LOMBARD EFFECT COMPENSATION AND NOISE SUPPRESSION FOR NOISY LOMBARD SPEECH RECOGNITION

Sang-Mun Chi, Yung-Hwan Oh

Department of Computer Science

Korea Advanced Institute of Science and Technology

373-1 Kusong-dong Yusong-gu Taejon 305-701 Korea

email: smchi@bulsai.kaist.ac.kr

ABSTRACT

The performance of speech recognition system degrades rapidly in the presence of ambient noise. To reduce the degradation, a degradation model is proposed which represents the spectral changes of speech signal uttered in noisy environments. The model uses frequency warping and amplitude scaling of each frequency band to simulate the variations of formant location, formant bandwidth, pitch, spectral tilt, and energy in each frequency band by Lombard effect. Another Lombard effect, the variation of overall vocal intensity is represented by a multiplicative constant term depending on spectral magnitude of input speech. The noise contamination is represented by an additive term in the frequency domain. According to this degradation model, the cepstral vector of clean speech is estimated from that of noisy-Lombard speech using spectral subtraction, spectral magnitude normalization, band-pass filter in LIN-LOG spectral domain, and multiple linear transformation. Noisy-Lombard speech data is collected by simulating the noisy environments using noises from automobile cabins, an exhibition hall, telephone booths in downtown, crowded streets, and computer rooms. The proposed method significantly reduces error rates in the recognition of 50 Korean word. For example, the recognition rate is 95.91% with this method, and 79.68% without this method at SNR (Signal-to-Noise Ratio) 10 dB.

1. INTRODUCTION

Although speech recognition in artificially constrained conditions has recently reached high levels of performance, problems still remain in the deployment of speech recognition technology in the real world. One of the problems is the performance degradation of speech recognizers when they are used in noisy environments such as offices, automobile cabins, streets, and computer rooms. The reason for this performance degradation is not only a contamination of speech signals by ambient noise, but also articulation variability as the speaker attempts to communicate more effectively in noisy environments, which is called the Lombard effect [6].

A number of approaches has been considered for robust speech recognition. Noise resistant features and distance measures, such as SMC (short-time modified coherence), RASTA (RelAtive SpecTrAl) processing, projection measures are used for suppress additive noise [5,7,8]. Speech enhancement from noisy speech using spectral subtraction [1], multiple linear regression transformation or artificial neural network and model parameter adaptation to noisy environment are also used for robust speech recognition.

Since the Lombard effect is a nonlinear distortion depending on the speaker, noise level, and noise type, it is not easy to analyze. The Lombard effect is modeled as an additive or multiplicative term in the cepstral domain, and it is estimated and canceled [2,3,9]. The multi-style training method uses Lombard speech for training data [9]. The dynamic feature is known to be robust to Lombard speech recognition [4], and the codebook adaptation and acoustic phonetic variability models for HMM adaptation are used for Lombard effect compensation [10,11].

In this paper, the Lombard effect and noise contamination are represented by several explicit distortions in the frequency domain, and these distortions are canceled in the feature extraction stage according to the speech degradation model in noisy environments. The cepstral vector of clean speech is estimated from that of noisy-Lombard speech using spectral subtraction [1], spectral magnitude normalization, band-pass filter in LIN-LOG spectral domain [5], and multiple linear transformation. To evaluate the proposed method, noisy Lombard speech is generated by having speakers listen to real world noises through headphones. Word recognition experiments are conducted with this noisy-Lombard speech.

This paper is organized as follows: Section 2 describes the database used for this experiment and section 3 describes feature processing for Lombard effect compensation and noise suppression. Section 4 describes the evaluation experiments, and section 5 concludes the paper.

2. SPEECH AND NOISE MATERIAL

2.1. Noise material

To develop effective robust speech recognition method, noisy speech uttered in the real world is required and the speech database should contain every possible distortions which could occur in noisy environments. But it is not feasible to collect speech data in various noisy environments. In this study, 22 noises obtained from automobile cabins, an exhibition hall, telephone

booths in downtown, crowded streets, and computer rooms with various SPL (Sound Pressure Level) are used for experiments. As can be seen from Figure 1, The SPL of the noises varies between 60dB and 90dB. Noise-free Lombard speech is produced by the speakers listening to these noises through headphones, and noisy-Lombard speech is produced by adding these noises to noise-free Lombard speech with various SNR.

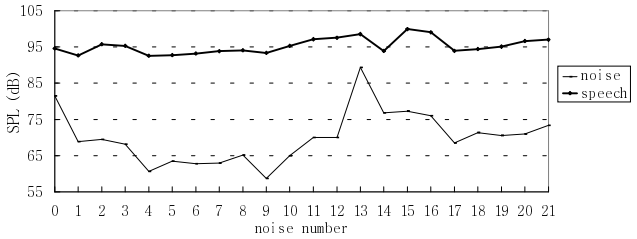

Figure 1: The average SPL of noises from headphones and speech uttered by 20 speakers in each noisy environments.

2.2. Speech material

50 Korean words under 22 simulated noisy environments are produced by 10 male and 10 female speakers. The average SPL of these Lombard speech in each noisy environment is plotted in Figure 1. It is observed that the SPL of the speech is proportional to the SPL of the noises by Lombard effect.

Although speakers are trying to change their vocal intensity according to the SPL of ambient noise, the SPL of speakers vary considerably. This perturbation depends on speaker, noise level, and noise type. As shown in Figure 2, the SNRs of 20 speakers in each type of noisy environment vary greatly. This is also a source of performance degradation.

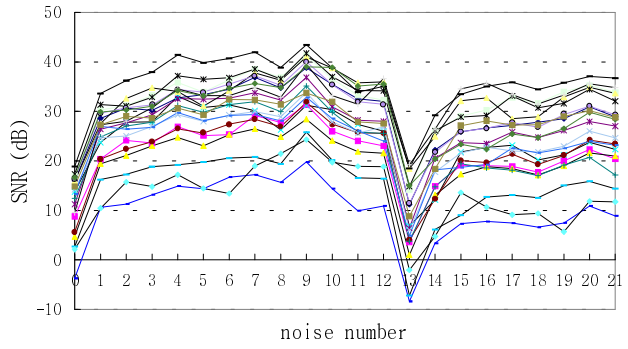

Figure 2: Average SNR of 20 speakers in 22 noisy environment.

3. LOMBARD EFFECT COMPENSATION AND NOISE SUPPRESSION

3.1. Degradation model

This section presents the model in frequency domain of Lombard effect and noise contamination which degrade speech recognition performance in noisy environments.

In noisy environment, speech is distorted by the Lombard effect and then noise is added to speech. The Lombard effect is modeled in two steps. First, the variations of formant location, formant bandwidth, pitch, spectral tilt, and energy in each frequency band is represented by nonlinear frequency warping $F(\cdot)$ and amplitude scaling of each frequency band $A(\cdot)$. The spectrum of clean speech $S(\omega)$ is distorted by $F(\cdot)$ and $A(\cdot)$, and becomes $Y_1(\omega)$.

$$Y_1(\omega) = A(\omega)S(F(\omega)) \tag{1}$$

Secondly, speakers increase their vocal intensity to communicate effectively according to the level of ambient noises as shown in Figure 1. The variation of overall intensity depends on noise level, speaker, and phoneme. This is also a degradation factor and is modeled by intensity variation factor G.

$$Y_2(\omega) = G \cdot Y_1(\omega) = G \cdot A(\omega)S(F(\omega)) \tag{2}$$

Finally, noise contamination can be represented as an additive term in the frequency domain. Let $N(\omega)$ be the spectrum of noise signal. Then the spectrum of Lombard speech $Y_2(\omega)$ becomes the spectrum of noisy-Lombard speech $Y_3(\omega)$ by noise contamination.

$$Y_3(\omega) = Y_2(\omega) + N(\omega) = G \cdot A(\omega)S(F(\omega)) + N(\omega) \tag{3}$$

3.2. Restoration of clean speech from noisy-Lombard speech

Distortions of noisy-Lombard speech can be eliminated by the inverse processing of degradation sequence.

Since the noise characteristics are assumed to change slowly relative to those of speech signal. the spectrum of noise $N(\omega)$ can be estimated in non-speech intervals. The spectrum of Lombard speech $Y_2(\omega)$ is obtained from spectrum of noisy-Lombard speech $Y_3(\omega)$ by subtracting $N(\omega)$ using Spectral subtraction method [1].

To cancel intensity variation factor G in $Y_2(\omega)$, G is defined by equation (4)

$$G = \frac{\text{average spectral magnitude of speech in input signal}}{\text{reference spectral magnitude}} \tag{4}$$

where reference spectral magnitude is defined by a fixed value which is the average spectral magnitude of all words. Intensity variation factor G depends on spectral magnitude of the input word, and thus depends on the speaker, the type and level of noise. $Y_1(\omega)$ is estimated by dividing $Y_2(\omega)$ by G, Every input signal has the same average spectral magnitude of speech interval after divided by G. This normalization not only cancel intensity variability, but also stabilize input signal to follow the LIN-LOG RASTA filtering, which is effective in both convolution and additive noise [5]. The optimal J value which is used in LIN-LOG RASTA processing is signal dependent and this dependency is another source of variability in speech recognition. Spectral mapping or multiple J values are used for compensating this variability. Those compensation methods are not used in the

proposed method since the input signals are processed by spectral magnitude normalization.

For LIN-LOG RASTA filtering, $Y_1(\omega)$ is transformed to spectral domain $Y(\omega) = \ln(1 + J \cdot Y_1(\omega))$ which is linear-like for small spectral values and logarithmic-like for large spectral values. To suppress additive and convolutional noise, $Y(\omega)$ is then filtered by band-pass filter $0.1 \cdot (2 + z^{-1} - z^{-3} - 2z^{-4}) / (1 - 0.94z^{-1})$. Finally, the filtered spectral value is transformed back by approximate inverse transform $Y_1(\omega) = e^{Y(\omega)} / J$.

The distortion factors, nonlinear frequency warping $F(\cdot)$ and amplitude scaling of each frequency band $A(\cdot)$ can be canceled in the cepstral domain. Let the cepstral coefficient of clean speech be C_n^{clean} and the corresponding spectrum of clean speech be $S(\omega)$. Let the cepstral coefficient of speech which is distorted by $F(\cdot)$ and $A(\cdot)$ be $C_k^{Lombard}$ and the corresponding spectrum be $Y_1(\omega) = A(\omega)S(F(\omega))$. Then we obtain Equation (5), (6). Substituting Equation (6) into (5), we get Equation (7). Thus by the multiple linear transformation, the cepstrum of clean speech is estimated from that of the speech distorted by $F(\cdot)$ and $A(\cdot)$.

$$C_n^{clean} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log(|S(\omega)|) e^{j\omega n} d\omega \quad (5)$$

$$\log(|A(\omega)S(F(\omega))|) = \sum_{k=-\infty}^{k=\infty} C_k^{Lombard} e^{-j\omega k} \quad (6)$$

$$\begin{aligned} C_n^{clean} &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \log(|S(F(F^{-1}(\omega)))|) e^{j\omega n} d\omega \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \left[\sum_{k=-\infty}^{k=\infty} C_k^{Lombard} e^{-jF^{-1}(\omega)k} - \log(|A(F^{-1}(\omega))|) \right] e^{j\omega n} d\omega \\ &= \sum_{k=-\infty}^{k=\infty} A(n,k) \cdot C_k^{Lombard} + B(n) \end{aligned} \quad (7)$$

The acoustic phonetic variability by Lombard effect depends on the speaker, noise level, noise type, and phoneme. This dependency is approximated in the acoustic feature domain by dividing cepstral space using vector quantization. To get transform matrix A and vector B, first, the cepstral vector of Lombard speech and that of clean speech are Viterbi aligned with reference HMM (Hidden Markov Model) respectively. Secondly, the pairs are made with the cepstral vector of Lombard speech and clean speech which are aligned in the same state of HMM. Thirdly, the cepstral vectors of Lombard speech are collected for each codeword. Finally, using the collected cepstral vector of Lombard speech and corresponding cepstral vector of clean speech, A and B are estimated by linear multiple regression method for each codeword.

The cepstral vector of clean speech is estimated by transforming the cepstral vector of Lombard speech using matrix A and vector B. To avoid the transformation being too dependent on training data, the cepstral vector of clean speech is obtained by averaging the transformed cepstral vector and the original cepstral vector.

4. EXPERIMENTAL EVALUATION

4.1. Experimental conditions

In order to validate the proposed method, speech recognition experiments were conducted using noisy-Lombard speech produced in 22 simulated noisy environments. The experimental conditions were as follows. The speech data was sampled at 16kHz, 16 bits, and pre-emphasized by a filter $1 - 0.95z^{-1}$. Using Hamming window with 32msec length, 14-order cepstral vectors were extracted every 16msec. Cepstral coefficients were computed from 19 Bark-scale filter bank using DCT (Discrete Cosine Transformation). 14th order cepstral vector, difference cepstral vector, normalized energy, first and second order difference energy were used for the feature vector. The size of 3 separate codebooks were 256, 256, and 32 respectively. 15 state discrete density HMMs were used as a recognizer. 2 repetition of 50 words uttered by 5 males and 5 females under clean environments were used for training data. Noisy-Lombard words uttered by another 5 males and 5 females under 22 noisy environments were used for evaluation.

The several features for comparison were as follows. (1) PROP: feature extracted by proposed method, (2) BARK-CEP: cepstral vector from log energy output of Bark-scale filter bank, DFT (Discrete Fourier Transformation) used for making filter bank, (3) SPEC-SUB: BARK-CEP with spectral subtraction, (4) LIN-LOG RASTA: LIN-LOG RASTA proposed in [5], (5) LPC-CEP: mel-scale cepstrum from LPC (Linear Predictive Coding) coefficients, (6) PROJ: projection measure to LPC-CEP, (7) SMC: mel-scale cepstrum from SMC LPC coefficient.

4.2. Evaluation by word recognition

The noisy-Lombard speech used for preliminary experiments was generated to have SNR 10 dB. Speech data recorded in street and computer room noises (noise number 7 to 12, 20, 21) were used.

As shown in Table 1, the proposed method PROP could effectively compensate Lombard effect and suppress noise. The recognition rates of LPC based methods such as LPC-CEP, PROJ, and SMC were worse than those of DFT based methods such as PROP, BARK-CEP, SPEC-SUB, and LIN-LOG RASTA. This is because LPC is more sensitive to noise than DFT. In LIN-LOG RASTA, 10^{-7} was used for J value. When 10^{-6} was used, the recognition rate was 87.5% and significantly degraded when 10^{-5} and 10^{-8} were used. The cepstral transformation matrixes and vectors are trained using the pair, noisy-Lombard speech under street and computer room noises and clean speech.

Table 1: Recognition rates of preliminary experiments (%). The SNR of speech data is 10dB.

Feature	PROP	BARK-CEP	SPEC-SUB	LIN-LOG RASTA	LPC-CEP	PROJ	SMC
Rec. rate	95.13	72.75	78.50	87.93	60.68	69.58	71.35

Table 2 indicates the rates of speech recognition experiments using speech data uttered in 22 noisy environments with various

SNR. SMC and LIN-LOG RASTA were excluded in these experiments because SMC has large computation load and LIN-LOG RASTA is difficult to choose J values.

SNR CLEAN means the speech data used for experiments were noise-free Lombard speech. SNR REAL means the data were contaminated by noises with the ratio of SPL of speech to SPL of noises from headphones. SNR 20dB, 10dB, 0dB mean that noises were added to speech with SNR 20dB, 10dB, 0dB respectively. In features PROP-LOM was also a proposed feature extraction but the training data was Noisy-Lombard speech uttered in noise environments 7 to 12 and 20, 21 with SNR real. Since this method contained the Lombard effect in training data, the cepstral transform was not used.

Table 2: Recognition rate of several feature extraction methods at various SNR

Feature / SNR	PROP	PROP-LOM	BARK-CEP	SPEC-SUB	LPC-CEP	PROJ
CLEAN	98.60	99.17	96.15	95.58	92.75	92.12
REAL	97.63	99.02	91.16	91.17	86.03	87.30
20dB	98.15	99.01	92.45	91.65	85.92	87.45
10dB	95.91	97.71	79.68	82.26	72.91	75.98
0dB	79.75	83.80	43.74	55.03	42.56	46.42

SNR CLEAN had the only distortion from the Lombard effect. As shown in Table 2, since proposed methods PROP and PROP-LOM can compensate the Lombard effect, they improved recognition rates compared with baseline feature extraction BARK-CEP. But the other noise suppression methods SPEC-SUB and PROJ degraded them when they were used in noise-free speech. The experimental results in SNR REAL, 20dB, 10dB, and 0dB showed that the proposed methods were effective in noise suppression and Lombard effect compensation. Since PROP-LOM used Noisy-Lombard speech, SNR REAL, the recognizer was trained on the data distorted by the Lombard effect and noise contamination. It showed the best recognition rates.

5. CONCLUSIONS

This paper described the use of Lombard effect compensation and noise suppression to reduce recognition performance degradation under noise conditions. The degradation model representing spectral changes of speech signal under noisy environments was proposed. Non-linear warping function and amplitude scaling function in the spectral domain represented variations of formant location, formant bandwidth, pitch, spectral tilt, and energy in each frequency band. These variations were approximated by linear transformation in the cepstral domain. The cepstrum of clean speech was estimated from that of Lombard speech by multiple linear transformation. Spectral magnitude was normalized to cancel the variation of vocal intensity. Spectral subtraction was used to suppress noise and the spectrum was filtered by band-pass filter to cancel slow varying noise.

Experimental evaluations were executed in speaker-independent isolated word recognition based on discrete density HMMs. Noisy-Lombard speech of 50 Korean words were spoken by 10 male and 10 female listening to real world noises through headphones. Recognition experiments were conducted with contamination by noise from automobile cabins, an exhibition hall, telephone booths in downtown, crowded streets, and computer rooms. From the experiments, the effectiveness of the proposed method has been confirmed.

REFERENCES

1. Boll, S.F. “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. ASSP* 27: 113–120, 1979.
2. Chen, Y. “Cepstral domain stress compensation for robust speech recognition,” *ICASSP*, 717–720, 1987.
3. Hansen, J.H.L. and Cairns, D.A. “ICARUS: source generator based real-time recognition of speech in noisy stressful and Lombard effect environments,” *Speech Communication* 16: 391–422, 1995.
4. Hanson, B.A. and Applebaum, T. H. “Robust speaker independent word recognition using static, dynamic and acceleration feature: experiments with Lombard and noisy speech,” *ICASSP*, 857–890, 1990.
5. Hermansky, H., Morgan, N., and Hirsh, H.G. “Recognition of speech in additive and convolutional noise based on RASTA spectral processing,” *ICASSP*, 83–87, 1993.
6. Junqua, J.C. “The Lombard reflex and its role on human listeners and automatic speech recognizer,” *J. Acoustic. Soc. Amer., Vol. 93, 1993, 510-524.*
7. Mansour, D. and Juang, B.H. “The short-time modified coherence representation and noisy speech recognition,” *IEEE Trans. ASSP* 37: 795–804, 1989.
8. Mansour, D. and Juang, B.H. “A family of distortion measure based upon projection operation for robust speech recognition,” *IEEE Trans. ASSP* 37: 1695–1671, 1989.
9. Paul, D.B., “A speaker-stress resistant hmm isolated word recognition,” *ICASSP*, 713–716, 1987.
10. Roe, D.B., “Speech recognition with a noise-adapting codebook,” *ICASSP*, 1139–1142, 1987
11. Suzuki, T. and Nakajima, K., and Abe, Y., “Isolated word recognition using models for acoustic phonetic variability by Lombard effect,” *ICSLP*, 999–1002, 1994.