Osmangazi Üniversitesi Müh. Mim. Fak. Dergisi C.XIII, S.2, 2000
Eng. & Arch. Fac. Osmangazi University, Vol. XIII, No 2, 2000

CORRELOGRAM BASED FEATURE SELECTION FOR SPEAKER IDENTIFICATION USING VOWELS

Figen ERTAŞ[1]

ABSTRACT: *A correlogram approach to the selection of text-dependent features in vowel sounds is investigated for speaker identification. In the approach, vowel sounds as the identity carrying parts in spoken utterances are represented in the form of a correlogram, in which the speaker dependent spectral and temporal information is coded. Psycho-physiologically motivated spectro-temporal correlation with a search algorithm is introduced to identify the regions where the relevant features are embedded that are suited to discrimination. We identify the feature regions for a set of individual vowel sounds, and present results on their effectiveness in identifying speakers. Particular to the approach is that it makes no explicit use of any individual speech features.*

KEYWORDS: *Correlogram, Vowels, Auditory modeling, Speaker identification.*

SESLİ HARF KULLANARAK KONUŞMACI BELİRLEME İÇİN KORELOGRAM TABANLI ÖZELLİK SEÇİMİ

ÖZET: *Konuşmacı belirleme amacı ile, ünlü seslerdeki metne bağlı özelliklerin seçimi için bir korelogram yaklaşımı araştırılmıştır. Bu yaklaşımda, kimlik bilgisi taşıyan ünlü sesler, konuşmacıya ait spektrum ve zamana bağlı bilgilerin içinde kodlandığı bir korelogram şeklinde temsil edilmektedir. Ayrıma elverişli özelliklerin gösterim içinde saklı olduğu bölgeleri tespit etmek için ise, literatürdeki psikofizyolojik deney sonuçlarından hareketle frekans-zaman ilintisi ve buna ilişkin bir arama algoritması tanıtılmıştır. Özellik bölgeleri bir grup ünlü ses için tespit edilmiş ve bunların konuşmacıyı belirlemede ne kadar etkili olduğuna ilişkin sonuçlar verilmiştir. Bu makalede kullanılan yaklaşımın özelliği ise, hiçbir ses özelliğini doğrudan kullanmamasıdır.*

ANAHTAR KELİMELER:*Korelogram, Ünlü sesler, İşitsel modelleme, Konuşmacı belirleme*

---

[1] *Uludağ Üniversitesi, Mühendislik-Mimarlık Fakültesi, Elektronik Mühendisliği Bölümü, Görükle Kampüsü, 16059 BURSA*

I. INTRODUCTION

Speaker-identification (SI) is a multiple-choice identification task and has received a great deal of attention in the last two decades [1]. As the performance of a SI system depends on the discriminatory quality of the chosen features [2], selection and extraction of acoustic features that effectively characterize speakers is therefore of crucial importance. Unfortunately, no feature set is known so far to allow perfect discrimination. However, findings of speech research in the literature support that temporal information within speech signals appear to have a good potential to contain speaker-dependent (SD) cues [3], and is shown to be useful for SI [4]. The importance of temporal information in speech signals is emphasized in [5]. Auditory modeling is well known in the literature, and has been employed in various speech applications as a front-end processor, mostly outperforming conventional techniques. But, the use of auditory modeling for SI has not been much explored in the literature except a few works some of which are [6][7][8], where success was reported over spectral based conventional techniques. But, common to these is that the features and the way they are used inherently neglect the SD temporal information contained in speech signal. However, for a template based classification, an auditory model may be used in conjunction with autocorrelation analysis, resulting in a correlogram [5], to capture the SD temporal as well as the spectral attributes of speech signals without making explicit use of any speech features.

In this paper, we investigate a correlogram approach to feature selection for text-dependent SI using vowels, in which the SD spectro-temporal features are embedded as coded in the correlogram representation of spoken utterances. Selection of SD feature regions is explained, and their effectiveness in SI are presented for a set of individual vowel sounds.

II. CORRELOGRAM REPRESENTATION

The first stage of the model of the auditory periphery used in this paper consists of a bank of bandpass cochlear gammatone filters representing the frequency-selective basilar membrane motion [9], which separates the acoustic signal into a number of frequency bands. A nonlinear stage, as the second stage, follows the output of each

cochlear filter to simulate auditory nerve fibers, which transduce the mechanical motion of basilar membrane to synchronous neural firing patterns. A well-established mechanical to neural transduction, or inner hair cell transduction, is given in [10]. These firing patterns contain useful *spectral* and *temporal* information. An important step in a SI process is to find a representation of speaker voices from which sufficient information can be extracted that is suited to discrimination. Autocorrelation is a signal processing technique and acts as a process of determining the relationships between the contents of a signal within itself. Applying the autocorrelation to each auditory filter output channel by channel, a correlogram for a neural response is computed as

$$R_x(\tau, f) = \sum_{i=0}^{N} x(t, f)\, x(t - \tau, f) \tag{1}$$

where $t=iT$ in which T is the sampling period, f is the characteristic (center or channel) frequency for the auditory filter (equally spaced in ERB-rate, spanning the 50-5000 Hz frequency range), τ is the autocorrelation delay whose smallest value is the sampling interval T, and x is the probability of a spike in the auditory nerve [10]. To be more specific, $x(t, f)$ is the neural activity as a function of time in response to the output of a gammatone filter in the filterbank whose center frequency is f. In this way, by applying autocorrelation to the output of each filter after the mechanical to neural transduction process as in (1), SD information in one-dimensional speech signal about how and where sounds manifest in time-frequency plane is coded into a two-dimensional visual representation regarded as an *auditory pattern* (AP) in this paper, for which an example is shown in Figure 1. Note that the vertical axis in the picture is the center frequencies of the filterbank in [50-5000 Hz]. However, this representation does not explicitly provide the acoustic differences directly related to individual speakers. Information is rather embedded, and the differences can then be explored by further processing. Since the identification of speakers by their voice translates in this way to an auditory pattern recognition, now, the task is to find out how to extract or measure these features from the AP representation.

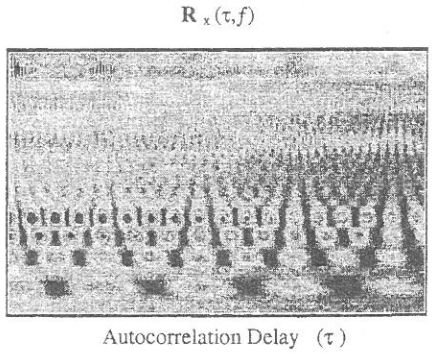

Figure 1. An example correlogram (AP).

III. FEATURE SELECTION

Since the SD within- and across-channel cues are preserved in the AP representation as spectro-temporally coded in the time-lag (τ) and in the channel frequency (f) variables, respectively, one does not need to explicitly deal with the embedded individual features themselves. Rather, as proposed in this paper, one can explore the regions where adequate cues about the invariant attributes of its speaker's voice for a given utterance is embedded as coded in τ-f plane. Since the APs have unique configurations as produced by individual speakers' vocal system, by identifying these regions, the embedded SD features can then be exploited through further processing to obtain the differences in speaker voices such that they can be used for discrimination.

There is a good evidence that listeners sometimes make comparisons across auditory filters (or across primary afferent fibers in an auditory nerve bundle), rather than listening through a single filter (or a fiber) [11]. This psycho-physiological evidence has motivated the exploration of the features of speaker identity in this paper by performing across-frequency comparisons in time-delay between an input speech transformed into a correlogram and the pre-stored correlogram templates, which are associated with speaker identities (through their utterances of a code sentence in the training session). This results in a *spectro-temporal* correlation (STC) of two correlograms, one for the input, $R_x(\tau, f)$, and the other for a reference template, $R_y(\tau, f)$, as

$$C_{XY}(\tau) = \frac{\sum_{i=1}^{M}[R_x(\tau,i) - \bar{x}(\tau)].[R_y(\tau,i) - \bar{y}(\tau,i)]}{\left[\sum_{i=1}^{M}[R_x(\tau,i) - \bar{x}(\tau)]^2 . \sum_{i=1}^{M}[R_y(\tau,i) - \bar{y}(\tau,i)]^2\right]^{1/2}} \quad (2)$$

where

$$\bar{x}(\tau) = \frac{1}{M}\sum_{i=1}^{M} R_x(\tau,i) \,, \qquad \bar{y}(\tau) = \frac{1}{M}\sum_{i=1}^{M} R_y(\tau,i) \quad (3)$$

Note here that (2) is the correlation coefficient as a function of τ for frequency slices taken from two APs at the same time-lag τ, hence the name *spectro-temporal*. We explore the features in two steps by performing *global* and *local* spectro-temporal correlation (STC) on APs by using (2) and (3), between the one for the input and the other for the reference. In the former one, it is performed on the whole AP as illustrated in Figure 2, which exploits the similarity of global features, while it is performed locally on a selected frequency band in the latter as shown in Figure 3, exploiting the similarity of local features that reveal the temporal variations across different frequency ranges of the speech signal.

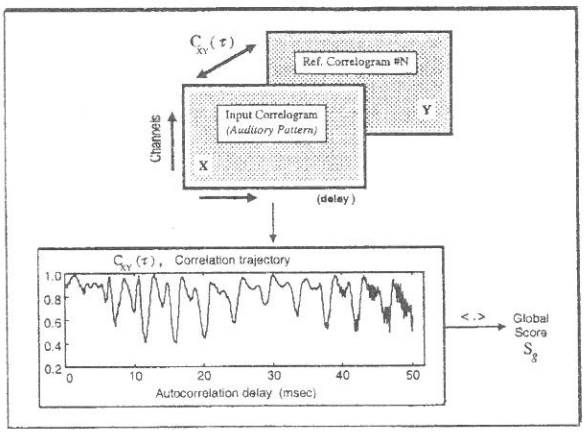

Figure 2. Spectro-temporal correlation on global features, a snapshot over 50 ms.

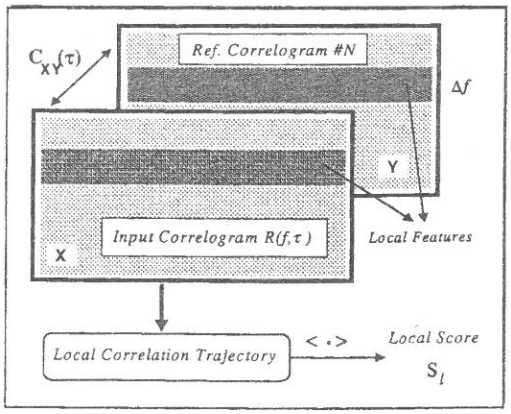

Figure 3. Correlation on local features.

The result of the STC is a spectral correlation trajectory. It is seen from Figure 2 and Figure 3, the measure of the similarity of APs is the global score S_g and the local score S_l, as the average of the global and the local correlation trajectories, respectively. Specifically, as a measure of the perception of SD similarity in two APs on a global scale, S_g is obtained by averaging the global STC trajectory as shown in Figure 2 with respect to a specific time-lag referred to as the best observation time. The best observation time is determined by forming the time varying average of the trajectory and returning the time-lag at which the average is maximum. This time-lag is also used for averaging the local STC trajectory to yield the S_l. The reason to look for a best observation time is to eliminate the speaker independent features, which are text-dependent. This is why the global and the local scores are taken as the average of STC trajectories. As anticipated, a local feature region can have any frequency channel span (Δf), which can be anywhere on the AP. Therefore its position on the AP should be identified, where a small intra-but a larger inter-speaker variations are observed. Namely, determining a channel span (or a time slice as shown in Figure 3) on the AP as a SD local feature region, one needs to search on the AP to find a slice that produce a good correlation with the speakers' own template but produce a weak correlation with the other speakers' templates by varying the width and the position of the slice. For an

input utterance, a search flowchart for the determination of the local feature regions is illustrated in Figure 4.

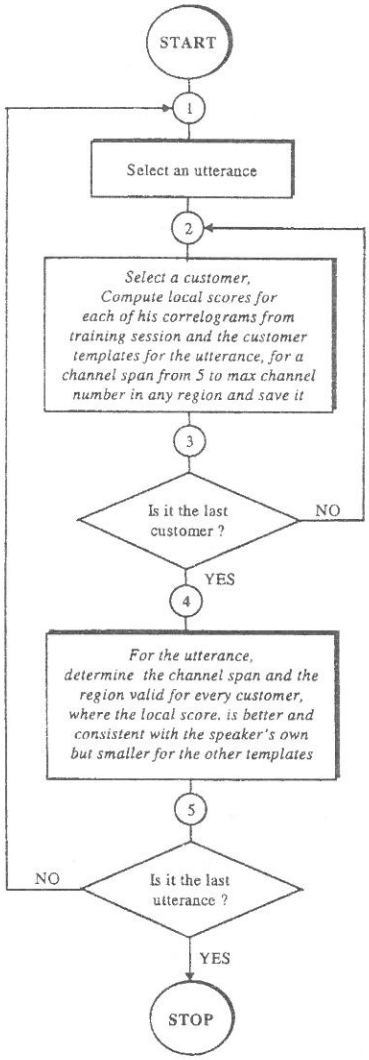

Figure 4. Flowchart for the search for local feature regions.

The experimental results on the recognition performance of vowels have revealed that the global similarity score is relatively sensitive to intra-speaker variance but can be compensated for by using local similarity score. The investigation has shown that

neither global nor local similarity on its own is consistently discriminatory, but rather they complement each other. For the discrimination of speakers, we measure the spectro-temporal acoustic differences in the utterances by using the local similarity score in combination with the global score as $S = S_g S_l$. This corresponds to AND logic, which imposes that the similarity must be satisfied in both global and local scale at the same time. If the similarity score is taken as the sum of the two scores as $S = S_g + S_l$, a misidentification may result since the sum of one lower and a larger incoherent score may well exceed a decision threshold, but their multiplication may be below the threshold leading to a correct identification. The sum logic has been tested and is found to produce poor results. Therefore, we use $S = S_g S_l$ as the final similarity score for discrimination.

IV. EXPERIMENTAL RESULTS

As reported to be the most effective speech sound that carry SD information [12], a closed-set SI experiment has been conducted with a speaker population of 10 customers (5 male and 5 female) by using only the vowel sounds extracted from their uttered sentences. The vowel sounds that are used in the identification experiment are /ɑ / in FATHER, /e/ in HATE, /i/ in EVE, /u/ in BOOT, /ɚ/ in BIRD, /æ/ in AT, and /ɔ/ in ALL. A set of correlogram templates with $\tau = 50$ ms was constructed for customers, and then local feature regions were determined for each vowel as illustrated in Figure 4, whose results are shown in Table 1 for a 32-channel auditory model. In the table, the frequency span of the regions are given in terms of auditory channel numbers (CHN-N) as well as center frequencies (CFs) of the channels. Then, a test was carried out to find the effectiveness of each of vowel in SI, by using a set of data collected also 4 months and 2 years after the training session, totaling 700 vowels (7x100).

Table 1. Local feature regions for a 32-channel system

	Local Feature Regions						
	V O W E L S						
	/ɑ/	/e/	/i/	/u/	/ɚ/	/æ/	/ɔ/
CHN-N	24-32	24-32	23-30	21-27	22-28	20-26	20-26
CF (*Hz*)	2225-5000	2225-5000	2004-4099	1619-3030	1802-3353	1452-2736	1452-2736

The result of the identification test using STC on global and local feature regions is shown in Table 2 for 16, 32, and 64-channel systems, in which N-CHN stands for number of channels (used in the auditory model). As an important aspect of the approach investigated, the sampling frequency for the recorded data was chosen 22050 Hz in order to exploit the temporal information in the fine structure of APs.

Table 2. Identification results

	Identification Results						
	V O W E L S						
N-CHN	/ɑ/	/e/	/i/	/u/	/ɚ/	/æ/	/ɔ/
16	80%	86%	90%	86%	93%	83%	63%
32	84%	93%	90%	93%	97%	83%	65%
64	80%	90%	90%	90%	93%	80%	70%

V. CONCLUSION

A correlogram approach to the selection of features is proposed for SI, in which no specific feature is in fact explicitly selected. The SD features are rather embedded in the APs on a local and global scale, and a procedure is given to identify relevant local feature regions suited to discrimination, where a small intra- but a larger inter-speaker variations are observed, by employing psycho-physiologically motivated spectro-temporal correlation to jointly exploit the embedded spectral and temporal information. Local feature regions for some individual vowel sounds are determined, and results on their effectiveness in speaker identification are also presented. It is seen from the results that for SI, /ɚ/ in BIRD is found to be the best vowel among the others, which agrees

with the result obtained in [13] by using the formant frequencies as features. It is observed from the results in Table 2 that increasing the number of channels from 16 to 32 improves the performance but further doubling it to 64 causes a loss in identification rate. This suggests that a further research should be performed to optimize the number of channels used in the filterbank that is suitable for SI. However, it is clear that increasing the number of channels does not always lead to a better performance. The reason for this is perhaps the inclusion of more speaker independent information in the AP and somewhat smearing out the speaker-dependent information.

In this paper, identification results are presented only for individual vowels. Later, we will present an identification strategy using a set of vowels, which explores the differences in APs by using a robust decision mechanism.

REFERENCES

[1] J. P. Campbell, "Speaker recognition: A tutorial", *IEEE Proceedings*, 85(9), pp 1437-1462, 1997.

[2] M. R. Sambur, "Selection of acoustic features for speaker identification", *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP-23(2), pp 176-182, 1975.

[3] C. C. Johnson, H. Hollien and J. W. Hicks, "Speaker identification utilizing selected temporal features", *J. Phonetics*, 12, pp 19-326, 1984.

[4] F. K. Soong and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition", *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP-36(6), pp 871-879, 1988.

[5] M. Slaney and R. F. Lyon, "On the importance of time – a temporal representation of sound", in *Visual Representations of Speech Signals*, by Martin Cooke, Steve Beet and Malcolm Crawford (eds.), Wiley, 1993, pp 279-284.

[6] J. M. Colombi, T. R. Anderson, S. K. Rogers, D. W. Ruck and G. T. Warhola, "Auditory model representation and comparison for speaker recognition", IEEE Proc. Int. Conf. On Neural Networks, 1993, pp 1914-1919.

[7] T. R. Anderson and R. D. Patterson, "Speaker recognition with the auditory image model and self organizing feature maps: A comparison with traditional

techniques", ESCA Workshop on Automatic Speaker Recognition, Martigny April 5-7, 1994, pp 153-156.

[8] X. Jiang, Z. Gong, F. Sun and H. Chi, "A speaker recognition system based on auditory model", World Congress on Neural Network, Int. Neural Network Society Annual Meeting, San Diego, 1994, (4), Ch. 128, D595-D600.

[9] F. Ertaş, "Ses sinyallerine karşı basilar membran hareketinin benzetimi", Elektrik-Elektronik-Bilgisayar Müh. 8. Ulusal Kongresi, Gaziantep, 1999, pp. 618-621,

[10] R. Meddis, "Simulation of auditory-neural transduction: Further studies", *Journ. of Acous. Soc. of America*, JASA 83(3), pp 1056-1063, 1988.

[11] B. C. J. Moore, "*An introduction to the psychology of hearing*", Academic Press, 1989.

[12] B. S. Atal, "Automatic recognition of speakers from their voices", *IEEE Proceedings*, 64(4), pp 460-475, 1976.

[13] K. K. Paliwal, "Effectiveness of different vowel sounds in automatic speaker identification", *J. Phonetics*, 12, pp 17-21, 1984.