0 feature from the speech representation resulted in an increase in recognition rate for noisy-Lombard test speech and decrease of recognition rate for normal speech, each of which was less than 1%.

3.2.2 *Relative improvements by confusable set*

To isolate the sources of the improved recognition rate for noisy-Lombard test speech the recognition rates were further categorized by the confusable subsets of the vocabulary (see Table 1). As shown in Fig. 4, for noisy-Lombard test speech lengthening the R_1 window increased recognition rate for every subset of the vocabulary. Adding R_2 and adding R_3 improved recognition rate for every subset except the S subset. Dropping R_0 improved recognition rates for A, E, N but decreased recognition rates for the O and S subsets, resulting in less than 1% net gain in overall recognition rate. In a similar categorization for normal speech (not shown in Fig. 4), it was found that increasing the R_1 window length improved the recognition rate for the E subset by 8% but reduced the recognition rate for each of the remaining subsets, resulting in the almost zero gain in recognition rate shown in Fig. 3. These results are reported in greater detail in [15].

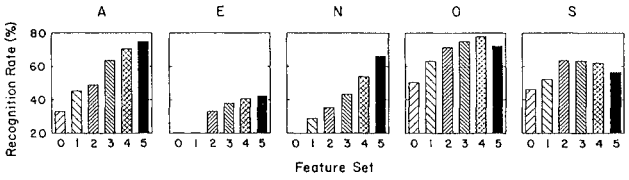

Figure 4 Recognition Rate for Subsets of the Confusable Vocabulary. Noisy-Lombard test speech. Feature sets are defined in Table 2.

4. DISCUSSION

Parametric recognition studies of regression window length for combinations of regression features were reported in [10] for the digits vocabulary. In the present work we adopted a more confusable vocabulary (Table 1) because we suspected that long regression windows, which gave the best recognition rates in [10], would not perform well when fine phonetic discriminations were required. However, most of the findings of [10] have been confirmed for this new vocabulary, despite the short duration of the discriminative portion of many of its contrasts.

The recognizer was trained on normal speech and tested on normal or noisy-Lombard speech. Under these conditions recognition of noisy-Lombard speech is sensitive to the R_1 window length but recognition of normal speech is not (see Fig. 2). Hence a noise-robust speech representation can be obtained by optimizing R_1 window length for noisy-Lombard test speech, where large improvements in recognition rate could be achieved, while tolerating a slightly sub-optimal R_1 window length for recognition of normal speech. The best recognition rates for noisy-Lombard test speech were obtained for R_1 windows over 200 msec in length. Other researchers have used shorter windows for first time-derivative features (D or R_1), as they have presumably optimized their recognition features for normal speech. We believe that long R_1 windows are the best compromise to handle both noisy-Lombard and normal speech input.

Further, the addition of R_2 and R_3 features was found to increase recognition rate for both normal and noisy-Lombard test speech. This has allowed us to select a speech representation which performs much better than the standard representation for both test speech conditions.

The advantages of long regression windows for recognition of noisy-Lombard speech are demonstrated here and in our earlier work. Two aspects of the regression feature window length influence recognition rate: the time length over which the regression is calculated [11], and the amount of smoothing due to averaging multiple frames of speech parameters within the regression window. These factors change together as the regression feature window length is varied. The relative contributions of these two aspects are still under study.

5. SUMMARY

This work examined recognition performance, on a confusable vocabulary, for normal and noisy-Lombard test speech when the recognizer is trained on normal speech data. Cumulative improvement of recognition rate for noisy-Lombard test speech was achieved by successively lengthening the R_1 window to over 200 msec, incorporating the R_2 and R_3 features and removing the R_0 feature from the feature set. Each of these changes, except removing the R_0 feature, also raised recognition rates for normal test speech. These changes to the speech representation raised the recognition rate by 20% for noisy-Lombard test speech and 5% for normal test speech.

REFERENCES

[1] Furui, S. and A.E. Rosenberg, "Experimental Studies in a New Automatic Speaker Verification System Using Telephone Speech," **ICASSP**, pp. 1060-1062, 1980.

[2] Furui, S., "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum," **ASSP**, vol. 34, pp. 52-59, 1986.

[3] Furui, S., "Speaker-Independent Isolated Word Recognition Based on Emphasized Spectral Dynamics," **ICASSP**, pp. 1991-1994, 1986.

[4] Soong, F.K. and A.E. Rosenberg, "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition," **ICASSP**, pp. 877-880, 1986.

[5] Junqua, J.-C., "Evaluation of ASR Front-end in Speaker Dependent and Speaker Independent Recognition," **JASA**, Supp. 1, vol. 81, p. S93, May 1987.

[6] Shikano, K., **Evaluation of LPC Spectral Matching Measures for Phonetic Unit Recognition**, CMU-CS-86-108, Carnegie Mellon Univ., Pittsburgh PA, 1986.

[7] Gupta, V.N., M. Lennig, and P. Mermelstein, "Integration of Acoustic Information in a Large Vocabulary Word Recognizer," **ICASSP**, p. 17.2, 1987.

[8] Lee, K.-F., **The SPHINX System**, CMU-CS-88-148, PhD. Thesis, Carnegie Mellon Univ., Pittsburgh, PA, 1988.

[9] Hanson, B.A. and T.H. Applebaum, "Robust Speaker-Independent Word Recognition using Static, Dynamic and Acceleration Features: Experiments with Lombard and Noisy Speech," **ICASSP**, pp. 857-860, 1990.

[10] Applebaum, T.H. and B.A. Hanson, "Robust Speaker-Independent Word Recognition Using Spectral Smoothing and Temporal Derivatives," in **Signal Processing V** - Proceedings of the Fifth European Signal Processing Conference (EUSIPCO-90), Barcelona, Elsevier, 1990.

[11] Furui, S., "On the Use of Hierarchical Spectral Dynamics in Speech Recognition," **ICASSP**, pp. 789-92, 1990.

[12] Ney, H., "Experiments on Mixture-Density Phoneme-Modelling for the Speaker-Independent 1000-Word Speech Recognition Task," **ICASSP**, pp. 713-716, 1990.

[13] Hermansky, H., B.A. Hanson, and H. Wakita, "Low-dimensional Representation of Vowels Based on All-Pole Modeling in the Psychophysical Domain," **Speech Communication**, vol. 4, pp. 181-187, 1985.

[14] Hanson, B.A. and H. Wakita, "Spectral Slope Distance Measures with Linear Prediction Analysis for Word Recognition in Noise," **ASSP**, vol. 35, pp. 968-973, 1987.

[15] Applebaum, T.H. and B.A. Hanson, "Features for Speaker-Independent Recognition of Noisy and Lombard Speech," presented at 120th Meeting of Acoustical Society of America, San Diego, Fall 1990.