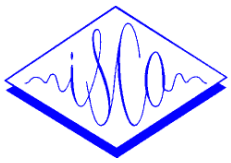

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

Phoneme-Level Voice Individuality Used in Speaker Recognition

Sadaoki Furui and Tomoko Matsui

NTT Human Interface Laboratories
3-9-11 Midori-cho, Musashino-shi, Tokyo, 180 Japan

ABSTRACT

This paper reports our recent investigation on voice individuality (speaker characterization) at the phoneme level of speech. The first half of the paper reports our perceptual experiments on speaker characterizing information included in continuous speech. Experimental results show that speaker-characterizing vowels differ from speaker to speaker. The latter half of the paper introduces a text-prompted speaker recognition method that we have recently proposed as an automatic speaker recognition method using phoneme units. In this method, the recognition system prompts each user with a new key sentence every time the system is used, and accepts the input utterance only when it decides that the registered speaker has uttered the prompted sentence.

1. INTRODUCTION

Individual information includes voice quality, voice pitch, loudness, speed, tempo, intonation, accent, and the use of vocabulary [1]. These arise both from individual hereditary differences in articulatory organs and from acquired differences in the manner of speaking. These features can also be classified into supra-segmental (prosodic) features and segmental features. Although supra-segmental features also characterize voice individuality, it has been shown that major speaker-characterizing information exists in the segmental features.

The segmental features can be classified into several levels such as syllable and phoneme levels. Among these, the phoneme level information is the most basic and principal information. If the phoneme-level voice individuality is fully clarified, it should be possible to improve automatic speaker recognition performance, and as a result, various new methods would probably be developed. However, phoneme-level voice individuality in continuous speech has not yet been deeply investigated, in either perceptual or automatic speaker recognition. We have conducted investigations from both the hearing and technological points of view.

Many papers in which certain phonemes are considered to be more reliable than others for use in perceptual and automatic speaker recognition have been published since 1960's. It has been generally observed that vowels and nasals provide relatively better performance than others [2, 3]. In daily life, it is generally observed that some people's voices are characterized by particular phonemes. For example, /a/ sounds similar to /e/, or /h/ and /s/ are reversed by some people especially in the Tokyo dialect. In other words, phonemes effective for speaker recognition are different from speaker to speaker. It is likely that human beings are naturally using these phoneme-dependent cues in perceptual speaker recognition. However, these phenomena have not yet been thoroughly and quantitatively investigated. In the first half of this paper, we report our investigation of perception from this point of view using syllable speech extracted from continuous utterances.

As described above, various new automatic speaker recognition methods using phoneme units can be considered. For example, if models of all the phonemes can be estimated for each speaker in the training stage, speech models of arbitrary sentences for each speaker can be made by concatenating these models. Therefore, sentences used for recognition can be changed every time the system is used. Thus one serious problem of conventional speaker recognition systems, that they can easily be defeated by playing back the recorded voice of a registered speaker, can be solved. We call this method the "text-prompted speaker recognition method," since the recognition system prompts each user with a new key sentence every time the system is used. The latter half of this paper describes the details and experimental results of this method.

2. PERCEPTUAL EXPERIMENTS ON PHONEME-DEPENDENT INDIVIDUAL INFORMATION

2.1 Experimental Methods

For the hearing investigation, we analyzed speaker characterizing information included in continuous speech. Subjective speaker identification experiments were conducted by using CVC (consonant-vowel-consonant) syllables extracted from continuous speech [4]. Since the variety of syllables that could be extracted from our speech database and the time length of experiments were limited, we focused on the analysis of center

vowels in the syllables. In order to reduce the effect of pitch individuality, eight male speakers having similar average pitch frequencies (165 - 175 Hz) were selected, and their CVC syllables were presented to 11 listeners. The listeners consisted of five males and six females. They were familiar with the voices of these speakers, and were requested to identify the speaker of each syllable. Each syllable was presented three times.

We used CVC syllables because the mean length of the center vowels was only 90.5 ms and it was very difficult to identify speakers by hearing only the vowel periods. In order to get the listeners to pay attention to the individual information in the vowel periods, we only used CVC syllables including a vowel longer than 50 ms, and truncated consonant periods beyond 30 ms using a smooth window. Because of the limitation on the length of the experiments, the combinations of preceding and succeeding consonants in these syllables were a subset of all the combinations in Japanese, and they were different from vowel to vowel. Since the number of syllables was 18 for each vowel for each speaker, the total number of syllable stimuli was 720(=5x8x18) .

2.2 Experimental Results

Table 1 shows speaker identification rates averaged over all the listeners and 18 syllables. Analysis of variance with two factors was applied to the identification rates of all the syllables after arcsin transformation. Results of the analysis of variance, shown in Table 2, indicate that all the factors including the cross factor between vowels and speakers are statistically significant (** denotes significance level at 1%). The cross factor means that speaker-characterizing vowels are different from speaker to speaker. To make clear the difference between the effective vowels across the speakers, we applied the Chi-squared test to the identification rates. The test results show statistical significance for seven speakers out of eight. Table 3 shows the most effective and the least effective vowels for each speaker.

Table 1 - Speaker identification rates for the normal (forward) syllables

Speaker	/a/	/e/	/i/	/o/	/u/	Average
S1	66.7	73.7	44.4	63.1	51.5	59.9
S2	56.1	20.2	33.8	49.5	38.4	39.6
S3	59.6	66.7	80.8	53.5	44.4	61.0
S4	40.9	44.4	50.0	58.6	67.2	52.2
S5	73.2	81.3	69.2	77.3	72.7	74.8
S6	93.9	84.3	42.4	78.3	93.4	78.5
S7	54.6	40.9	33.8	41.4	27.8	39.7
S8	59.6	48.5	37.4	77.8	48.5	54.3
Average	63.1	57.5	49.0	62.4	55.5	57.5

Table 2 - Analysis of variance table for the results shown in Table 1

Factor	Degree of freedom	Mean square ratio	Expected value of mean square ratio
Speakers	7	28.74 **	2.64
Vowels	4	6.15 **	3.32
Cross	28	4.55 **	1.70

Table 3 - Most effective and least effective vowels for each speaker

Speaker	Most effective	Least effective
S1	/e/	/i/
S2	/a/	/e/
S3	/i/	/u/
S4	/u/	/a/
S5	No significant difference	
S6	/a/, /u/	/i/
S7	/a/	/u/
S8	/o/	/i/

In order to investigate the effect of dynamic information including prosody in these CVC syllables in contrast with static information, a supplementary experiment was performed in which the syllables were played back in reverse. Tables 4 and 5 show the results for the reverse (backward) signals. They are similar to the results for normal conditions (forward signals) shown in Tables 1 and 2, except that the vowel effect for the backward signal is not significant. The most effective and the least effective vowels for each speaker were also analyzed, and it was found that they were exactly the same as those in Table 3. The similarity between the results for forward and backward signals indicates that static features in CVC syllables are perceptually more important than dynamic features.

Table 4 - Speaker identification rates for the reverse (backward) syllables

Speaker	/a/	/e/	/i/	/o/	/u/	Average
S1	62.6	78.3	54.6	60.1	52.0	61.5
S2	57.6	20.7	30.8	42.4	34.3	37.2
S3	56.1	67.7	84.3	52.5	59.6	64.0
S4	40.9	26.3	53.0	56.1	60.1	47.3
S5	75.3	71.7	61.1	66.7	66.2	68.2
S6	77.8	79.3	44.4	79.8	83.8	73.0
S7	61.6	39.4	40.4	46.0	30.8	43.6
S8	49.0	36.4	51.0	79.8	55.6	54.3
Average	60.1	52.5	52.5	60.4	55.3	56.2

Table 5 - Analysis of variance table for the results shown in Table 4

Factor	Degree of freedom	Mean square ratio	Expected value of mean square ratio
Speakers	7	19.23**	2.64
Vowels	4	2.72	3.32
Cross	28	3.78**	1.70

2.3 Discussion

Through these experiments, we have found that speaker-characterizing vowels are different from speaker to speaker, and that static features in CVC syllables play dominant roles in identifying speakers by listening.

The influence of the vowel duration on the identification rate for each syllable was analyzed, and it was found that the hypothesis of no correlation between them was rejected at a significance level of 1% for only two speakers out of the eight speakers.

We further investigated the relationships between the perceptually confusable speaker pairs and their physical distances in the cepstral domain. However, there was no clear correspondence between the perceptual and physical spaces.

3. AUTOMATIC SPEAKER RECOGNITION USING PHONEME UNITS

3.1 Text-prompted Speaker Recognition Method

As a result of our technological investigations, we proposed a text-prompted speaker recognition method [5]. This method is facilitated by using speaker-specific phoneme models as basic acoustic units. In this method, the recognition system prompts each user with a new key sentence every time the system is used, and accepts the input utterance only when it decides that the registered speaker has uttered the prompted sentence. Because the vocabulary is unlimited, prospective impostors cannot know in advance what sentence they will be asked to repeat. Thus a pre-recorded voice can easily be rejected. One of the major issues in this method is how to properly create the speaker-specific phoneme models with training utterances of a limited size for each speaker.

In our method, the phoneme models are represented by Gaussian mixture continuous HMMs or tied-mixture HMMs, and they are made by adapting speaker-independent phoneme models to each speaker's voice. Since the text of training utterances is known, these utterances can be modeled as the concatenation of phoneme models, and these models can be automatically adapted by an iterative algorithm. In the recognition stage, the system concatenates the phoneme models of each registered speaker to create a sentence HMM, according to the prompted text. Then the likelihood of input speech against the sentence model is calculated and used for the speaker recognition decision. If the likelihood of both speaker and text is high enough, the speaker is accepted as the claimed speaker.

Figure 1 shows a block diagram of the method. In order to properly adapt the models of phonemes that are not included in the training utterances, we have recently proposed a new adaptation method based on tied-mixture HMMs [6].

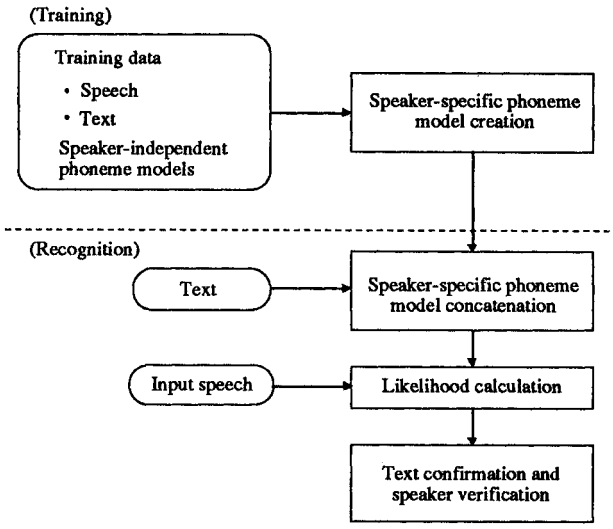

Fig. 1 - Block diagram of the text-prompted speaker recognition method.

3.2 Likelihood Normalization

The most significant factor affecting automatic speaker recognition performance is variations in signal characteristics from trial to trial (intersession variability, variability over time) [7]. Variations arise from those generated by the speaker him/herself, and differences in recording and transmission conditions as well as noise. The likelihood has a wide range, especially in the case of text-prompted speaker recognition where speech with different texts are uttered at different sessions.

Higgins et al. proposed a normalization method for distance (similarity, likelihood) values that uses likelihood ratio [8], and we proposed a normalization method based on a posteriori probability [5] :

$$\log L = \log p(X|S=S_c, T=T_c) - \log \sum_{S,T} p(X|S,T) \quad (1)$$

where X is the observed measurements, S is a speaker, S_c is the claimed speaker, T is text, and T_c is the prompted text.

The difference between the normalization method based on a posteriori probability and that based on likelihood ratio exists in whether or not the claimed speaker is

included in the speaker set for normalization; the normalization term (the right most term in Eq. (1)) for the a posteriori-probability-based method is calculated by using all the reference speakers including the claimed speaker, whereas the speaker set in the likelihood-ratio-based method does not include the claimed speaker. Experimental results indicate that both normalization methods are almost equally effective.

We have recently proposed a new method in which the normalization term is approximated by the likelihood for a Gaussian mixture which models the parameter distribution for free-text utterances by all the reference speakers [9]. This method has been shown to give much better results than either of the above-mentioned normalization methods.

3.3 Experimental Results

Recognition experiments were performed to evaluate the effectiveness of this method. Various sentences uttered by 30 speakers (20 male and 10 female) at five sessions over a period of roughly ten months were used. 10 male and 5 female speakers were used as customers and the remainder were used as impostors.

Experimental results show that, when the adaptation method for tied-mixture-based phoneme models and the likelihood normalization method were used, a speaker and text verification rate of 98.9% was obtained. The error rate was reduced to roughly 1/3 of that obtained without normalization.

3.4. Discussion

It was shown that the text-prompted speaker recognition method using speaker-specific phoneme units works very well. Various improvements are still possible. One is to select key sentences (phrases) containing a high proportion of 'effective' phonemes. Alternatively, a front-end classifier could be used to automatically identify effective phonemes, enabling appropriate biasing to be applied in the recognition stage.

4. CONCLUSION

This paper reported our recent investigation on perceptual speaker-characterizing information carried by phonemes and on a phoneme-based automatic speaker recognition method. Recent advances in speaker recognition are mainly due to the improvements in techniques for making speaker sensitive feature measures and models, and they have not necessarily come about as an outgrowth of new or better understanding of speaker characteristics or how to extract them from the speech signal. Our experimental results do not show any clear relationships between perceptual and physical speaker characteristics. However, for more effective speaker recognition systems, we expect that better understanding of speaker characteristics in the speech signal will play important roles.

ACKNOWLEDGMENT

The authors wish to thank Professor Irwin Pollack for his guidance and stimulating discussions during the perceptual investigation reported in the first half of this paper.

REFERENCES

[1] S. Furui: "Speaker-dependent-feature extraction, recognition and processing techniques," Speech Communication, 10, pp. 505-520 (1991)

[2] J. P. Eatock and J. S. Mason: "Phoneme performance in speaker recognition," Proc. Int. Conf. Spoken Language Processing, Banff, pp. II-1411-1414 (1992)

[3] J. P. Eatock and J. S. Mason: "A quantitative assessment of the relative speaker discriminating propertied of phonemes," Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, Adelaide, pp. I-133-136 (1994)

[4] T. Matsui, I. Pollack and S. Furui: "Perception of voice individuality using syllables in continuous speech," Proc. Fall Meeting of Acoust. Soc. Jap., 2-9-9 (1993) (in Japanese)

[5] T. Matsui and S. Furui: "Concatenated phoneme models for text-variable speaker recognition," Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, Minneapolis, pp. II-391-394 (1993)

[6] T. Matsui and S. Furui: "Speaker adaptation of tied-mixture-based phoneme models for text-prompted speaker recognition," Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, Adelaide, pp. I-125-128 (1994)

[7] S. Furui: "An overview of speaker recognition technology," ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, pp. 1-9 (1994)

[8] A. L. Higgins, L. Bahler and J. Porter: "Speaker verification using randomized phrase prompting," Digital Signal Processing, 1, pp. 89-106 (1991)

[9] T. Matsui and S. Furui: "Similarity normalization method for speaker verification based on a posteriori probability," ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, pp. 59-62 (1994)