LETTER

On the Importance of Transition Regions for Automatic Speaker Recognition

Bong-Jin LEE[†], ***Student Member***, **Chi-Sang JUNG**[†], **Jeung-Yoon CHOI**[†], ***and*** **Hong-Goo KANG**[†], ***Nonmembers***

SUMMARY This letter describes the importance of transition regions, e.g. at phoneme boundaries, for automatic speaker recognition compared with using steady-state regions. Experimental results of automatic speaker identification tasks confirm that transition regions include the most speaker distinctive features. A possible reason for obtaining such results is described in view of articulation, in particular, the degree of freedom of articulators. These results are expected to provide useful information in designing an efficient automatic speaker recognition system.

key words: *automatic speaker recognition, speech transition regions, phoneme class*

1. Introduction

There have been various research activities on assessing the relative importance of phonetic classes to speaker recognition. Fundamentally, they use the fact that some classes of phonemes, such as vowels and nasals, include more speaker-related information than obstruents such as stops, affricatives, and fricatives [1]. One of the approaches assigns different weights to each phoneme depending on how much each phoneme includes speaker related information [2]. Pelecanos et al. found an optimal feature type and a frame length of each phoneme for speaker identification [3]. Gutman and Bistritz investigated phoneme-adapted Gaussian mixture models (GMM) to enhance overall speaker verification performance [4].

In addition to using intra-phonetic information, there have been other approaches introducing inter-phonetic information such as coarticulatory effects in recognition [5]. In fact, conventional research in speech has focused on finding invariant cues for uncovering phonetic identity from the acoustic signal, and has traditionally regarded coarticulatory effects as a nuisance. We can then question whether those regions of speech which seem most problematic in yielding phonetic information should be rich in providing other unrelated information, such as speaker identity. In past studies in acoustic phonetics, "steady-state" characteristics corresponding to phonemes in the acoustic signal have been suggested as "targets" for articulatory realization. It has been also shown that speakers enjoy some degree of freedom in the production of a given phoneme [6], [7]. During the production process, as articulatory targets are continuously being supplied, the articulatory organs attempt to produce the configurations for the targets; however, they are not constrained by any targets during those intervals in between. These transition regions, which straddle the interval between acoustically steady portions of the speech signal, display much of the coarticulatory effects that occur between phonemes, and are most free to be produced by whatever combination or sequence of mechanisms each speaker wishes to employ, subject only to physical constraints. Thus, it is expected that these regions contain the major part of inter-speaker variability in the signal, which may be exploited in schemes for speaker identification. Actually, in [8], the authors have shown that the transition frames have more discriminative power than stationary frames. However, they only focused on building efficient speaker recognition system by utilizing difference between steady-state regions and transition regions without in-depth investigation of transition regions. Hence, in this paper, we focus on transition region itself and verification of speaker discriminative capability of transition regions.

The motivation of this paper is to re-examine the discriminative capability of each phonetic class, especially transitions. At first, we verify the relevance of our assumption by evaluating the recognition error rates as the number of test frames used in each class varies. To further investigate the impact of transition regions in detail we classify transition regions into four categories depending on the manner types of the preceding and the following phonemes, and compare recognition rates for each type of transition. Experimental results confirm the discriminative power of transition regions, especially in the segment boundary between two vowels. Finally, we describe possible reasons for obtaining the outcomes by linking them to the physical characteristics of voice production mechanism, especially to the degree of freedom while pronouncing sounds. The outcome of the experimental results and the analysis will provide useful information for designing speaker recognition systems.

2. Speaker Discriminating Capability of Phonetic Information

In this section, speaker discriminative capability of various phonetic classes is re-examined. We perform three experiments. First, we confirm the speaker recognition performance of each phoneme class when the test segment is a

Manuscript received July 14, 2009.
Manuscript revised September 11, 2009.

[†]The authors are with Electrical and Electronic Engineering, Yonsei University, 134 Shinchon-dong Seodaemoon-gu Seoul, 120–749, Korea.

DOI: 10.1587/transinf.E93.D.197

sentence. We can confirm the contribution of each phoneme class in practical speaker recognition tasks. Next we fix the length of the test segment for a fair comparison of speaker discriminative performance of each phoneme class. Finally, we test the speaker recognition performance of each transition type.

We set experiments as follows for the above three tests. First, we analyze speech signals in TIMIT corpus every 10 ms with 20 ms analysis frame length. All frames are then classfied into seven phonetic classes, i.e. stops, affricates, fricatives, nasals, glides, vowels, and transitions based on the indices that are already included in the TIMIT corpus [9]. If two or more phonemes are included within the analysis frame, we consider the frame as a transition class. While most speaker recognition applications adopt Mel-frequency cepstral coefficients (MFCCs) which have more filterbanks in low frequencies we adopt linear frequency cepstral coefficients (LFCCs) which have evenly placed filterbank in frequency domain to obtain the characteristic of the phoneme classes. After extracting LFCCs, we set up a Gaussian mixture model (GMM) based speaker identification system which is proposed by Reynolds and Rose [10]. During the training process, we use five sentences of TIMIT database to train each speaker and the number of mixtures of GMM is set to 16. Remaining five sentences are used for testing.

2.1 Practical Condition: One Test Sentence

A practically designed text-independent speaker recognition system requests the speaker to speak some words or sentences. Thus, we may not know the relative amount of tokens in each phoneme class and cannot control them. We perform an experiment to verify the speaker discriminative information of each phoneme in this practical condition. In the experiment of simulating practical condition, we analyze the identification performance of each phonetic class in the trial. In other words, we first need to concatenate test features corresponding to the same phonetic classes in a single trial sentence. It is reasonable to say that the test length of each class (i.e. the total length of all tokens per class) is different for this case. Table 1 shows the results. The table shows the relative amount and speaker recognition error rate of each phoneme class. We can see that almost half of the sentence consists of vowels and they also give the best speaker identification performance. Transitions also show good speaker identification performance even though the amount of transitions is much less than that of vowels. We can estimate that if the amount of transitions is similar to the amount of vowels, transitions may achieve a better result than vowels. Nasals, which are well-known as a good feature for speaker recognition [1] are the third but its error rate is much worse than vowels and transitions. This is because the total length of the nasal tokens in an ordinary sentence is much less than the length of vowels. As the table shows, the proportion of nasals is only 6% in a sentence.

Table 1 Speaker identification error rate of each class in the sentence.

Phoneme Classes	Error Rate (%)	% in TIMIT database
Stops	88.09	9
Affricates	89.44	2
Fricatives	63.47	16
Nasals	44.12	6
Glides	60.51	8
Vowels	5.90	40
Transitions	7.65	19

2.2 Controlled Condition: Same Test Length

As shown in the previous subsection, the proportion of each phoneme class in a sentence is different depending on the frequency of each phoneme. However, we need to fix the length of the test segment to be equal for all classes for fair comparison. Previous studies have fixed the length of the test segment [1], [3] or showed the results as an average likelihood versus the number of frames in each phoneme [2] for a similar reason. Thus, we also perform speaker identification test with a fixed test length. The length of the test segment varies from 1 frame to 70 frames and frames in each test segment are selected sequentially in each test sentence. During the experiment, we omit affricates because the length of the affricates is quite small. Figure 1 shows the results. In the figure, the x-axis denotes the number of test frames and the y-axis represents the speaker identification error rate. The error rates of all classes decrease as the number of test frames increases. The decrease of nasals, vowels, transitions is especially remarkable. Nasals give the best performance when the test length is very short while transitions give the best performance when the test length is long enough. As the table shows, the error rate of nasals is best when the number of test frames is 1 frame or 10 frames. We can guess that the reason of this is the variability of transitions. As the number of test frame increases, transitions can include much more information than nasals because there are many combinations in transitions while nasals have only a fixed number of phonemes. Therefore,

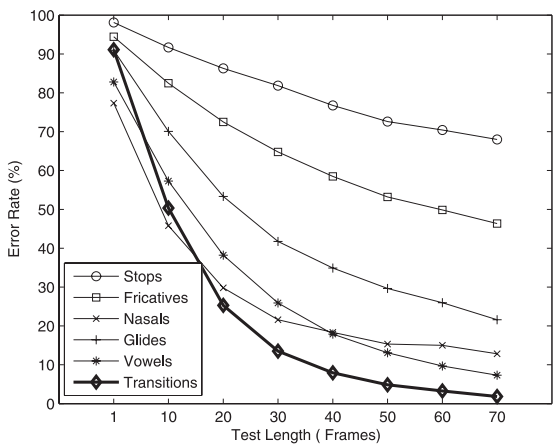

Fig. 1 Speaker identification error rates of each class as the length of test frames varies.

Table 2 Error rates of each transition type.

Speaker identification error rate (%)						
Class1 \ Class2		S	F	N	G	V
92.10	S				71.29	42.24
83.00	F				85.29	57.87
46.65	N		58.62			37.29
70.88	G					36.04
57.41	V	46.45	76.03	35.12	43.50	25.76

we need to know the type of transitions to verify the reason of the performance improvement.

2.3 Dependency on the Combination Type of Transition

In this experiment, we first classify all the transition regions by the preceding phoneme and the following phoneme with several types. For example, a transition frame may start with a vowel and end with a consonant, or vice versa. According to this transition type, speaker identification performance may be different. After classifying the transition types, we perform speaker identification tests. In these tests, the length of the test segment is set to 10 frames and we also omit affricates. Table 2 shows the start/end class and the error rates of each phonetic transition pair. In the table, Class1 denotes the class of the first half, and Class2 the second half of the transition. The error rates of non-transition classes are included in the first column of the table for comparison. If a given transition type has less than 10 frames for any speaker, it remains blank in the table. As shown in the table, most transitions are vowel-related transitions and they contribute to speaker identification performance more than others. Especially, as the table shows, vowel-to-vowel transitions achieves 25.76% error rate while the average error rate is about 50% in Fig. 1. Moreover, most vowel-related transitions give better results than steady-state vowels and nasal regions which achieve 57.41% and 46.65% error rates. This implies that transitions include more speaker discriminative information than nasals and vowels. In the following subsection, we will explain possible reasons why the transition regions seem to be the most important to speaker recognition.

2.4 Analysis of Experimental Results

In this subsection, we would like to analyze the simulation results given in the previous subsections by linking them with the concept of degrees of freedom in articulation. The results given in Table 1 show that vowels include much higher speaker-specific information than other classes in a practical condition. It has been verified by various experiments [11], [12]. Johnson et al. show that there is a significant difference in the articulatory gestures between speakers but is consistent within speaker in saying the same vowel, which can be interpreted similarly for our results [12].

From the results given in Fig. 1 and Table 2, we may conclude that transition regions include more speaker-discriminative information than any other phoneme class if the test condition is same. The results can be explained by the degree of freedom in articulation. We may assume that the degree of freedom increases in transition regions because human should dynamically change the articulatory organs to pronounce the sound. In a situation where the start and the ending phonemes are fixed targets, speakers enjoy more freedom in between. Since the articulatory gestures are different between speakers and the degree of freedom in articulation is increased in the transition region, we may expect that the transition region should include more speaker-related information. Moreover, vowel to vowel transitions which have a higher degree of freedom than other combinations include much more speaker related information than others. Therefore, vowel to vowel transitions seems to show the largest improvement in speaker recognition performance.

3. Conclusion

This letter describes the importance of transition regions in speaker identification tasks. We defined the transition frame as a frame straddling two or more phonemes in speech frames obtained from conventional short-time speech analysis processing and performed speaker identification tests for each phoneme class and the transition class. From various experiments, we found that transitions had the best speaker discriminative capability. Moreover, among the transitions, vowel to vowel transitions gave the best improvement.

The outcomes obtained in this letter can be used for designing efficient automatic speaker recognition systems. For example, we may enhance recognition accuracy by increasing the number of features extracted in transition regions and/or by assigning higher weights to the transition region relative to other regions. The analysis we have made in this letter can be further extended if we find a relationship between speaking rate and speaker recognition performance as well as the effects of adopting variable lengths of analysis frames at the feature extraction stage.

Acknowledgement

This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Biometrics Engineering Research Center (BERC) at Yonsei University (R112002105070040 (2009)).

References

[1] J. Eatock and J. Mason, "A quantitative assessment of the relative speaker discriminating properties of phonemes," Acoustics, Speech, and Signal Processing, 1994. ICASSP '94, 1994 IEEE International Conference on, vol.1, pp.I/133–I/136, 1994.

[2] R. Auckenthaler, E. Parris, and M. Carey, "Improving a GMM speaker verification system by phonetic weighting," Acoustics, Speech, and Signal Processing, 1999. ICASSP '99, Proceedings, 1999 IEEE International Conference on, vol.1, pp.313–316, 1999.

[3] J. Pelecanos, S. Slomka, and S. Sridharan, "Enhancing automatic speaker identification using phoneme clustering and frame based parameter and frame size selection," Signal Processing and Its Appli-

cations, 1999. ISSPA '99, Proc. Fifth International Symposium on, vol.2, pp.633–636, 1999.

[4] D. Gutman and Y. Bistritz, "Speaker verification using phoneme-adapted Gaussian mixture models," EUSIPCO-2002 the XI European Signal Processing Conference, pp.85–88, 2002.

[5] Y.L. Chow, R. Schwartz, S. Roucos, O. Kimball, P. Price, F. Kubala, M. Dunham, M. Krasner, and J. Makhoul, "The role of word-dependent coarticulatory effects in a phoneme-based speech recognition system," Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '86, ed. R. Schwartz, pp.1593–1596, 1986.

[6] B.E.F. Lindblom and J.E.F. Sundberg, "Acoustical consequences of lip, tongue, jaw, and larynx movement," J. Acoust. Soc. Am., vol.50, no.4B, pp.1166–1179, 1971.

[7] P. Mermelstein, "Articulatory model for the study of speech production," J. Acoust. Soc. Am., vol.53, no.4, pp.1070–1082, 1973.

[8] J. Louradour, K. Daoudi, R. Andre-Obrecht, and P. Sabatier, "Discriminative power of transient frames in speaker recognition," Acoustics, Speech, and Signal Processing, 2005. ICASSP '05, IEEE International Conference on, pp.613–616, 2005.

[9] J.S. Garofalo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, and N.L. Dahlgren, "The darpa timit acoustic-phonetic continuous speech corpus cdrom," Linguistic Data Consortium, 1993.

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

[11] K. Johnson, P. Ladefoged, and M. Lindau, "The degrees of freedom in controlling articulations," J. Acoust. Soc. Am., vol.89, no.4B, p.1870, 1991.

[12] K. Johnson, P. Ladefoged, and M. Lindau, "Individual differences in vowel production," J. Acoust. Soc. Am., vol.94, no.2, pp.701–714, 1993.