Functional Dissipation for Speaker Recognition

D. Napoletani[1,2], L. Romito[3], D.C.Struppa[1]
[1] Schmid College of Science, Chapman University, USA;
[2] Center for Applied Proteomics and Molecular Medicine, George Mason University, USA;
[3] Phonetics Laboratory, University of Calabria, Italy.

ABSTRACT

We describe a new type of functional dissipation, that proves to be effective for speaker recognition. Functional dissipation is based on standard signal transforms, but uses the transforms recursively to uncover new features. We generate a variety of random masking functions and extract features with several signal transform iterations. In each iteration of the recursive process, we modify several coefficients of the transformed signal with the largest absolute values according to the specific masking function, and we collect basic statistics of the distribution of the coefficients of the modified transformed signal. We show that the statistics generated by the recursive functional dissipation are able to distinguish sentences uttered by different speakers, even when the sentences are noisy and undersampled, providing a new approach to speaker recognition.

BACKGROUND

SPEAKER RECOGNITION.
The voice is more than simply a sum of sound. It is intrinsically articulated and its complexity is due to the relationship between meaning, intentions emotion, own health, social status, own self-esteem etc. [Laver, 1994]. Given the richness and variability of features embedded into a single acoustic signal, it is not surprising that the problem of robustly recognizing a speaker from spontaneous utterances is still an unsolved problem [Furui, 2005]. One particular type of speaker recognition is the problem of matching an input speech to one of several reference speakers, what goes under the name of speaker identification.

Speaker identification problems are particularly hard when speeches are undersampled and noisy, and yet these are precisely the conditions found in a variety of forensic and commercial applications. These limitations, and the tremendous variability of in-class signals, make speaker identification an ideal setting to test the limits of classification methods

FUNCTIONAL DISSIPATION.
We generalize in this project functional dissipation, a feature enhancement technique that we developed in [Napoletani et al., 2007] to classify texture in images, to approach the problem of recognizing different speakers on the basis of sentences that they utter.
Functional dissipation is based on one of the simplest and yet most successful ways to analyze signals, i.e the process of recursive extraction of features from the signal itself. The general implementation of this process, on which functional dissipation is based, is the following.

Given an initial approximation G=0 of F, and an initial residual signal R=F-G.

- We expand R in some dictionary D={g_1, ..,g_P} of simpler signals.

- We select the element g_k in the dictionary for which the norm of inner product |<R,g_k>| is maximum.

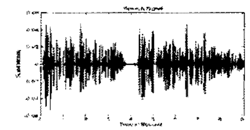
- Set G=G+<R,g_k>g_k and R=R-<f,g_k>g_k.

The process is repeated on the residual signal several times to extract successively different relevant structures from the signal.

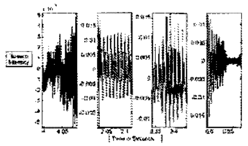
Functional dissipation makes a simple, but fundamental change to the last step of this process: the largest coefficients in the dictionary are modified, before building the residual signal. In this way, this standard approximation process is turned into a slow, non-linear dissipation of the structure of the signal, with no approximation purposes.

OVERALL PROCESSING OF INDIVIDUAL SPEECHES

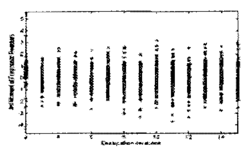
1. Given a speech signal



2. Extract many short speech fragments.



3. Perform several functional dissipation iterations on all speech fragments (see FUNCTIONAL DISSIPATION box) Compute 3[rd] and 4[th] statistical moments of each fragment residue, for each iteration.



4. Classify the given speech by comparing its distribution of statistical moments, for all dissipation iterations, with the corresponding distributions of moments of a set of training speeches for each class of speakers (see CLASSIFICATION box).

NOTE
- For each iteration, moments are normalized so that the aggregates of moments from the fragments of all available training speeches have mean 0 and variance1.
- Residues degrade to noise after ~20 iterations Differences among speeches arise in the slow process of degradation of information in the initial iterations.

References and Patents:
[Laver, 1994] Laver J., Principles of Phonetics, Cambridge University Press, 1994

[Furui, 2005] Furui S., 50 years of progress in speech and speaker recognition, Proc. SPECOM 2005 Patras, Greece, pp 1-9 (2005)

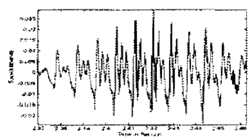
[Napoletani et al., 2007] D. Napoletani, D. C. Struppa, T. Sauer, V. Morozov, N. Vsevolodov and C. Bailey, Functional Dissipation Microarrays for Classification, Pattern Recognition 40(12): 3393-3400 (2007)

D. Napoletani, D. C. Struppa, T. Sauer, United States Patent 7,724,933 Issued May 25, 2010 Functional dissipation classification of retinal images

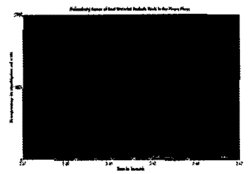
D. Napoletani, D. C. Struppa, T. Sauer, United States Patent 7,650,024 Issued January 19, 2010 Dissipative functional microarrays for classification

FUNCTIONAL DISSIPATION

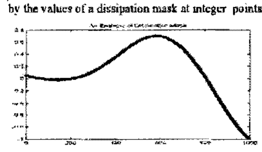
(a) Given a speech fragment



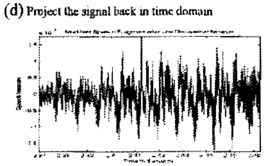
(b) Find its best basis coefficients in a wavelet packets basis (to detect spikes) or in a cosine packets basis (to detect localized frequency information)



(c) Multiply the 1000 wavelet packets coefficients with largest norm by the values of a dissipation mask at integer points.



(d) Project the signal back in time domain



(e) Normalize the modified fragment to have norm 1 and repeat steps (a) to (e)

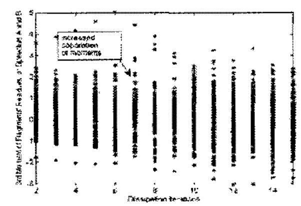
NOTE

- Dissipation masks are created by filtering Gaussian white noise with low-pass filters of variable frequency support
- Wavelet packets basis and cosine packets basis are alternated in the iterated dissipation process, optimizing each time the basis to the signal

CLASSIFICATION

For each dissipation mask and dissipation iteration, we can compare the distributions of statistical moments of fragments of any two speeches A and B

In some cases, when speech A and speech B are from different speakers, the two sets of moments will not fully overlap



For a given mask I and dissipation iteration J, we compute the spread associated to the pair of speeches A and B. Let M(A,I,J) and M(B,I,J) be the means of the fragments' moments, and S(A,I,J), S(B,I,J) their respective standard deviations, then the spread is defined as

spread(I,J)=|M(A,I,J)-M(B,I,J)|/(S(A,I,J)+S(B,I,J))

Let S(A,B) be the maximum spread across all masks and dissipation iterations. We define S(A,B) as the pseudo-distance of A and B.

We can use K-Nearest Neighbor classification methods to classify an input training speech according to its pseudo- distance to the K closest speeches in a set of training speeches from all speakers (we use K=5).

RESULTS

SETTING.
- 40 training speeches for each of 4 speakers classes, split randomly into several sets of 35 training speeches and 5 testing speeches for each speaker.
- Gaussian white noise added to the signals, with standard deviation equal to 5% of their mean
- Each speech instance is a repetition of the same phrase. Note however that the method is not in principle text-dependent.
- 100 fragments for each speech. 50 dissipation masks. 15 dissipation iterations.

ERROR RATES