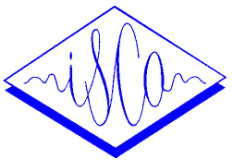

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

# SPEAKER INFORMATION ENHANCEMENT

*Fangxin Chen*

**Eloquent Technology Inc.**
**Ithaca, NY, USA**
**E-mail: Fangxin@eloq.com**

## ABSTRACT

This study consists of two experiments*. Experiment A investigates speaker-information distribution in the parametric domain. Experiment B compares different weighting strategies for speaker-information enhancement. The results indicate that weightings based on speaker-information distribution in the parametric domain yield better speaker recognition performance.

## 1. INTRODUCTION

Speech contains both phonetic (linguistic) and speaker information components. Phonetic information has three distinct linguistic functions: the *meaning-differentiating function*, which distinguishes the individual units of meaning; the *culminative function*, which indicates the important linguistic units contained in a particular utterance; and the *delimitative function*, which signals the boundaries between the linguistic units [1]. Speaker-information, on the other hand, is related to a speaker's voice quality, which permits the hearer to identify the individual speaker. Speaker-information is usually not under the speaker's conscious control and it consists mostly of the invariant aspects of a speaker's anatomical structure and involuntary nature of vocal settings.

Though interrelated, speaker and phonetic information have their own distinctive representations in the acoustic and parametric domain. In speech recognition, speaker idiosyncrasy is mostly a confounding factor and needs to be suppressed. However, in speaker recognition, this information should be maximally enhanced to achieve better performance. The present study investigated speaker-information distribution in the parametric domain In Experiment A, and then in Experiment B applied different weighting strategies for optimal speaker-information enhancement.

## 2. EXPERIMENT A DESIGN

The first experiment in this study is to investigate speaker information distribution in the parametric domain. *Mel frequency cepstrum coefficients* (MFCC) was chosen for parametric representation of the speech signal because of its wide use in speech signal processing.

The basic assumption of this experiment is that the variance in each MFCC coefficient contains certain amount of speaker-information, which may contribute to the separation of one speaker's voice from the other's. To estimate the amount of speaker-information coded in each MFCC coefficient for a particular speaker, we can intentionally exclude the variance in each MFCC coefficient from a speaker recognition test and see how it statistically affects the separation of this speaker from the rest speakers in the database. The degree of its effect can be used as an indicator of the amount of speaker-information coded in that particular MFCC coefficient.

To achieve this purpose, the *speaker identification error rate measurement* was used to estimate the speaker-information distribution in the MFCC coefficients. The speaker identification error rate measurement is in effect a closed-set text-dependent VQ speaker identification test.

First, a set of VQ word models were built for each speaker. In the testing phrase, the program was designed in such a way that it would automatically identify the input word's index and load all the speakers' trained VQ codebooks with the same word index. Then, this word would be matched with all those codebooks to find the one with the least VQ distortion score. If the codebook had the same user ID as the one of the input word, then, this utterance would be counted as a correct identification, otherwise as an identification error. A speaker's identification error rate (IER) is the percentage of the identification errors. For an estimation of speaker-information coded in each MFCC coefficient, the baseline performed a speaker identification test for each speaker, with the variances of all the MFCC coefficients included in the VQ distortion measure. The subsequent tests repeated that same testing procedure except that the variance in one of the MFCC coefficients was excluded from the VQ distortion measure. The normalised IER score (NIER), which was the difference score between the IER score that was calculated with the exclusion of the variance in one of the MFCC coefficients and the baseline IER score, was used as an indicator of how that particular coefficient could affect the speaker identification performance, or in other words, how much speaker-information was coded in that particular MFCC coefficient. The NIER score depended on the amount of speaker idiosyncrasy contained in that particular coefficient. There are three possibilities: if a coefficient contained significant speaker-information, the NIER score would be high; if a coefficient contained little speaker-information, the NIER score would basically remain at zero; if the MFCC coefficient contained significant confounding variation for speaker recognition, the NIER score could be negative.

The speech database used in this experiment is TI-46. This database consists of two sets of vocabulary: TI-ALPHA and TI_20. In the present investigation, only the TI-20 was used. TI-20 data corpus contains 20 isolated short words with 20 repetitions by 16 speakers: 8 male speakers labelled M1 to M8 and 8 female speakers labelled F1 to F8. There were nine recording sessions for each speaker. The first session recorded 200 tokens, 10 repetitions for each word. In the other eight sessions, each one recorded 40 tokens in a different random order, with 2 repetitions for each word. TI-20 database has total

26 repetitions for every word item from 9 separate recording sessions by each speaker, which provides an adequate number of sessions and tokens for the present study. The data assignment in this study was as follows: Session 1 was designated to the speaker VQ model training; Session 2-5 were used for the measurement of speaker-information distribution in MFCC; Session 6-9 were reserved for Experiment B. Since TI-20 data were originally designed for speech recognition purpose, the vocabulary consists mostly of single-syllable short words. This utterance length is not long enough for yielding high speaker recognition performance due to lack of sufficient speaker information coded in the speech signal. However, our study was only interested in maximally extracting speaker-information from a limited speech source. This data provided the necessary challenging environment.

## 3. EXPERIMENT A RESULTS

The result of the averaged NIER distribution in the MFCC coefficients over all male speakers is presented in Figure 1; Figure 2 is for the female counterpart; Figure 3 shows the averaged NIER distribution over all the speakers. The results of individual speakers' NIER score distributions are not presented in this paper due to limited space.

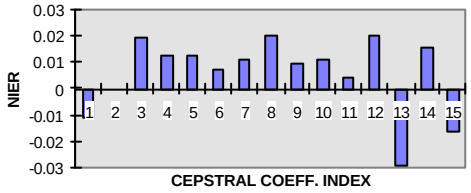


**Figure 1:** Averaged NIER score distribution over the male speaker group.

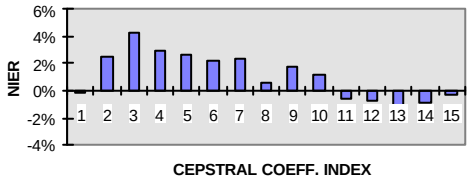


**Figure 2:** Averaged NIER score distribution over female speakers.

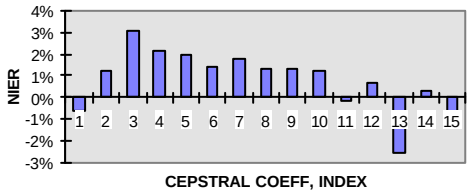


**Figure 3:** Averaged NIER score distribution over all speakers.

From the results of this experiment, we can come to following tentative conclusions:

- In general, the lowest- and the higher-order MFCC coefficients tend to contain less speaker-information compared with the lower- and middle-order ones; For many speakers, the variances contained in these coefficients (especially for the lowest coefficient C1) are confounding factors for speaker recognition;
- Female speakers tend to have much less speaker-information distributed in the higher-order MFCC coefficient region as compared with male speakers;
- Though there exist the above general tendencies in the distribution pattern of speaker-information, the amount of speaker-information in each coefficient is largely speaker-dependent. In other words, each individual speaker's idiosyncrasy is coded in the MFCC coefficients in its unique way.

## 4. EXPERIMENT B DESIGN

A baseline and three different weighting methods were applied in the closed-set speaker identification tests for comparison. The baseline speaker identification test used only the standard inverse-variance weighting [2]. For the other three methods, different weighting functions were added respectively, in addition to the inverse-variance weighting.

***Weighting Function I*** used the raised sine function [3], a popularly used weighting strategy for speech information enhancement;

***Weighting Function II*** assigned weight to each MFCC coefficient based on the averaged distribution pattern of speaker-information in the MFCC coefficients. Raised sine function weighting used by *Weighting Function I* was originally intended for speech recognition. Therefore, the lower-order coefficients were given low weights because the variances in those coefficients tend to be more related to speaker variation, as well as to the transmission channel variation. For speaker recognition, however, the variability due to speaker variation should not be suppressed. Furthermore, the average distribution pattern of speaker-information (see Figure 3) shows that the speaker-information distribution is not a sine function and the higher-order coefficients tend to have little or even negative speaker-information. A better approach for speaker-information enhancement, then, should adopt a strategy which weights the MFCC coefficients according to the distribution pattern of speaker-information, which is obtained from the speech training data. *Weighting Function II* is based on this assumption, and weights each MFCC coefficient according to its ranking in the amount of speaker-information. The coefficients were sorted in an increasing order according to their respective NIER scores. The weights rank from 1 to 15. The lowest-ranking coefficient is assigned the weight value of 1 and with the increment of 1 for each subsequent coefficient. If two or more coefficients have the same NIER scores, they receive the same weight. In that case, the highest weight will be lower than 15. This weighting function is applied to all the speakers in the testing phase.

***Weighting Function III*** assigned weights based on the individual speakers' distribution patterns of speaker-information in the MFCC coefficients. *Weighting Function II* assumed that the average distribution pattern of speaker-information was applicable to all the speakers. Therefore, the same set of weights was applied to all the speakers indiscriminately in the testing phase. As we have already pointed out, although there exists a general tendency, speaker-information distribution is largely speaker-dependent. Some individual speakers' speaker-information distribution patterns are quite inconsistent with the averaged distribution pattern of speaker-information. For example, the middle-order coefficients contain much speaker-information in the averaged speaker-information distribution pattern. However, some individual speakers' pattern showed negative NIER or NIDD scores in that region. Optimal speaker-information enhancement, then, has to depend on individual speakers' distribution patters of speaker-information. *Weighting Function III* adopted this approach. The weight assigned to each coefficient was based on the ranking of NIER scores. In this respect, *Weighting Function III* is the same as *Weighting Function II*. However, the ranking of speaker-information in *Weighting Function II* was based on the average NIER score distribution, and the weighting assigned to each coefficient was the same to all the speakers. The ranking of speaker-information in *Weighting Function III*, on the other hand, was based on each individual speaker's NIER score distribution and the weighting assigned to each coefficient was individual speaker-dependent.

The above three weighting methods can also be classified into two categories: *the general weighting* approach and *the* individual-*speaker-based weighting* approach**.** In *the general weighting* approach, the same weighting function was applied to all the speakers indiscriminately in the VQ Euclidean distance measurement.

$$d(x, x') = \sum_{i=1}^{p} (w_i | xi - x'i|^2 \ Mi)$$

where *wi* is the *ith* MFCC coefficient's inverse-variance weight; $M_i$ , in the case of *Weighting Function I*, is the *ith* MFCC coefficient's raised sine function weight, and in the case of Weighting Function II, the *ith* MFCC coefficient's general speaker-information weight.

In the *general weighting* approach, $M_i$ is the same for all the speakers. *The general weighting* approach includes *Weighting Functions I* and *II*.

In *the* individual-*speaker-based weighting* approach, weighting was based on individual speaker's speaker-information distribution patterns, and different weightings were applied to different speakers in the Euclidean distance measurement.

$$d(x, x') = \sum_{i=1}^{p} (w_i | xi - x'i|^2 \ M_{ji})$$

where *wi* is the *ith* MFCC coefficient's inverse-variance weightand $M_{ji}$ is the *jth* speaker's *ith* coefficient speaker-information weight.

*Function III* belongs to *the individual-speaker-based weighting* approach.

## 5. EXPERIMENT B RESULTS

The IERs of the baseline and all the weighting methods are fairly high because of using very short words for speaker identification training and testing. However, this in principle should not affect the purpose of the current experiment, which was to compare different weighting methods for speaker identification improvement. The averaged IERs for the speakers of the TI-20 data with the baseline and three different weighting approaches are listed in the following table:

| Method | Baseline | Weighting I | Weighting II | Weighting III |
|---|---|---|---|---|
| Average IER | 33.81% | 30.36% | 27.40% | 26.70% |

**Table 1:** Averaged Identification Error Rate (IER) for different weighting approaches

*Weighting Function I* reduced the averaged IER by 3.45%, compared with the baseline. Since this weighting function was originally designed for enhancing phonetic information, the effect of this function in improving speaker recognition indicates that there exists a fairly strong correlation between the distribution patterns of speaker and phonetic information in the parametric domain. As pointed out by O'Shauhgnessy [6]: "Most of the parameters and features used in speech analysis contain information useful for the identification of both the speaker and the spoken message. (p.480)". In the speech signal, the same acoustic phenomena, such as formant frequencies, carry both phonetic and speaker cues. Studies on speech perception [5,6] indicate that speaker-information actually facilitates listeners' phonetic processing for some perceptual tasks. This suggests that the human's perceptual system treats speaker-information as an integrated component of the acoustic cues for speech recognition, and there is an inherent relationship existing between phonetic and speaker information.
In spite of the fact that *Weighting Function I* improved speaker recognition performance in general, there were three speakers (F4, F6 and M4) whose error rates actually increased and another two speakers (F8 and M3) whose error rates remained the same, compared with the baseline .

*Weighting Functions II* was based on the average distribution pattern of speaker-information in MFCC according to either NIER scores and it achieved better performance than the baseline with an IER reduction of 6.41%. It also outperformed *Weighting Function I* with error rate reduction of 2.84%. This result supports our argument that speaker-information has its distinct distribution pattern in the acoustic and parametric domain, which can be identified and enhanced effectively for speaker-recognition.

*Weighting Function III* was based on the individual speakers' NIER score distributions. In Comparison with the baseline and

*Weighting Function I*, this approach performed better. It also yielded slightly overall better performance than the Weighting Function II. One advantage of this approach is that this individual-speaker-information-distribution-based approach reduced nearly all speakers' identification error rates except speaker M3. As for this particular speaker, there may be an explanation for the lack of improvement even with the use of this individual-speaker-dependent weighting strategy. The error rate for this particular speaker in the baseline is the lowest (1.88%). Compared with the average error rate 33.81%, improvement for this speaker's error rate might have already been saturated.

## 6. CONCLUSIONS

Based on the above experimental results, we tentatively conclude that

- The weighting approaches based on speaker-information-distribution perform better than the conventional speech weighting method (*Weighting Function I*) for speaker-information enhancement;
- The weighting approach based on individual speakers' speaker-information-distribution has one important advantage over the general weighting approach (*Weighting Function I* and *II*), that is, it basically effective for all individual speakers.

The better speaker identification performance by using individual speakers' distribution patterns conforms with a voice perception theory that different acoustic cues are used in distinguishing different voices [7]. According to Lancker et al. the critical parameter(s) for speaker-information are not the same for all the voices. Speech contains a constellation of potential cues from which the listener "selects" a subset to use for identifying a given voice. The acoustic cue(s) essential for distinguishing one speaker's voice may be expendable in the case of distinguishing another speaker's voice. Loss of certain parameter(s) will not impair recognisability if a voice is sufficiently distinctive on some other dimensions.

In conclusion, this study developed ideas on how to improve speaker recognition via a technique designed to enhance those elements in the speech parameters most relevant to discriminate speakers. By applying appropriate weights in distance measures, the speaker-information is enhanced, which improves speaker recognition performance.

Though the present study investigated speaker-information in MFCC, the same principle for enhancing the speaker-related variability in the parametric domain should also be applicable to other speech parameters. In this respect, the study provides a methodology for speaker-information enhancement. The same technique should also be applicable for enhancing the phonetic-information in speech recognition. How well this approach could improve speech recognition performance is our next research interest.

* The experiments in this paper were conducted as part of the author's Ph.D work at University of Alberta, Edmonton, Canada.

**REFERENCES**


1. Trubetzkoy, N. S. (1969). Principles of phonology . Berkeley and Los Angeles: University of California Press.

2. Tohkura, Y. (1986). A weighted cepstral distance measure for speech recognition". ICASSP-86, p.761 –764.

3. Juang, B-H. , Rabiner, L. R. , & Wilpon, J. G. (1986). On the use of bandpass liftering in speech recognition, IEEE Trans. on ASSP, v.35, p.947-954.

4. O' Shaughnessy, D. (1987). Speech communication: human & machine. New York: Addison-Wesley.

5. Palmeri, T. J. , Goldinger, S. D. , & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. J. Exp. Psychol. 19. p.1-20.

6. Nygaard, L. C., Sommers, M. S , & Pison, D. B. (1994). Speech perception as a talker-contingent process. Psychol. Sci. 5. p.42-46.

7. Lancker, D. V. , Kreiman, J. , & Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters. Part I: Recognition of backward voices. Journal of Phonetics, 13, p.19-38.