Robust Formant Features for Speaker Verification in the Lombard Effect

Ileun Kwak* and Hong-Goo Kang†
Department of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea
* E-mail: ileun@dsp.yonsei.ac.kr Tel/Fax: +82-10-30039146
† E-mail: hgkang@yonsei.ac.kr Tel/Fax: +82-02-21232766

***Abstract*—This paper presents a voice controlled speaker verification system for hand-held devices in noisy environments. In noisy environments, users unintentionally increase their voice intensity because of the ear-mouth feedback mechanism i.e., the Lombard effect; thus, the characteristic of the input signal is much different from that in a quiet environment. To enhance the accuracy of a speaker verification system, this paper proposes a robust formant feature that represents the physical nature of the voice production system of the speaker. It utilizes the analysis results that the impact of Lombard varies depending on the types and levels of the background noise, but the third and fourth formant frequencies are insensitive even in the Lombard condition. Experimental results show that optimal performance can be achieved when the above formant frequencies and corresponding bandwidths are combined.**

I. Introduction

As the size of wearable smart devices becomes smaller, it is becoming increasingly important to design an efficient mechanism to control or interface with the device. Since most wearable devices do not have enough space to install multiple sensors, a microphone based voice interface system, such as automatic speech or speaker recognition, is a good candidate for fulfilling the purpose. Recently, the accuracy of voice recognition systems has increased significantly, but there are still robustness issues in noisy environments. Because the operating environment of wearable devices cannot be limited to a small number of restricted conditions, it is not easy to build a statistical training model that can be generalized to various conditions.

On the other hand, the configuration of speech and speaker recognition in watch-type devices is somewhat different from conventional recognition systems because of the user's active involvement in the production of voice commands. In a noisy environment, most users closely position the device near their mouths to clearly speak commands [1]. However, this causes the characteristics of the input speech signal to change due to the Lombard effect. Although the loudness of input speech in Lombard condition becomes louder but its speaking style also changes, thus the accuracy of automatic recognition system decreases significantly. Rajasekaran *et al*. claimed that performance degradation caused by the Lombard effect was much more severe than that caused by background noise [2].

The impact of the Lombard effect on speaker recognition and its compensation methods have been studied for several years. Karlsson *et al*. introduced a structured training set which consists of utterances with different speaking styles, but the performance of the system degraded to the utterances that were not included in the speaking style of training set [3]. Another approach was to set up two models for neutral and Lombard conditions, and applying a different model to each condition [4]. However, this is unrealistic because speech characteristics also change depending on the types of noise at different levels and individual speakers [5] [6]. Furthermore, the signal characteristic also varies by the type of words used (*e.g.* verb, object, etc.) [7]. For these reasons, it is not easy to design an adaptation model that is suitable for various effects. Therefore, finding robust features that are applicable for both neutral and Lombard conditions can be a solution for speaker recognition under Lombard situations.

In a view of the source-filter model based speech production system, both source and filter features represent the physical characteristics of the speaker. Several previous studies on the analysis of source characteristics of the Lombard condition showed that the variation of source characteristics was much higher than that of filter ones [8]–[10]. Consequently, they concluded that the features representing filter characteristics or vocal tract were useful for speaker recognition in Lombard conditions. There were several studies on analyzing the effects of Lombard speech on formant frequencies [11] [12]. Bond *et al*. showed that the first formant frequency increased, but the second formant frequency decreased in 95 dB pink noise [13]. However, there has been no analysis on the variation of higher formant frequencies and other types of features such as the variation of formant bandwidth.

This paper proposes filter related features that have more robust characteristics than source related features in various Lombard conditions. The features utilize the experimental results that the statistical variation of formant features in the high frequency band is small in Lombard conditions, and the features also represent characteristics of the speaker's vocal tract as well. From the analysis of the Lombard speech database, we selected features that are frequently used in speaker verification tasks, then grouped them into the category of source or filter related ones. Analysis results confirm that source characteristics vary significantly in Lombard conditions. However, some features representing filter characteristics such as second to fourth formant frequency, show con-

sistent characteristics in all six Lombard conditions. Speaker verification tasks are performed with formant features that use long-term formant (LTF) distribution features [14] [15], which confirms the superiority of the proposed features in Lombard conditions.

II. Lombard Speech Database and Analysis

A. Database

To perform speaker verification experiments in the Lombard condition, we recorded speech data by following the setup depicted in Fig. 1 and Table I. In fact, the system setup is similar to the procedure described in the recording process of UT-SCOPE database [16]. Ten subjects (6 male and 4 female) were participated in the recording process. Each subject recorded twenty instruction sentences related to wearable devices, ten phrases consisted of two words that are generally used for web search or control commands, and three long news scripts. Each speaker produced a speech while listening noise signals [17] through a headphone. The sampling frequency of the recorded signal was 44.1 kHz, then they were downsampled to 16 kHz.

B. Database analysis

Several features that were frequently used for speaker verification tasks were used to analyze the characteristics of a Lombard speech database. Fundamental frequency (F0) and formant features were extracted by the Wavesurfer software [18] for this analysis. The extracted features were categorized into source or filter related ones, then the characteristic variations of features were analyzed in various types (Babble, Pink) and levels (65, 70 and 75 dB-SPL) of noise. The source related features were F0, energy, and vowel duration. The filter related features were formant frequency and bandwidth. Only the features obtained in the region of vowel duration were used for analysis. The analysis results are depicted in Fig. 2–4, and a detailed explanation follows in the next subsections.

1) Variation of source related features: It is well-known that the source related features are good for speaker verification systems [19] [20]. However, under the Lombard condition, the characteristics of the source related features vary inconsistently depending on the types and levels of noises as is shown in Fig. 2. It shows that the percentage variation of source related features is much higher in Lombard than

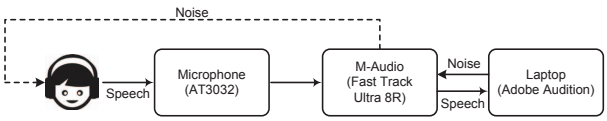

Fig. 1. Recording Scenario

TABLE I
Noise type and level for Lombard condition

	Noise type	Level(dB-SPL)	Subjects
Neutral	None	None	4 female, 6 male
Lombard	Babble Pink	65/70/75	

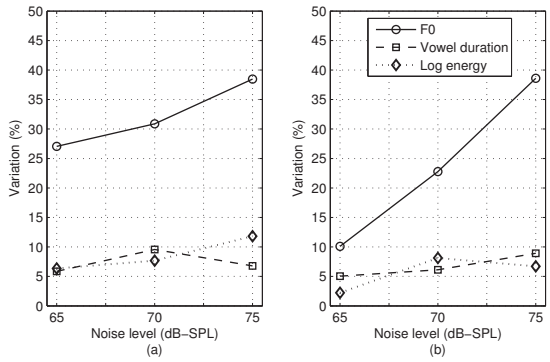

Fig. 2. Average variation of source characteristics in different noise levels and types of Lombard condition. (a) babble noise (b) pink noise

in neutral conditions. It also shows that the variation of the F0 is the largest among three source related features. The percentage variation monotonically increases as noise level increases, but the degree of variation differs with the type of noise. The variation of log-energy and vowel duration is not dramatic compared to that of F0, but the variation is somewhat inconsistent or irregular to the noise types and levels.

2) Variation of filter related features: To analyze the variation of filter related features in Lombard conditions, we chose the first to fourth formant frequencies and their corresponding bandwidths. Fig. 3 shows the average variation of formant frequencies in six Lombard conditions. The first formant (F1) frequency increases as the noise level increases. The second formant (F2) frequency becomes lower in Lombard conditions, but there is no variation in the third and fourth formant (F3 and F4) frequencies even if the noise level increases. Note that the results are well matched with the one reported in the previous study such that F1 and F2 frequencies change under Lombard conditions [13]. However, the frequencies of F3 and F4 do not change significantly, therefore it is expected that they may be regarded as good features in Lombard conditions. Fig. 4 depicts the average variation of formant bandwidth, which shows that all of the formant bandwidths are decreased in Lombard conditions. From the results of formant frequency and bandwidth in Lombard conditions, it clearly shows that the formant frequency of high frequency band is more consistent than others.

To find suitable features for the speaker verification system in Lombard conditions, Kruskal-Wallis analysis was performed to the formant frequency features. Twenty sentences of neutral speech in the recorded database were used for the analysis. Table II shows the Kruskal-Wallis analysis results of the formant frequency distribution in ten subjects. Higher value of chi-square denotes that the feature distribution is much different from each other. F3 and F4 frequencies have higher chi-square values than others, which also confirms that F3 and F4 frequencies are key features to represent the characteristics of the speakers. Table III depicts the variance of the formant frequency distribution of four female subjects in Lombard conditions. The analysis results show that F1 frequency has

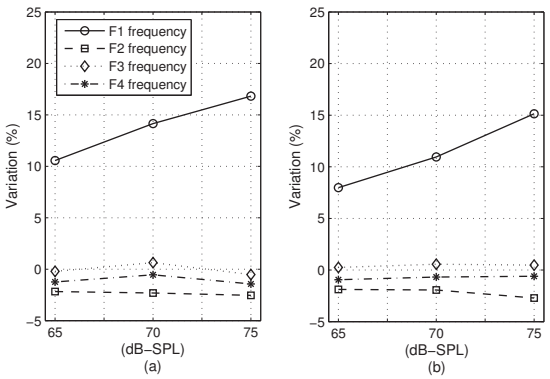

Fig. 3. Average variation of formant frequency in different noise levels and types of Lombard condition. (a) babble noise (b) pink noise

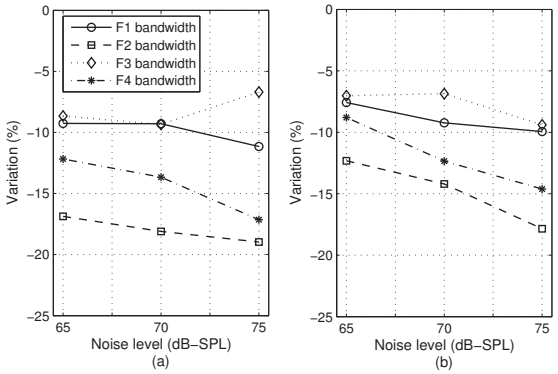

Fig. 4. Average variation of formant bandwidth in different noise levels and types of Lombard condition. (a) babble noise (b) pink noise

higher value of chi-square than F2, F3, and F4 frequencies. In terms of F2 frequency, the frequency shows the lowest value in three out of four cases in Table III, but the F2 frequency shows the lowest value in Table II. This results mean that, although the F2 frequency is a robust feature in Lombard conditions, F2 frequency does not better represent the characteristics of the vocal track of the speaker than other formant frequencies. Note that F3, F4 frequencies show higher chi-square values between subjects in neutral condition, and the lower value under different Lombard conditions than other features. Therefore, they are good candidate features for speaker verification.

III. Long Term Formant Distribution

From the analysis results obtained from the Lombard speech database in the previous section, it concludes that features related to the formant frequency have robust characteristics. However, since the formant frequency and bandwidth vary depending on the type of phonetic information, it may not be a good choice to apply the formant features to text independent speaker verification systems. In forensic applications, such problem is overcome by taking a formant distribution of frequency with a long sentence period, which is called an LTF distribution feature [10]. The LTF distribution represents vowel formant frequencies over an entire input sentence. To obtain the LTF distribution feature, first, the vowels that have clear formant structures need to be extracted from the recorded

TABLE II
Kruskal-Wallis analysis of formant frequency between ten subject under the neutral condition

	F1	F2	F3	F4
chi-square	2.38E3	1.96E3	4.25E3	4.06E3

TABLE III
Kruskal-Wallis analysis of formant frequency under the Lombard condition (Pink 75 dB) for four female subjects

	Subject	F1	F2	F3	F4
chi-square	subject2	723.57	31.13	47.23	11.02
	subject3	462.05	4.81	64.78	10.92
	subject5	807.89	4.33	110.31	78.71
	subject8	518.20	0.19	47.81	40.81

speech signal. The formant features (frequency, bandwidth) are extracted from linear prediction coefficients (LPC) analysis, then the features are modeled with Gaussian mixture model (GMM). In order to better represent the characteristics of the speakers using LTF distribution, the appropriate length of the training utterance has to be defined. Fig. 5 depicts the variation of EERs by the length of utterances. The performance does not change much if the length of the utterance is longer than 15 seconds. Honglin *et al* showed similar results that a speaker model well represented the individual vocal tract characteristics, if more than 20 seconds of vowel duration were used [21].

IV. Experiments

The LTF frequencies and its corresponding bandwidths were used for speaker verification experiments in a neutral condition and six Lombard conditions. To compare the performance of the proposed method with a conventional system, a Gaussian mixture model-universal background model (GMM-UBM) system with Mel-frequency cepstral coefficient (MFCC) features was also implemented.

In the speaker verification experiments, features were extracted from utterances using 20 ms Hamming window and 10 ms shift. Only the vowel region of each utterance was used for extracting features. For the LTF features, the F1 to F4 frequencies and its corresponding bandwidths were extracted.

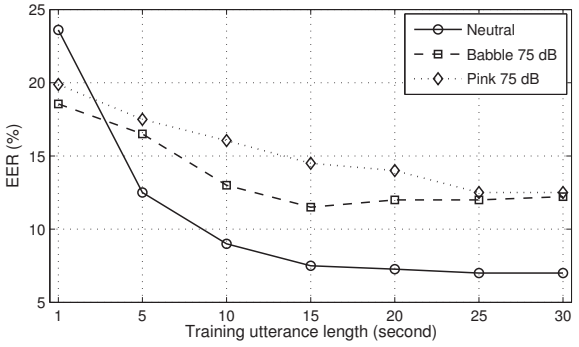

Fig. 5. EER versus training utterance length in three test conditions

TABLE IV
EER FOR DIFFERENT FORMANT FREQUENCY FEATURES UNDER EACH NOISE TYPE

Features	EER (%)								
	N	B65	B70	B75	P65	P70	P75	AVG.	STD.
F1–2 freq.	30.00	30.50	31.11	32.94	33.50	33.00	32.50	31.94	1.38
F2–3 freq.	13.00	19.88	23.00	25.16	21.50	20.83	22.00	20.77	3.82
F3–4 freq.	12.11	**11.88**	16.66	17.11	15.00	16.00	15.50	16.18	2.76
F1–3 freq.	15.38	20.16	23.50	24.27	21.16	21.50	21.66	21.09	2.88
F2–4 freq.	**8.66**	12.94	**15.00**	**15.61**	**13.00**	**15.27**	**15.00**	**13.64**	2.45
F1–4 freq.	11.50	13.00	16.88	15.50	13.50	15.66	14.94	14.93	**1.32**

TABLE V
EER FOR DIFFERENT FORMANT FREQUENCY AND CORRESPONDING BANDWIDTH FEATURES UNDER EACH NOISE TYPE

Features	EER (%)								
	N	B65	B70	B75	P65	P70	P75	AVG.	STD.
F1–2 freq. & bw.	24.00	26.22	28.94	29.33	28.33	29.27	31.50	28.23	2.43
F2–3 freq. & bw.	10.00	17.00	17.50	22.05	15.61	17.55	17.38	16.73	3.57
F3–4 freq. & bw.	7.50	9.00	12.00	13.11	10.00	**12.77**	12.50	10.98	**2.16**
F1–3 freq. & bw.	8.50	17.00	18.00	20.11	18.00	18.50	16.44	16.65	3.78
F2–4 freq. & bw.	6.50	**8.50**	**11.27**	13.11	**10.00**	13.00	**10.22**	**10.37**	2.38
F1-4 freq. & bw.	**6.22**	9.22	13.83	**12.00**	12.05	14.00	12.33	11.38	2.77

Thirteenth dimensional MFCC (13-MFCC) and its differential coefficients were also extracted. These features were utilized for training a UBM and speaker models. One hundred male and one hundred female subjects from the TIMIT database were used for training the UBM of which has 512 mixtures [22]. Speaker models were trained by applying a *maximum a posteriori* (MAP) adaptation [23] to the UBM. Note that only neutral speech was used for training both UBM and MAP adaptation. Eighteen neutral sentences among twenty sentences for each subject were used for training and the remaining two sentences from seven test conditions (a neutral and six Lombard conditions) were used for testing. Ten fold cross validation technique was introduced to avoid overfitting. For the speaker verification test, twenty sentences from each seven test conditions were used in a trial set. The trial set consisted of two true sentences and eighteen imposter sentences. Imposter sentences were composed of two sentences from each nine imposters. In the figures and tables, the neutral condition is marked as 'N', babble noise as 'B', and pink noise as 'P'. The numbers following each character denote noise levels in dB-SPL. The formant frequency is marked as 'freq.' and bandwidth as 'bw.'

Table IV shows the equal error rate (EER) of the verification test results. As we expected in section 2, experiments with combinations of two frequencies show that the combination of F3 to F4 frequencies perform higher performance than other combinations. Also, experiments with the combination of three frequencies show that the combination of F2 to F4 frequencies shows higher performance than other combinations. The verification test results with both formant frequency and its corresponding bandwidth are summarized in Table V. The results show the similar trend to the first experiment, but the EER has improved in this experiment compared to the experiment with frequency features only. Table VI summarizes experimental results with MFCC and a combination of formant

TABLE VI
EER FOR FORMANT FEATURES AND MFCCS UNDER EACH NOISE TYPE

Features	EER (%)								
	N	B65	B70	B75	P65	P70	P75	AVG.	STD.
13-MFCC	4.00	7.72	11.94	14.11	9.33	9.77	11.38	9.75	3.26
12-MFCC	6.00	6.05	7.83	9.38	9.11	8.05	8.88	7.90	1.39
24-MFCC	5.77	5.00	6.83	**7.11**	9.00	7.05	8.05	6.97	**1.33**
12-MFCC & F2–4 freq. & bw.	**3.50**	**4.00**	**6.00**	7.50	**5.77**	**7.00**	**7.61**	**5.91**	1.64

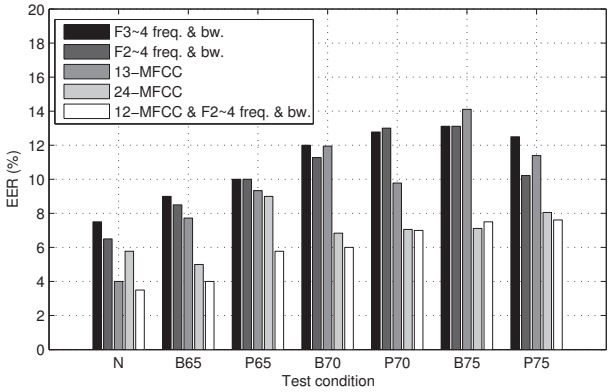

Fig. 6. EER for each features in seven test conditions

features. The first order of MFCC represents the energy of the speech signal, thus it needs to be removed in Lombard conditions. Experimental results also show that twelfth dimensional MFCC (12-MFCC) feature that does not include the first coefficient show higher performance than 13-MFCC in noisy conditions. The average and standard deviation of EER for each feature confirm that the features including second to forth formant frequencies and their corresponding bandwidths show very good performance in speaker verification tasks. Fig. 6 depicts the EER of five types of features. Although formant related features are only sixth dimension, the EER of the formant related features are similar to the one using 13-MFCC in the Lombard condition. We also found similar results when MFCC and formant related features were combined; the EER of 12-MFCC with sixth dimensional formant features is lower than that of twenty-fourth dimensional MFCC (24-MFCC) feature which consists of 12-MFCC and its differential coefficients. It confirms that the formant feature represents speaker characteristic more efficiently than the MFCC feature in Lombard conditions.

V. CONCLUSIONS

This paper analyzes a Lombard speech database on source or filter related features. Under six Lombard conditions, formant related features (especially F2 to F4 frequencies) showed more consistent characteristics than other features. In speaker verification experiments, LTF distribution features with F2 to F4 frequencies and their corresponding bandwidths showed the best performance in a neutral and six Lombard conditions. Analysis and experimental results confirm that formant related features obtained from long intervals are one of the key features for speaker recognition in noisy environments.

REFERENCES

[1] J. J. Dreher and J. ONeill, "Effects of ambient noise on speaker intelligibility for words and phrases," *The Journal of the Acoustical Society of America*, vol.29, no.12, pp.1320–1323, 1957.

[2] P. Rajasekaran, G. Doddington, and J. Picone, "Recognition of speech under stress and in noise," *in Proc. of ICASSP*, vol.11, pp.733–736, 1986

[3] I. Karlsson, T. Banziger, J. Dankovicova, T. Johnstone, J. Lindberg, H. Melin, F. Nolan, and K. Scherer, "Speaker verification with elicited speaking styles in the verivox project," *in Speech Communication*, vol.31, no.2, pp.121–129, 2000.

[4] J. H. Hansen and V. Varadarajan, "Analysis and compensation of lombard speech across noise type and levels with application to inset/out-of-set speaker recognition," *IEEE trans. on Audio, Speech, and Language Processing*, vol.17, no.2, pp.366–378, 2009.

[5] J.-C. Junqua, S. Fincke, and K. Field, "The lombard effect: A reflex to better communicate with others in noise," *in Proc. of ICASSP*, vol.4, pp.2083–2086, 1999

[6] V. S. Varadarajan and J. H. Hansen, "Analysis of lombard effect under different types and levels of noise with application to in-set speaker id systems." *in Proc. of InterSpeech*, 2006.

[7] R. Patel and K. W. Schell, "The influence of linguistic content on the lombard effect," *Journal of Speech, Language, and Hearing Research*, vol.51, no.1, pp.209–220, 2008.

[8] G. Bapineedu, B. Avinash, S. V. Gangashetty, and B. Yegnanarayana, "Analysis of lombard speech using excitation source information." *in Proc. of Interspeech*, pp. 1091–1094, 2009.

[9] T. Drugman and T. Dutoit, "Glottal-based analysis of the lombard effect." in *in Proc. of Interspeech*, pp. 2610–2613, 2010.

[10] F. Nolan and C. Grigoras, "A case for formant analysis in forensic speaker identification," *International Journal of Speech Language and the Law*, vol.12, no.2, pp.143–173, 2007.

[11] J.-C. Junqua, "The influence of acoustics on speech production: A noise-induced stress phenomenon known as the lombard reflex," *in Speech Communication*, vol. 20, no. 1, pp. 13–22, 1996.

[12] R. Schulman, "Dynamic and perceptual constraints of loud speech," *The Journal of the Acoustical Society of America*, vol.78, no.S1, pp.S37–S37, 1985.

[13] Z. Bond, T. J. Moore, and B. Gable, "Acoustic–phonetic characteristics of speech produced in noise and while wearing an oxygen mask," *The Journal of the Acoustical Society of America*, vol.85, no.2, pp.907–912, 1989.

[14] T. Becker, M. Jessen, and C. Grigoras, "Forensic speaker verification using formant features and gaussian mixture models." *in Proc. of Interspeech*, pp.1505–1508, 2008.

[15] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Trans. on Speech and Audio Processing*, vol.3, no.1, pp.72–83, 1995.

[16] V. Varadarajan, J. H. Hansen, and I. Ayako, "Ut–scope–a corpus for speech under cognitive/physical task stress and emotion," *in Proc. of LREC*, pp. 72–75, 2006.

[17] "Noisex–92 database," *http://spib.rice.edu/spib/select noise.html.*

[18] K. Sjolander and J. Beskow, "Wavesurfer-an open source speech tool," *in Proc. of Interspeech*, pp.464-467, 2000.

[19] B. Yegnanarayana, K. Sharat Reddy, and S. P. Kishore, "Source and system features for speaker recognition using aann models," *in Proc. of ICASSP*, vol.1, pp. 409–412, 2001.

[20] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and mfcc features for speaker recognition," *IEEE Signal Processing Lett.*, vol.13, no.1, pp.52–55, 2006.

[21] C. Honglin and K. Jiangping, "Speech length threshold in forensic speaker comparison by using long-term cumulative formant (ltcf) analysis," *in Proc. of ICCCN*, pp. 418–421, 2012.

[22] J. S. Garofolo, L. D. Consortium, "TIMIT: acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium* , 1993.

[23] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol.10, no.1, pp.19–41, 2000.