Available online at www.sciencedirect.com
SCIENCE DIRECT®
Physica D 200 (2005) 75–80
PHYSICA D
www.elsevier.com/locate/physd

Topological voiceprints for speaker identification

M.A. Trevisan[a], M.C. Eguia[b], G.B. Mindlin[a,*]

[a] *Departamento de Física, FCEyN, Universidad de Buenos Aires Ciudad Universitaria, Pab. I C1428EGA Buenos Aires, Argentina*
[b] *Centro de Estudios e Investigaciones, Universidad Nacional de Quilmes, Roque Sáenz Peña 180, Bernal, B1876BXD Buenos Aires, Argentina*

Received 7 June 2004; received in revised form 27 September 2004; accepted 30 September 2004

Communicated by C.K.R.T. Jones

Abstract

Despite its noninvasive nature, subject identification by voice is not as popular as other biometric procedures (i.e. fingerprinting). In part, this is due to the difficulty of establishing how close is close enough when comparing spectral features. In this work, we address this issue by showing how to characterize spectra by means of sets of integers, borrowing topological tools used in the theory of dynamical systems. On the other hand, we report an empirical result: within a relatively small bank of speakers, there are subsets of integers that seem to strenghten the speakers' identity information. These results suggest a new direction in the identification of subjects by voice: one in which arrangements of integers define voiceprints that stand on their own, despite any acceptance/rejection thresholds.

Keywords: Speaker recognition; Biometrics; Topological indexes

1. Introduction

Many techniques used in the problem of voice identification are based on spectral properties [1]. The reason is that during the production of voiced sounds, a spectrally rich sound signal (produced by the modulation of airflow by the vocal folds) is filtered by the vocal tract. The resonances of this passive filter are determined by ergonomic features of the speaker, and therefore can be used to identify him/her.

The purpose of this work is twofold. On one hand, we borrow from the theory of dynamical systems a set of tools of topological nature that will allow us to characterize the spectra of voiced sounds. On the other hand, we report an empirical result: with the use of these topological indexes we can define a voiceprint, i.e. a set of integers uniquely associated (within the extent of our exploration) with a speaker.

The production of human voice involves a wide variety of physical processes, from turbulence (typically present in non-voiced sounds) to low dimensional dynamics (like in voiced sounds). During the production of voiced sounds (like vowels), the airflow induces periodic oscillations in the vocal folds. These oscillations generate time varying pressure fluctuations at the input of a passive linear filter, the vocal tract. The physics of voiced sounds is usually described in terms of the

* Correspoding author. Tel.: +54 11 4576 3390.
E-mail address: gabo@df.uba.ar (G.B. Mindlin).

0167-2789/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.physd.2004.09.008

standard *source-filter* theory [1–3]. The separation between source and filter assumes that the feedback into fold oscillations is negligible, an hypothesis that has been extensively validated for normal speech regime [4]. The spectrally rich input pressure presents harmonics of a fundamental frequency of about 100 Hz. The vocal tract selects some frequencies out of these harmonics. In this way, the spectrum of a voiced sound carries information about the vocal tract and therefore it is used as a biometric characterization of the speaker.

A typical approach in the field is to use feature vectors with quantities that characterize different subjects, perform multidimensional clustering and separate the clusters associated with the different subjects by means of some metric on the feature vectors. In the framework of the spectral characterization of the voice, a way to perform an identity validation is to construct a distance between properties computed from utterances (distorsion measures), such as the integral of the difference between the two spectra on a log magnitude. Another distorsion measure is based upon the differences between the spectral slopes [1] (first-order derivatives of the log power spectra pair with respect to frequency).

It is interesting to notice that the spectral comparison is sometimes implemented by means of another set of coefficients called *cepstrum* that are just the Fourier amplitudes of the spectral function. In other words, the spectrum is treated as a "time" series where f plays the role of time. This suggested to us that the techniques used in the theory of dynamical systems in order to compare two periodic orbits can be used in the analysis of voiced sound spectra. In particular, we are interested in exploring the use of topological tools, designed to capture the main morphological features of orbits regardless of slight deformations [8].

In the analysis of three-dimensional dynamical systems, the periodic orbits are closed curves that can be characterized by the way in which they are knotted and linked to each other and to themselves [5–7]. Since uniqueness theorem warranties that periodic orbits cannot cross, linking and knotting properties cannot change and threrefore can be used as fingerprints. For the purpose of applying this analysis to the problem of speaker identification we will treat the power spectrum of voiced sounds $|H(f)|^2$ on a log scale as a periodic string of data, using techniques commonly applied to the analysis of periodic "time" series. A three-dimensional orbit can be constructed from this string

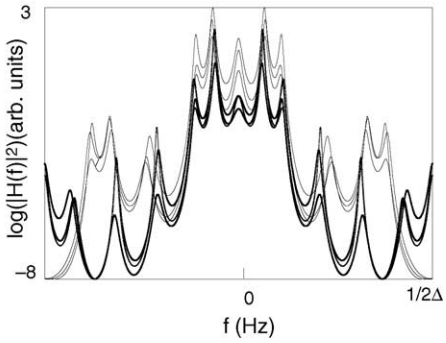

Fig. 1. Three examples of $\log(|H(f)|^2)$ using the maximum entropy approximation for two different speakers (using light and dark lines for each of them) over the complete period of the function. Beyond the second *formant*, the spectra naturally cluster in two different groups. The original sound segments correspond to the spanish vowel [*a*] extracted from normal speech utterances. Sampling rate is $r = 1/\Delta = 11,025$ Hz.

of data using a delay embedding [8,9]. In Fig. 1 we show the log power spectra of three vocalizations of two speakers. Notice that these spectra naturally cluster in two sets. Could the topological properties of their embeddings be a pertinent tool for identity validation?

In order to answer that question, we will deal with *relative rotation rates*, topological invariants introduced to help in the description of periodically driven two-dimensional dynamical systems [5], which can also be constructed for a large class of autonomous dynamical systems in R^3: those for which a Poincaré section can be found.

2. Topological voiceprint implementation

In order to describe the vocal tract frequency response, we computed the maximum entropy approximation of the power spectrum for each of the stored voiced segments. This can be performed by calculating *m linear predictor* coefficients [10] for the voiced segment $\{y_n\}$, sampled with rate $r = 1/\Delta$:

$$y_n = \sum_{k=1}^{m} d_k y_{n-k} + x_n, \tag{1}$$

where the *lp* coefficients $d_1, d_2, \ldots, d_m$ are assumed constant over the speech segment, and are chosen so that x_n is minimum. These *lp* coefficients can be used to estimate the power spectrum $|H(f)|^2$ as a rational

function with m poles:

$$H(f) = \frac{d_0}{1 - \sum_{k=1}^{m} d_k \, \mathrm{e}^{ik2\pi f\Delta}} \tag{2}$$

which is periodic in $[-1/2\Delta, 1/2\Delta]$, the Nyquist interval. Several examples of reconstructed spectra are shown in Fig. 1 for two male speakers of the same age.

The log of power spectral function $\log|H(f)|^2$ was approximated using Eq. (2) with $m = 13$ coefficients. This value results from the optimization of our speaker recognition implementation over a database of 108 utterances. With this number of poles, the reconstructed spectra capture the main features of the formant's envelope. The fluctuations in the spectra of the same speaker and the same voiced sound (Fig. 1) correspond to variations arising from normal speech.

The function $|H(f)|$ is symmetric with respect to $f = 0$. Therefore, only one half is relevant to our analysis. We washed out the difference between $\log|H(0)|$ and $\log|H(1/(2\Delta))|$, adding a linear function and substracting the average. The final function $F(f)$ is periodic with half the original period. In Fig. 2, a few examples of $F(f)$ for different utterances of the same speaker are shown, along with a *reference* function.

The resulting function $F(f)$ can be embedded in phase space using many different techniques [11]. We used a δ delay embedding. An example of such an orbit using $\delta = 40$ Hz is displayed in Fig. 3.

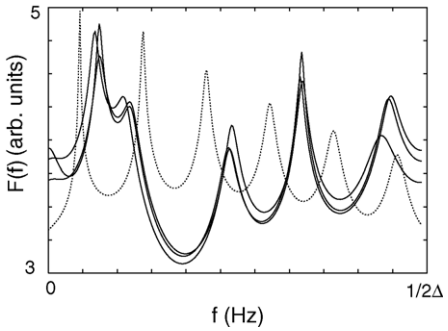

Fig. 3. Periodic functions used for the embedding from a single speaker (solid lines) and a universal reference (dotted line). These functions are constructed from the original $\log(|H(f)|^2)$ as in (Fig. 1), using half of the original period.

These delay-embedded orbits in phase space $(F(f), \; F(f-\delta), \; F(f-2\delta))$ always display a hole around the line $F(f) = F(f-\delta) = F(f-2\delta)$. Therefore, a good Poincaré section is given by the semi-plane defined by $F(f) = F(f-2\delta)$; $F(f-\delta) \leq F(f-2\delta)$.

A family of topological tools are available in order to classify and characterize these orbits. We explored both the self relative rotation rates and the relative rotation rates with respect to a reference orbit. The self relative

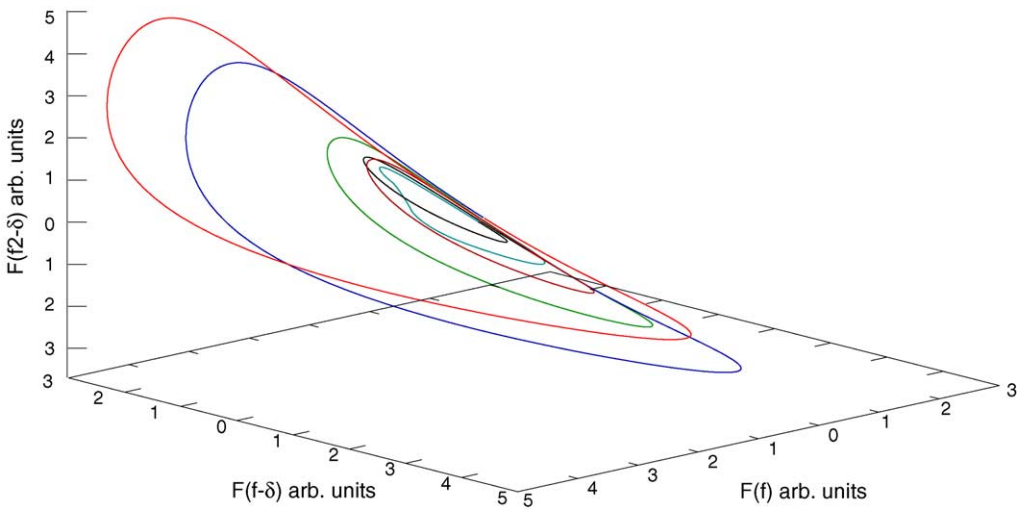

Fig. 2. Delay embedding ($\delta = 40$ Hz) of the function $F(f)$ computed from one voiced fragment. Different line types represent different segments of the orbit between succesive crossings of the Poincaré section.

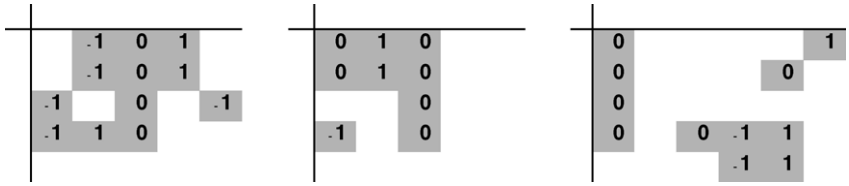

Fig. 4. Vowelprints for three male speakers of nearly the same age, constructed from short vowel segments (∼100 ms) of around 10 utterances taken in different enrollment sessions.

rotation rates showed sensitivity to natural fluctuations of speech spectra. Since we were computing relative rotation rates of an orbit relative to itself, slight variations in the height of some formants were reflected as different topological indexes. On the other hand, relative rotation matrix empirically proved to contain more robust numbers if computed against a constructed reference orbit. This reference was derived from the formants of a plain, non-articulated vocal tract (a *zero hypothesis* for voiced sounds). Our election of the reference as an open-closed tube of 17.5 cm is based in the observation that the geometry of a real vocal tract can be thought as departures from an open-closed tube.

In the following, we will refer to the number of the intersections of an orbit with the Poincaré section as its period. Each section of the orbit in between successive intersections will be called segment. The relative rotation of these embedded spectra can be calulated in the following way: assume that the orbits have periods p_A and p_B. We build a *relative rotation matrix* $M \in Z^{p_A \times p_B}$, where element M_{ij} corresponds to summing the signed crossings of the ith segment of the orbit A relative to the jth segment of the orbit B. The signed crossings can be calculated projecting the two orbits onto a two-dimensional subspace. In that projection, tangent vectors to the two periods just over the cross are drawn in the direction of the flow. The upper tangent vector is rotated into the lower tangent vector, assigning a $+1\,(-1)$ to the crossing if the rotation is right (left) handed.

This relative rotation matrix is related to the relative rotation rates through:

$$R_{ij}(A, B) = \frac{1}{p_A p_B} \sum_{k=0}^{p_A p_B - 1} M_{i+k, j+k} \tag{3}$$

taking periodic boundary conditions for the matrix. Relative rotation rates (Eq. (3)) are averages over the matrix elements M_{ij} until the same initial conditions (i, j) are reached. The elements of matrix M, on the other hand, allow us to keep track of the rotation of each pair of segments, and therefore can be used to identify spectral variations for each pair of formants of the utterance i. In order to construct a signature of the speaker, we begin characterizing each of his/her vowels. We do so superposing all the relative rotation matrices corresponding to the same voiced sound and the same speaker, and looking for coincidences (rotation numbers which do not change when computed from different utterances). These coincidences are named *robust rotation numbers*. We called *vowelprint* the arrangement of the robust rotation numbers placed in the original matrix sites. We called a *voiceprint* the collection of the speakers' *vowelprints*. In Fig. 4, we display three vowelprints corresponding to the spanish vowel [*a*] for three male subjects of nearly the same age.

3. Results and discussion

In order to test our topological approach to the problem of speaker recognition, we performed the following implementation. A voice bank was constructed as follows. We recorded six repetitions of a sentence containing the five Spanish vowels for each one of 18 speakers. Then, we build topological matrices from short fragments (∼100 ms) taken from those vowels. The bank consisted of the *voiceprints* computed from the topological matrices for each speaker.

Topological matrices computed from an utterance by a speaker who claims to be in the bank are computed. These candidate matrices are compared with the corresponding vowelprints in the bank. The speaker is identified as a member of the bank only if the set of candidate matrices fully matches a single stored voiceprint. In this context, full matching means that all the robust numbers in all the vowelprints are present in the corre-

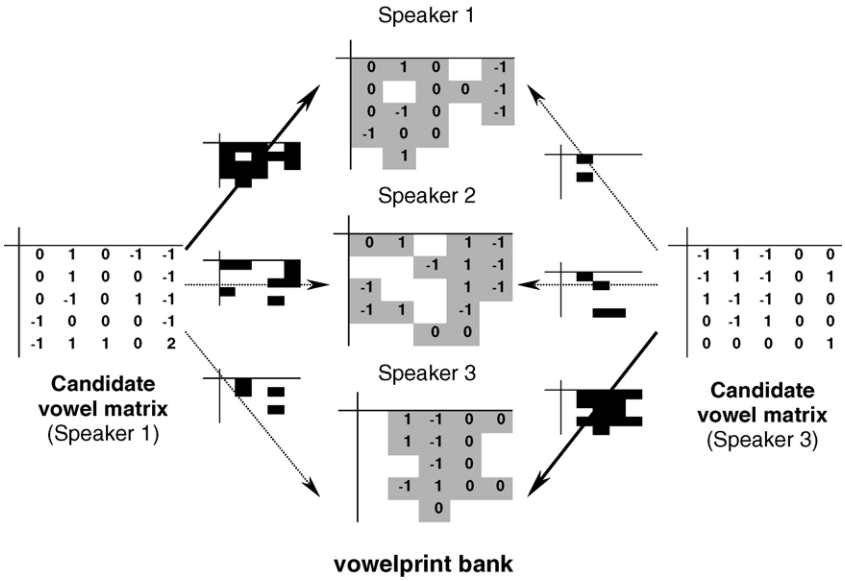

Fig. 5. Sketch of the recognition procedure: two candidates (we display a single *vowel print* for each of them) are compared with the bank of molds. A speaker is identified as a member of the bank if the set of his/her candidate matrices fully matches a single stored voiceprint. The grey areas in the molds correspond to positions in the matrices that contain robust numbers. Identification of a candidate as a member of the bank (i.e., full matching) requires the numbers in those positions of the candidate's matrix being equal to the robust numbers in the mold.

sponding candidate matrices. This process is illustrated in Fig. 5, for a single vowelprint. In turns, each of the 108 utterances of our bank was used as a candidate for identification. We obtained perfect recognition performance (no false positive/negatives).

Our choice of a subset of the rotation numbers in the construction of a voiceprint could suggest that some information is lost. In order to test this hypothesis, we replaced each *voiceprint* in the bank with the collection of the complete individual matrices used to construct them, in such a way that all the topological information is kept. Each of the 108 utterances of our bank was used as a candidate for identification. We evaluated the number of coincidences between the candidate matrices and the set of matrices characterizing each speaker in the bank. The result was a lower performance method, since several false positives and negatives were found. The topological robust numbers seem to strenghten the relevant spectral information, discarding the unuseful information carried by the indexes that vary the most from one utterance to the next.

We also compared our topological approach with a naïve metric method, in which the quadratic distance between spectra is calculated and coincidences are computed below an optimized threshold. In this case, the *voiceprint* of each speaker in the bank was replaced by the spectral functions used to construct the rotation matrices. Again, the performance of this metric method as a speaker recognizer was worse than the topologic one.

Our approach presents many interesting advantages over the usual ones. In a metric strategy (in which some distance between spectra are computed), a threshold has to be defined, and this is a bank dependent quantity. The use of topological voiceprints constructed with integers, along with the full-matching criterion, introduces a novel strategy, which is bank-independent, with no-threshold needed to verify the acceptance.

Implementations of the topological approach running on standard PCs also showed to be very fast. Once an utterance is recorded, voiced sounds segments can easily be extracted. Their relative rotation matrices can be built using simple cross-counting algorithms [8] and voiceprints are then computed by simply counting coincidences over a collection of small matrices. On the other hand, once the bank is constructed, the whole recognition task consists in the matching of small matrices like the ones showed in Fig. 4.

We explored the change in the number of robust numbers as a function of the training set size. We found that for training sets larger than 10 vowels, the number of robust numbers converge to approximately 8. These numbers describe the relative heights of the peaks of the spectral function of a voiced sound with respect to the spectrum of a reference, that do not change from utterance to utterance. The robust numbers of a subject in our base were compared with the topological indexes obtained from an utterance recorded when the subject had a strong cold. We found that the information in the matrix of robust numbers degrades gracefully: only the indexes associated with the highest frequencies changed, while a large part of the voice print remained unaltered.

In summary, we propose a novel approach to the problem of subject identification by voice based on topological indexes. The strategy is inspired in a successful program of characterization of recurrences in dynamical systems [8]. We found empirically that a set of indexes can be used to characterize subjects, since they do not change from utterance to utterance and are different from subject to subject within the bank explored. Larger studies are required to compare the performance of this approach to more mature techniques, but our strategy allow us to advance towards a threshold independent solution of the problem of subject identification by voice.

Acknowledgments

This work was partially funded by UBA, CONICET and ANCyT, Argentina.

References

[1] L. Rabiner, B. Juang, Fundamentals of Speech Recognition, Prentice-Hall, 1993, pp. 24–256.
[2] I. Titze, The Physics of Voice Production, Allyn and Bacon, 1994.
[3] K. Ishizaka, J.L. Flanagan, Bell Syst. Techn. J. 51 (1972) 1233.
[4] R. Laje, T. Gardner, G.B. Mindlin, Phys. Rev. E 64 (2001) 056201.
[5] H.G. Solari, R. Gilmore, Phys. Rev. A 37 (1988) 3096.
[6] G.B. Mindlin, X. Hou, H. Solari, R. Gilmore, N.B. Tufillaro, Phys. Rev. Lett. 64 (1990) 2350.
[7] G.B. Mindlin, Gilmore, R., Physica D 58 (1992) 229.
[8] R. Gilmore, Topological analysis of chaotic dynamical systems, Rev. Mod. Phys. 70 (1998) 1455.
[9] R. Gilmore, M. Lefranc, The Topology of Chaos, Wiley, 2002, pp. 131–160.
[10] H.W. Press, et al., Numerical Recipes in C: The Art of Scientific Computing, Cambridge University, 1999, pp. 564–574.
[11] M. Lefranc, P. Glorieux, Int. J. Bifurcation Chaos 3 (1993) 643.