THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA VOL.24, NO.1E 2005. 3 pp. 16-20

Quantitative Measure of Speaker Specific Information in Human Voice: From the Perspective of Information Theoretic Approach

Samuel Kim*, Jung-Tae Seo**, Hong-Goo Kang*
*Department of Electrical and Electronic Eng., Yonsei University, **Information Control Engineering, Chungju University
(Received May 17 2004; revised September 3 2004; accepted February 24 2005)

Abstract

A novel scheme to measure the speaker information in speech signal is proposed. We develope the theory of quantitative measurement of the speaker characteristics in the information theoretic point of view, and connect it to the classification error rate. Homomorphic analysis based features, such as mel frequency cepstral coefficient (MFCC), linear prediction cepstral coefficient (LPCC), and linear frequency cepstral coefficient (LFCC) are studied to measure speaker specific information contained in those feature sets by computing mutual information. Theories and experimental results provide us quantitative measure of speaker information in speech signal.

Keywords: *Speaker recognition, Information-theoretic performance analysis, Speaker information, Homomorphic analysis*

I. Introduction

Apparently, human voice contains various information; not only phonetic sound but also the linguistic messages, intention or emotion of speaker, and even speaker's identity. There have been many researchers who tried to transmit or to interpret those information in speech signal. While speech communication systems and speech recognition systems are devised for transmitting phonetic sounds and extracting the linguistic messages, respectively, the speaker recognition system is designed to identify the person who is talking to machine[1].

The performance of the speaker recognition system has been remarkably improved by many devoted researchers. The state of art speaker recognition system has two main parts: feature extraction and classification. An important design issue is to choose the type of features that represents speaker identities, which should normally be transformed into low dimensionality with low computational complexity. The selection of features determines the separability of the speakers, and it also has large influence on the classification step because the classifier must be tuned to the given feature space. It is very difficult, however, to measure how discriminative the feature set is in a theoretic way. Furui measured distances between intra speakers and inter speakers and described the separability of cepstral feature set in[2]. Battiti used mutual information for feature selection in neural network classification system[3]. In [4], Kwak and Choi insisted that the error probability of a general classifier should be lower-bounded with the mutual information, where the probability densities were estimated by the Parzen window technique.

In this paper, we propose an information theoretic scheme to quantitatively measure the speaker information from speech signal. We also discuss about the amount of speaker specific information with various parametric features commonly used in speaker recognition systems. Even though our scope in this paper is limited to the speaker recognition application, the uniqueness

Corresponding author: Samuel Kim (worshipersam@mcsp.yonsei.ac.kr)
DSP Lab. B601 Dept. of E.E. Yonsei Univ.
134 Shinchondong Seodaemoongu Seoul 120-749 Korea

of this work can be found in the extendibility of the approach to other applications which need to measure speaker specific information from speech signals in near future.

II. Feature Extraction

As it was described in previous section, the main purpose of the feature extraction is how to compress the speech signals into lower dimension vectors without losing speaker specific information. It is widely believed that the vocal tract information through the homomorphic analysis contains fairly good speaker specific information, and many speaker recognition systems hire the cepstrum coefficient as their speaker specific feature vectors[2]. Lee et al. experimentally showed, however, that the cepstral coefficients are not the optimal feature vector by the performance comparison with line frequency spectrum (LSF)[5].

Several algorithms have been proposed to extract the cepstral coefficients, and they can be categorized according to the methods, such as mel frequency cepstral coefficient (MFCC), linear prediction derived cepstral coefficient (LPCC), and linear frequency cepstral coefficient (LFCC)[2,6,7]. Since many literature described on the details of it, we leave the readers to refer the articles[8,9].

Even though the cepstral vectors are widely used in many application and we use them in this paper, there is no way to prove the cepstral vectors are the optimal feature vectors. Intuitively, the higher level information, such as the pitch contour, intonation, and accents, can tell a lot about the speaker and can be transformed into feature vectors[10]. Those are, however, beyond the scope of this paper.

III. Quantitative Measure of Speaker Information

3.1. Mutual Information and Error Probability

Our previous work showed that the performance of a speaker recognition system is closely connected to the mutual information, $I(S; C)$, which represents the relationship between two random variables, i.e. feature vector set, C, and speaker, S. [11] We also proposed the upper and lower bounds for the performance could be derived from the information[12]. The classification error probability P_e is related to the mutual information between speaker and features as

$$\log N - H(P_e) - P_e \log(N-1) \le I(S;\mathbf{C}) \le \log N - H(P_e), \quad P_e \le 0.5 \tag{1}$$

where N is the number of registered speakers and $H(P_e)$ denotes the entropy. The first inequality holds equality if we have perfect symmetry, i.e. if we guess the correct speaker with a probability $1-P_e$, and any other speaker is guessed with the same probability $P_e/(N-1)$. The second inequality becomes equality if we guess the correct speaker with a probability $1-P_e$, and one of the other speakers totally dominates the rest of the probability, so that one erroneous speaker has probability P_e, and the others have probability of zero. See also[12] for the proofs.

Since it can not be solved easily unless we have prior knowledge of the actual problem, however, we propose approximated relationship between the mutual information and error probability as follows.

$$I(S;\mathbf{C}) \approx \log_2 N - \frac{H(P_e)}{1-P_e} \tag{2}$$

Figure 1 shows the lower and upper bound of the probability of error (solid lines) and the proposed relationship (dotted line) versus mutual information.

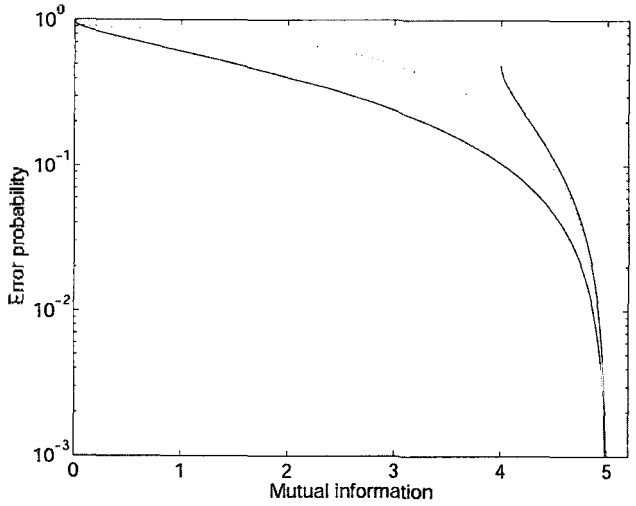


Fig. 1. Bounds for the probability of error versus mutual information.

3.2. Computing the mutual information

We now proceed with the computation of the mutual information. Note that we assume the features $C = \mathbf{c}_1, \cdots, \mathbf{c}_M$ described in the previous subsection contains all the feature vectors for the classification. It is reasonable assumption because many feature vectors derived from several seconds of speech. In order to simplify the calculations below, we will also assume that M equals 1 so that the classification is based on a single feature vector $\mathbf{c}$.

To compute the mutual information, we must rely on training database. We devide the N partitions, one for each speaker s: $\{\mathbf{c}_{s,t}\}_{t=1}^{T_s}, s = 1, \cdots, N$ where T_s denotes the number of features for speaker s. Noting that the overall pdf is a mixture of the pdf's for the speakers, with equal probability for each speaker, we compute the mutual information as

$$\begin{aligned} I(S;\mathbf{c}) &= h(\mathbf{c}) - h(\mathbf{c}|S) \\ &= -E\left[\log \sum_s \frac{1}{N} f(\mathbf{c}|s)\right] - D\left(\sum_s \frac{1}{N} f(\mathbf{c}|s) \middle\| \sum_s \frac{1}{N} f^{real}(\mathbf{c}|s)\right) \\ &\quad + E[\log f(\mathbf{c}|s)] + D\left(f(\mathbf{c}|s) \| f^{real}(\mathbf{c}|s)\right) \end{aligned} \tag{3}$$

where $D(f\|f^{real})$ is the relative entropy (or Kullback Leibler distance), which is the distance between two densities f and f^{real} that represents modeling errors[11]. We use the expectation maximization (EM) algorithm with Gaussian mixture modeling (GMM) to estimate the feature pdf's for each speaker[7]. If we assume that our EM algorithm works fine enough to ignore the modeling errors, then the mutual information can be represented as following.

$$\begin{aligned} I(S;\mathbf{c}) &= h(\mathbf{c}) - h(\mathbf{c}|S) \\ &\approx -E\left[\log \sum_s \frac{1}{N} f(\mathbf{c}|s)\right] + E[\log f(\mathbf{c}|s)] \\ &\approx -\frac{1}{\sum_{s=1}^{N} T_s} \sum_{s=1}^{N} \sum_{t=1}^{T_s} \log \sum_{s=1}^{N} \frac{1}{N} f(\mathbf{c}_{s,t}|s) + \frac{1}{N} \sum_{s=1}^{N} \frac{1}{T_s} \sum_{t=1}^{T_s} \log f(\mathbf{c}_{s,t}|s) \end{aligned} \tag{4}$$

3.3. Data Processing Inequality

The data processing inequality can be used to show that no processing of the data can increase the amount of information that we first get from the data. Suppose three random variable, X, Y, and Z, are said to form a Markov chain in the given order (denoted by $X \rightarrow Y \rightarrow Z$), then we have the data processing inequality,

$$I(X;Y) \geq I(X;Z) \tag{5}$$

which is easy to prove using the chain rules for mutual information[11]. A special case is when Z is a function of Y, $Z = g(Y)$, and therefore trivially fulfills the conditions in the inequality. Then we have

$$I(X;Y) \geq I(X;g(Y)) \tag{6}$$

with equality holds if and only if $g(Y)$ is an invertible function in the support region of Y. The original signal, the speech waveform, contains all information about the speaker, and each consecutive step can only decrease the information, or leave it unchanged. An invertible function will not reduce any information, while non-invertible functions will. The feature extraction can therefore not lead to any increased information about the speaker, but it will be able to reduce the complexity of the classifier.

Fig. 2 shows the diagram of the feature generation used in this paper. Let X, Y, and Z be the frequency representation of speech signal, smoothed one, and feature vector, respectively[6]. During the procedure, we will not consider FFT processing, which is an invertible process, because mutual information would not be changed as described in (15). Even though any process, either smoothing or truncation, cannot increase the mutual information between speaker and feature vector, we are still interested in these procedures because we would like to analyze

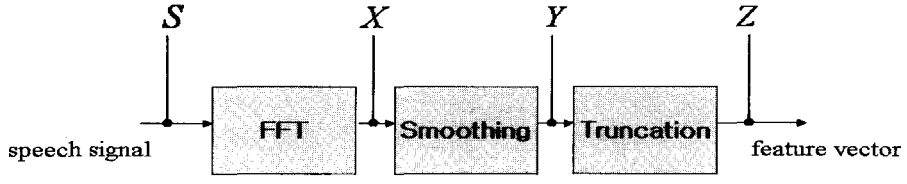


Fig. 2. Diagram of feature generation of homomorphic analysis.

what kind of process, or feature vector, can preserve the speaker information in it.

IV. Experiments

4.1. Setups

In this section, we analyze the results of several different features for speaker identification. To keep the total amount of information in each feature set approximately equal, we will use the same dimensionality of the feature vectors in the comparison. Gaussian mixture models will be carefully trained for each feature set and each speaker, and the mutual information between the features and the speaker is computed, according to the theory described in Section 3. We have used the YOHO database for the experiments[13]. There are 138 speakers, and for each speaker there are 4 sessions of 24 utterances for enrollment.

The pdf of each speaker is modeled by a GMM (optimized by the expectation maximization algorithm) using data from the enrollment sessions, after removing silence regions at the beginning and the end of each file. We use hamming-windowed speech with a frame length of 25 ms, and the frame step is 10 ms.

4.2. Results

Fig. 3 shows the mutual information versus the number of mixtures when we use the 12th order of MFCC. Intuitively, it implicates that the more mixture we use, the less effects of modeling errors we have. There are, however, a trade-off between the complexity and the performance, hence the number of mixtures should be varied depending on the application systems.

Fig. 4 depicts the mutual information of various feature sets versus feature dimension when we use 32 Gaussian mixtures. It shows that MFCC contains more speaker information than LPCC and LFCC, which implicates that the smoothing process of mel frequency filtering is more efficient than that of AR modeling and linear frequency filtering. Note that LPCC hires the linear prediction in which could be a leak of speaker specific information to the residual signals, while the others use FFT-based methods, which is a well known invertible process. Besides the loss of linear prediction, we can easily recognize the superiority of mel frequency filtering over the linear frequency filtering by comparing the mutual information of MFCC and LFCC.

To prove the theories described in this paper experimentally, we perform the speaker identification tasks with a single feature vector. Fig. 5 illustrates the classification error rate versus feature order, and showed MFCC outperforms the others. It confirms that the mutual information and the classification error rate are highly correlated.

The results provided in this paper are valuable information for narrow-band speech signal, but we may have different conclusion if we apply the idea to wide-band speech signal. It will be our future work.

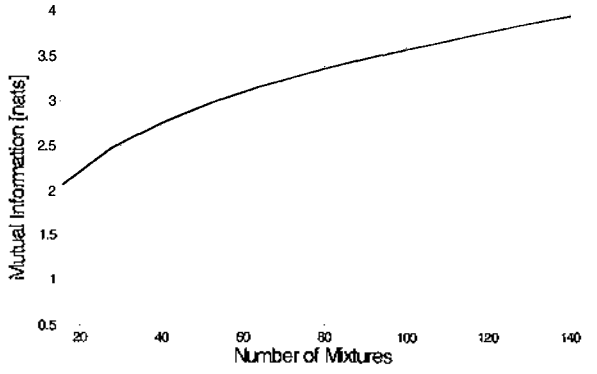


Fig. 3. Mutual information versus number of mixture for MFCC.

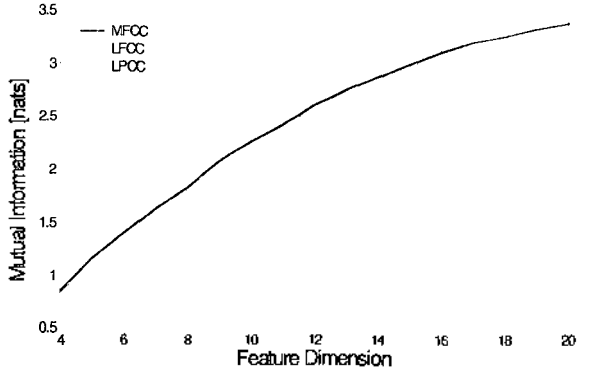


Fig. 4. Mutual information versus feature dimension for various feature vectors.

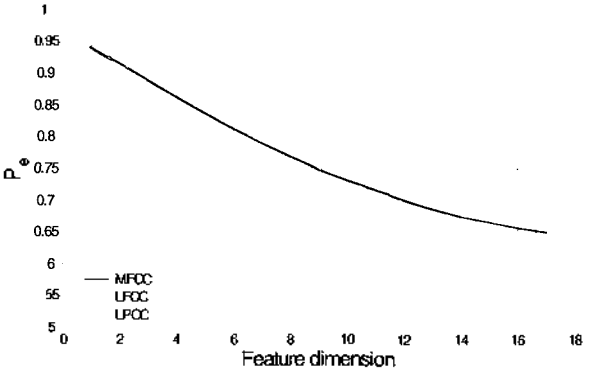


Fig. 5. Classification error with single feature vector versus feature dimension for various feature vectors.

V. Conclusion

We proposed a novel scheme to quantitatively measure the speaker information in speech signal with the information theoretic point of view. We measured the speaker information in various features which were commonly used and showed the experimental speaker identification error rate of them using narrow-band speech signal database. Since, we performed the measurement only for parametric feature sets so far, further work will be on measuring non-parametric feature sets.

VI. Acknowledgements

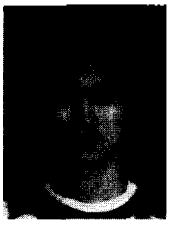
This work was supported by New Faculty Supporting Program at Yonsei University.

References

1. D. A. Reynolds, "Speaker identification and verification using Gaussian mixture models," Speech Communication, 17, 91-108, 1995.
2. S. Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Transaction on speech and audio processing, ASSP-29 (2), 254-272, Apr. 1981.
3. R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," in IEEE Transactions on Neural Networks, 5 (4), 537-550, 1994.
4. N. Kwak and C.-H. Choi, "Input Feature Selection by Mutual Information Based on Parzen Windows," IEEE Transactions on Pattern Analysis and Machine Intelligence, 24 (12), 1667-1671, 2002.
5. B. J. Lee, S. Kim, and H. G. Kang, "Speaker recognition based on transformed line spectral frequencies," Submitted to Intenational Symposium on Intelligent Signal Processing and Communication System, 2004
6. T. F. Quatieri, Discrete time speech signal processing, Prentice Hall, 2002.
7. D. A. Reynolds and R. C. Rose, "Robust text independent speaker identification using Gaussian mixture models," IEEE Trans. Speech and Audio Processing, 3, 72-83, 1995.
8. B.-H. Juang, L. R. Rabiner, J. G. Wilpon, "On the use of bandpass liftering in speech recognition," IEEE Transaction on acoustic, speech, and signal processing, ASSP-35 (7), 947-954, July 1987.
9. W. Verhelst and O. Steenhaut, "A new model for the short-time complex cepstrum of voiced speech," IEEE Transaction on acoustic, speech, and signal processing, ASSP-34, 43-51, Feb. 1986.
10. D. A. Reynold et al, "The SuperSID project: exploiting high-level information for high-accuracy speaker recognition," Proc. Internat. Conf. Acoust. Speech Signal Process., 784-787, 2003.
11. T. M. Cover, and J. A. Thomas, *Elements of information theory*, (Wiely, 1991).
12. T. Eriksson, S. Kim, and H.-G. Kang, "Theory for speaker recognition over IP," submitted to Interspeech 2004 - ICSLP, Apr. 2004.
13. J. P. Campbell, "Testing with the YOHO CD-ROM voice verification," Proc. Internat. Conf. Acoust. Speech Signal Process., 341-344, May 1995.

[Profile]

•Samuel Kim



He received the B.S. and M.S. degrees in electronic engineering from Yonsei University, Seoul, Korea, in 2003 and 2005, respectively. His major interests are focused on speech signal processing include recognition and enhancement. Currently, he is a researcher at CITY (Center for IT of Yonsei University) pursuing his Ph. D. degree.

•Jeong-Tae Seo

He received the B.S., M.S., and Ph.D degrees in electronic engineering from Yonsei University, Seoul, Korea, in 1985, 1987, and 1995, respectively. He was a researcher at Samsung Electronics from 1988 to 1990. Since 1995, he has been a assistant profess at the Information Control Engineering, Chungju University.

•Hong-Goo Kang

He received the B.S., M.S., and Ph.D degrees in electronic engineering from Yonsei University, Seoul, Korea, in 1989, 1991, and 1995, respectively. He was a Senior Technical Staff Member of AT&T Labs-Research, from 1996 to 2002. In 2002, he joined the Department of Electrical and Electronic Engineering, Yonsei University, where he is currently an Assistant Professor.