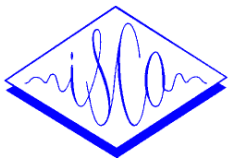

ESCA-NATO Workshop on
Speech under Stress
Lisbon, Portugal
September 14-15, 1995

Recognition Strategies for Lombard Speech

T. H. Applebaum, B. A. Hanson and P. Morin

Speech Technology Laboratory, Panasonic Technologies, Inc.
3888 State St., Santa Barbara, California 93105 USA
Email: {ta, han, phm}@STL.Research.Panasonic.COM

ABSTRACT

Environmental noise is a stress which induces the Lombard style of speech. The effects of this change in speech style is observed on three automatic speech recognition systems. A low complexity word recognizer using reliably found regions of high phoneme similarity was found to perform as well, under normal and Lombard speech styles, as an HMM-based recognizer using cepstral coefficients and their derivatives.

1. INTRODUCTION

The accommodation that talkers make to a noisy environment (Lombard effect) presents an additional challenge to automatic speech recognition systems. This paper characterizes the difficulties of recognizing Lombard speech, and attempts to identify robust speech representations which are less sensitive to the mismatch between normal training and Lombard test conditions.

Derivatives of Perceptually-based LP (PLP) cepstral coefficients calculated over regression windows longer than 200 ms have previously been shown to be relatively robust to Lombard speech variations (e.g. [1–3]). A multiple phoneme similarity speech representation has been shown to be relatively insensitive to the variation between speakers [4]; use of speech representations based on multiple phoneme similarity values for recognition of Lombard speech is considered in this paper.

The impact of the Lombard speaking style on three diverse automatic speech recognition systems is reported. The first recognition system uses a discrete density Hidden Markov Model, based on PLP cepstral coefficients and their derivatives [2]. Two other recognition systems based on multiple phoneme similarity representations are also evaluated: a system using a frame-based dynamic programming matching procedure [5], and a multi-stage high phoneme similarity region based system [6]. Recognition results for clean and Lombard test speech are compared for these three recognition strategies.

2. DATABASE

The lexicon for this study consisted of 21 confusable English alpha-digits, grouped by vowel class as shown in Table 1. The speech data were collected in a sound proof room, and digitized at 10 kHz. Each talker recorded two normal and two Lombard style repetitions of the lexicon. Lombard style speech was collected while the talkers were listening through headphones to 85 dB SPL white Gaussian noise.

Word references were trained with normal speech only. Testing was done under normal and Lombard speech conditions. The forty-eight speakers were divided into two disjoint gender-balanced teams. Data from each team was used once for training and once for testing. For use with the phoneme similarity based recognition systems (MSM, RC-TC) the speech data were downsampled to 8 kHz.

Vowel Class	Vowel Features	Words
IY	Front Long	b c d e g p t v z three
OW	Back Long	oh no go
EY	Front Long	a j k
EH	Front Short	f s x m n

Table 1 Lexicon by vowel class

3. RECOGNITION METHODS

3.1 Discrete Density HMM

The discrete density HMM system [2] uses five states per word, 128 codevectors per feature, and either one or three features. Version **HMM-1** uses only static PLP cepstral coefficients. Version **HMM-3** uses static PLP cepstral coefficients and their first two derivatives: $R_1(210)$ and $R_2(250)$.

3.2 Model Speech Method (MSM)

The Model Speech Method system [5] is an English language adaptation of the system described by Hoshimi [4]. It represents words as six phoneme identifiers and similarity values per centisecond frame. (Phoneme similarities are computed over 10 consecutive LPC cepstra, and thus capture both static and dynamic spectral characteristics). Matching to an unknown word is performed frame-by-frame, by dynamic time-warping.

3.3 High Similarity Region Based Methods

The high similarity region based methods [6], unlike MSM, exploit specific time segments in the phoneme similarities. By keeping only the regions that very consistently indicate the presence of a phoneme, more compact word models are obtained.

The **Region Count (RC)** method represents a word as the mean and inverse variance of the number of high phoneme similarity regions found for each of 55 phoneme units, in the beginning, middle and end of the word.

The **Target Congruence (TC)** method represents a word as a list of phoneme targets, plus two global statistics. Targets represent the reliably found regions of high phoneme similarity, and consist of the phoneme identifier, average peak phoneme similarity value, average left and right frame locations, and target probability (% occurrence in the training data). The training method is adjusted to produce about 50 phoneme targets per second.

The **Multi-stage (RC-TC)** method is based on the combination of scores from the RC and TC stages, as described by Morin [6].

	HMM-3	MSM	RC-TC	
			RC	TC
Speech Rep.	PLP cepsta and derivatives	Multiple phoneme similaries (from 10 consecutive LPC cepstra)		
Time increment	Frame (100/s)	Frame (100/s)	High similarity region (variable rate)	
Match method	Discete density HMM	DTW	Region count vector distance	Congu-ence of aligned regions
Param per 0.5 sec. word	5x128x3 (1920)	6x2x50 (600)	3x2x55 (330)	5x50+2 (252)

Table 2 Recognition methods

4. RECOGNITION RESULTS

Table 3 shows word error rates for each recognition system under normal and Lombard test conditions. Each recognition score results from 1008 trials. The two HMM results show that incorporating the first and second time-derivatives of cepstral coefficients in the speech representation significantly improves recognition rate for Lombard speech conditions [1–3]. Frame-by-frame matching of phoneme similarity values (MSM) did poorly in both conditions, but the high similarity region based methods did well. The multi-stage method (RC-TC) performed as well as the best HMM method in both normal and Lombard test conditions.

5. DISCUSSION

5.1 Lombard Speech Effects

The Lombard speaking style results in numerous speech distortions, such as changes to the overall spectral tilt, formant locations, phoneme durations and pitch. As summarized by Junqua [7], in Lombard speech the frequency of the first formant increases for

Recognition Method	Test Speech Style	
	Normal	Lombard
HMM-1	34	51
HMM-3	24	35
MSM	37	44
RC	27	39
TC	37	44
RC-TC	23	34

Table 3 Word error rates for two test speech styles

vowels, glides, liquids and nasals. Also, vowels are lengthened, while consonants are shortened and more distorted. The increase in mean duration is greater for long vowels than for short vowels [8].

5.2 Within vs. Between Vowel-Class Errors

For each of the recognition methods shown in Table 3, ninety percent or more of all word errors were made within the same vowel class. Tables 4 and 5 show the number of word recognition errors made by the HMM-3 and RC-TC recognition systems, tabulated by vowel class. For normal test speech, the majority of between vowel-class errors come from either IY-EY or EY-IY confusions For Lombard test speech, the majority of between-class errors come from IY-class words recognized as EY-class words. The reverse confusion was not observed. The **IY-EY** confusion (outlined in bold in the tables) is consistent with the vowel lengthening and raising of F1 frequency in Lombard speech, as EY is generally longer and has higher F1 than IY.

Comparing the number of word errors for normal and Lombard test conditions shows that the increased number of errors introduced by Lombard speech style is primarily due to within vowel-class errors for HMM-3 and due to between vowel-class errors for RC-TC. Specifically, the fraction of increase in word error rate due to between class errors is 38% for HMM-3 and 62% for RC-TC. This suggests that HMM-3 does a better job of modeling vowel variations introduced by the Lombard speech style than does RC-TC.

Test Class	Recognized Class							
	Normal				Lombard			
	IY	OW	EY	EH	IY	OW	EY	EH
IY	166				216		37	
OW		10				17		
EY	3		12				32	7
EH				48			2	43

Table 4 Number of word errors by vowel class for HMM-3 recognizer

Test Class	Recognized Class							
	Normal				Lombard			
	IY	OW	EY	EH	IY	OW	EY	EH
IY	158		3		182		53	
OW		15				13		4
EY	10	1	6			2	14	1
EH	3		3	38	7	2	21	50

Table 5 Number of word errors by vowel class for RC-TC recognizer

5.3 Lombard Induced Word Errors

Word confusions which are significantly different between normal and Lombard test conditions (i.e. with $p < 0.05$ by McNemar's test) are shown in Table 6. Several IY-to-EY class word confusions are significant for the high phoneme similarity based methods (see cells outlined in bold). The most common such contrast is the g-to-j confusion. To understand the basis for this confusion, it is useful to look at Target Congruence (TC) word prototypes.

HMM-3	MSM	RC	TC	RC-TC
g→p <0.01	3→a <0.01	n→a <0.01	g→j <0.01	n→a <0.01
g→b <0.01	p→b <0.01	g→j <0.01	c→3 <0.01	3→a <0.01
p→t <0.01	x→s <0.01	a→k 0.02	m→f <0.01	g→j <0.01
a→d 0.01	n→e 0.03	3→a 0.02	a→k <0.01	k→e 0.03
	d→a 0.03	p→b 0.03	p→k 0.01	k→p 0.03
	3→z 0.04	c→3 0.03	t→k 0.03	x→s 0.04
			e→g 0.03	

Table 6 Word confusions which are significantly different between normal and Lombard test conditions. The number of confusions for Lombard speech was higher (white) or lower (grey) than for normal speech. Each cell shows the test word-recognized word combination and significance level of the difference of normal and Lombard test results

Figure 1 shows major discriminating phonemes in TC word prototypes for normal words "g" and "j", as well as for Lombard "g". (The Lombard word prototype represents a composite view of the test data, and was not used in the previous recognition experiments). Phoneme similarity is plotted versus time

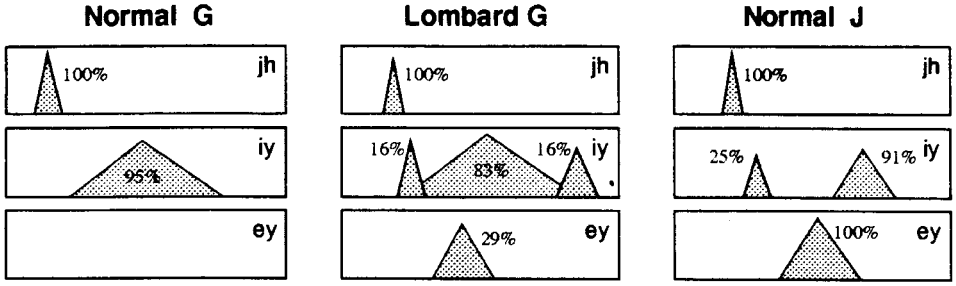

Figure 1 Major discriminating phonemes in Target Congruence (TC) word prototypes for normal and Lombard "g", and normal "j" word utterances. Each panel shows similarity versus time for the indicated phoneme. Triangles depict targets, which represent frequently found high similarity regions. The frequency of occurrence is indicated in percent.

for three of the 55 phoneme symbols. Phoneme targets are shown as triangles, with height equal to the average peak phoneme similarity and bases indicating the average left and right frame locations. Target probabilities are indicated numerically.

The basis for possible confusions of Lombard "g" with "j" is evident. While vowel targets for normal "g" have little resemblance to the vowel targets for normal "j", the Lombard "g" is intermediate between the normal "g" and "j" word prototypes. Lombard "g" has a lengthened IY and has some EY present (with frequency of occurrence 29%). These changes are consistent with the Lombard speech effects noted above.

5.4 Cepstral Normalization

Improved robustness to Lombard effect has been obtained for standard HMM-based recognition techniques in a variety of ways, including use of temporal derivatives calculated over long regression windows and various cepstral normalizations [9]. As each of the speech recognition methods considered in this paper relies on a cepstral representation (at least as an intermediate speech representation), they may also be combined with cepstral normalization techniques. Initial evaluation on the MSM method failed to find significant improvement to normal or Lombard speech recognition for RASTA, Cepstral Mean Removal or Mel-Cepstral scaling. Investigation of cepstral normalization of the high phoneme similarity based recognition methods is yet to be done.

6. CONCLUSIONS

The high phoneme similarity region based recognition method RC-TC has been shown to achieve comparable recognition performance to the higher complexity HMM-3 method in normal and Lombard test conditions. While both techniques have similar overall recognition rates, they make different types of errors.

REFERENCES

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

[2] Applebaum, T. H. and B. A. Hanson, "Robust Speaker-Independent Word Recognition Using Spectral Smoothing and Temporal Derivatives", Proc. EUSIPCO, pp. 1183–1186, 1990.

[3] Applebaum, T. H. and B. A. Hanson, "Features for Speaker-Independent Recognition of Noisy and Lombard Speech", J. Amer. Voice I/O Soc., vol. 14, pp. 73–80, 1993.

[4] Hoshimi, M., M. Yamada, and K. Niyada, "Speaker Independent Speech Recognition Method Using Phoneme Similarity Vector," Proc. ICSLP, vol. 3, pp. 1915-1918, 1994.

[5] Ohno, Y., M. Hoshimi, S. Hiraoka, K. Niyada, and T. H. Applebaum, "A Study of English Model Speech Method," Proc. Acoustical Society of Japan, Spring 1995 (in Japanese).

[6] Morin, P. and T. H. Applebaum, "Word Hypothesizer Based on Reliably Detected Phoneme Similarity Regions", to appear in Proc. Eurospeech, 1995.

[7] Junqua, J.-C., "The Lombard Reflex and Its Role on Human Listeners and Automatic Speech Recognizers," JASA, pp. 510- 524, 1993.

[8] Junqua, J.-C., "A Duration Study of Speech Vowels Produced in Noise " Proc. ICSLP, pp. 419–422, 1994.

[9] Hanson, B. A. and T. H. Applebaum, "Subband or Cepstral Domain Filtering for Recognition of Lombard and Channel-Distorted Speech", Proc. ICASSP, Vol. 2, pp. 79–82, 1993.