MULTI-RESOLUTION FOR SPEECH ANALYSIS

Marie-José Caraty, Claude Montacié

LIP6 - Université Pierre et Marie Curie
4, place Jussieu - 75252 Paris cedex 05 - France

ABSTRACT

In the purpose to deal with artifact on observations measurements resulting from usual speech processing, we propose to extend the representation of the speech signal by taking a sequence of sets of observations instead of a simple sequence of observations. A set of observations is computed from temporal Multi-Resolution (MR) analysis. This method is designed to be adapted to any usual mode and technique of analysis. Its originality is to take into account two main variations in the analysis, -the center of the frame and -the duration of the frame. In speech processing, multi-resolution analysis has many applications. MR analysis is a basic representation -to locate the stationary and non-stationary parts of speech from the inertia computation, -to select the best representative observation from centroid or generalized centroid.

Preliminary experiments are presented. The first one consists in the MR analysis of pieces of the French and the English-American speech databases (i.e., TIMIT, BREF80) and on the inertia as a criterion of location of stationary and non-stationary parts of the speech signal. The second one is on the computation of the phoneme prototypes of the two speech databases. At last, some perspectives are discussed.

1. INTRODUCTION

Speech is known to be essentially a non-stationary signal. Production of speech involves articulatory configuration changes on the order of the phonemic rate estimated about 10 phonemes per second. Because of these changes, the short-time Fourier transform is one of the major speech measurements. To be meaningful, this measurement is taken over a finite time interval (i.e., frame) for which the articulatory apparatus may be considered motionless. The duration of the signal frame is generally chosen from experience and typically spreads from 20 to 40 msec. The speech continuum is usually represented by a sequence of observations (i.e., short-time spectral vectors). The observations are computed from pitch-asynchronous analysis typically measured once every 10 msec over a given short-time duration frame. Because of its computation cost, less usual is the pitch-synchronous analysis. To our knowledge, a variable frame duration is an alternative unused for asynchronous or synchronous analysis.

In the purpose to deal with artifact on the observations coming from these various types of speech processing used for the measurements, we propose to extend the representation of speech signal by taking a set of vectors of observation instead of a single vector of observation. This set of observation is computed from multi-resolution analysis [1]. This method is designed to be adapted to any analysis mode (i.e., asynchronous/synchronous) and to any spectral vector measurement obtained via standard methods as the Fast Fourier Transform (FFT), the Auto-Regressive models (AR), the Auto-Regressive-Moving Average (ARMA) models. For speech processing, the multi-resolution analysis is a well fitted representation to enhance the following principles of decision :

- Stationary parts or non-stationary parts of the speech signal.
- Best representative observation, in terms of stationary observation, at a located point or a located part of the speech signal.

2. MULTI-RESOLUTION ANALYSIS

The multi-resolution analysis we propose is available whatever the technique of analysis is (e.g., LPCC, MFCC, formants), and adaptable to usual synchronous or asynchronous modes.

2.1. Multi-Resolution Principle

We designed the multi-resolution analysis so as to retain information on the local stationarity of the signal. This information is simply kept through the representation of the signal by a sequence of sets of observation instead of a sequence of observations. The set of analysis results from several analysis in the neighborhood of a considered point of the signal, each analysis being carried out on various frame durations.

Let A(s, d) be the analysis of a frame of a duration (d) centered on the speech sample (s). The multi-resolution analysis of a speech continuum located at sample (t) consists in a set of short-time frames analysis {A(s, d)} (s (resp., d) taking a value in the set S (resp., D)) such as the center (s) and the duration (d) of the analyzed frames vary. An example of values is as follows for a 10 msec rate pitch-asynchronous multi-resolution analysis of a speech continuum located at the sample (t).

$$S = \{ t + i * \delta \} \quad i=(-10, ..., 10)$$

$$D = (20 + i * \delta \} \quad i=(0, ..., 20) \quad \text{with } \delta = 0,5 \text{ msec}$$

441 observations are computed like this. The values of the parameters (s, d) are chosen to have an homogeneous processing of the speech continuum. These values aim at the representation of the local stationarity in a neighborhood spreading up to 40 msec. For vowels for instance, taking into account the phonemic rate and assuming an equal distribution of the coarticulation effects (i.e., beginning of transition, steady state, end of transition), we may expect a local stationary part of about 33 msec (0,1/3 msec).

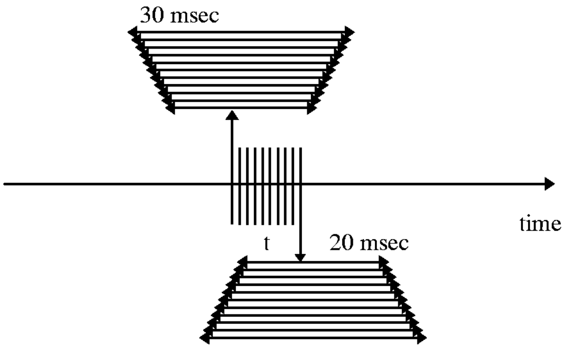

Figure 1: Multi-resolution analysis (one by two) at the located point t of the speech signal. 21 measurements are computed in the neighborhood of through and through t with a 0,5 msec rate. For each one of these centers of MR analysis, the measurement is computed for 21 frame durations spreading from 20 msec to 30 msec with a 0,5 msec rate.

2.2. Centroid and Generalized Centroid

The stability measures the adequacy between the considered application (e.g., speech/speaker recognition, modeling, synthesis), the technique of speech analysis (e.g., LPCC, MFCC, formants) and the analyzed speech continuum. The stability can be quantified by a measure spreading on a scale from non-stability to stability. In case of non-stability (respectively, stability), the continuum can't (respectively, can) be represented by an (respectively, any) observation of the set of observations. In most cases, the measure of stability is intermediary and techniques such as centroid or generalized centroid can be applied to obtain the best prototype of the MR set of observations. The automatic selection of the best representative observations of the signal is of a major interest in the principles of decision involved in speech recognition.

3. EXPERIMENTS

Preliminary experiments using multi-resolution analysis are carried out on the French [2] and the English-American [3] speech databases (BREF80, TIMIT). Formant parameter space is chosen, a specific distortion measure (Clustering of Spectral Peaks measure) is used. The first experiment consists in the MR analysis and the computation, from the MR set of observations, of the inertia and of the best representative observation. Location of stationary and non-stationary parts of the speech is discussed. The analysis of the results brings to a discussion on the accuracy of the speech database labeling.

3.1. Parameter Space and Distortion Measure

Formants are chosen to represent the speech signal, this analysis is well known to be very sensible in its measurement. An advantage of formant parameters lies on the amount of knowledge we have on the characterization, the discrimination and the perception of the vowels. Moreover, formants are physically interpretable. The spectral representation is based, by analogy with formants, on the characteristics of the spectral maxima/peaks detected on the LPC spectrum. In this representation space, the inter-spectra distortion measure used is the Clustering of Spectral Peaks (CSP) measure we designed [4]. This measure is based on perceptive criteria of the sound perception.

Spectral Peaks Parameter Space

A short-time spectrum S is represented by the set of its spectral peaks (i.e., local maxima) $\{P_k\}_{(k=1,..,K)}$ characterized by their central frequency f_k (Hz), their 3 dB bandwith b_k (Hz) and their magnitude m_k (dB) :

$$S = \{P_k(f_k, b_k, m_k)\}_{(k=1,..,K)}$$

The signal is analyzed by a 16th order linear prediction. The peak parameters are computed from the LPC spectrum.

Clustering of Spectral Peaks Distortion Measure

The original Clustering of Spectral Peaks measure has proved to have many advantages [5]. Its computation is summarized as follows.

A local inter-peaks distortion measure δ (P_i, P_j), between two peaks of distinct spectra, is introduced for the measurement of a local spectral distortion and computed with the following formula :

$$\delta\,(P_i,P_j) = \omega_f(f_i).e_f(P_i,P_j)+\omega_b(f_i).e_b(P_i,P_j)+\omega_m(f_i).e_m(P_i,P_j)$$

$$e_f(P_i,P_j) = \frac{|f_i - f_j|}{f_i + f_j}\ ;\ e_b(P_i,P_j) = \frac{|b_i - b_j|}{b_i + b_j}\ ;\ e_m(P_i,P_j) = |m_i\text{-}m_j|$$

The weighting tables on frequency $\{\omega_f(f_i)\}$, bandwith $\{\omega_b(f_i)\}$ and magnitude $\{\omega_m(f_i)\}$ are computed from the probability distribution functions of the first four formants on the frequency scale $(\{\Pi_{\mathcal{F}1}(f_i)\}, \{\Pi_{\mathcal{F}2}(f_i)\}, \{\Pi_{\mathcal{F}3}(f_i)\}, \{\Pi_{\mathcal{F}4}(f_i)\})$ considered [4] and from protoype values of weighting coefficients for the first four formants $\{\Omega_f(\mathcal{F}_i), \Omega_b(\mathcal{F}_i), \Omega_m(\mathcal{F}_i)\}_{(i=1,..,4)}$.

The global inter-spectra distortion measure D (S^R, S^T) is computed from the optimal clusterings of peaks of the given spectra S^R and S^T. The clustering of a peak P_i^R of S^R to a peak P_j^T of S^T is measured by the inter-peaks measure δ (P_i^R, P_j^T) and is defined optimal when :

$$P_j^T = \text{Argmin}_{\{P^T \in S^T\}} \{\delta (P_i^R, P^T)\}$$

Let Δ be the matrix of the inter-peaks measures of the two given spectra :

$$\Delta = \{\Delta_{ij} = \delta (P_i^R, P_j^T)\}_{(i=1,..,I\ ;\ j=1,..,J)}$$

The global inter-spectra measure D (S^R, S^T) is computed as the average of the distinct optimal clusterings located on the rows and the columns of the matrix Δ.

3.2. Inertia and Prototype from MR Analysis

The inertia is a measure of the stability of the set of observations computed from multi-resolution principle. The inertia is computed by the summation of the squared generalized distances of the MR observations set. Used as criterion of location of the stationary and highly non-stationary parts of speech, the inertia is a priori a well fitted measurement. The continuum of speech is analyzed by multi-resolution in a 10 msec asynchronous mode. The tables 1 and 2 give the results of the MR analysis on two sentences, the first one in French, the second one in English-American. For each labeled segment, the lowest inertia (I) is given with the best representative observation computed as the centroid of the MR observations set at the location of the lowest inertia. Only the first three peaks/formants of the generalized centroid are given. As expected, fricatives, stops, glides are not stationary as the inertia shows it. We find in table 3, for the vowels, the diphthongs and the semivowels the prototypes computed as the centroid of the representatives of the corresponding phonetic segments having a low inertia. The prototypes are planed to be used for identification experiments.

Ph.	I	P1	P2	P3
[l]	>99	(1590, 270, -7)	(2628, 46, 0)	(4209, 251, -17)
[a]	1.44	(580, 121, 0)	(1618, 113, -11)	(2592, 180, -21)
[d]	1.58	(1610, 373, 0)	(2641, 550, -10)	(3837, 374, -12)
[ə]	1.49	(405, 68, 0)	(1347, 121, -18)	(2623, 164, -26)
[m]	0.47	(2302, 698, -4)	(3554, 449, -1)	(3849, 201, 0)
[ã]	0.37	(485, 132, 0)	(929, 132, -9)	(1919, 268, -32)
[d]	2.38	(327, 87, 0)	(1387, 163, -23)	(2776, 158, -28)
[ə]	0.75	(317, 129, 0)	(1519, 105, -17)	(2633, 176, -23)
[t]	26.8	(1335, 680, 0)	(2406, 674, -2)	(3221, 793, -4)
[r]	14.7	(1286, 149, 0)	(3138, 727, -20)	(3668, 320, -15)
[a]	0.38	(737, 135, 0)	(1341, 176, -7)	(2649, 134, -17)
[v]	3.43	(1286, 222, 0)	(2594, 307, -9)	(3372, 249, -10)
[a]	0.39	(664, 113, 0)	(1547, 195, -11)	(2611, 148, -15)
[j]	1.16	(399, 55, 0)	(2069, 149, -27)	(3039, 186, -28)
[d]	0.85	(1797, 296, -5)	(2832, 652, -15)	(3760, 61, 0)
[e]	1.43	(363, 60, 0)	(1978, 191, -26)	(2748, 95, -20)
[p]	28.6	(1148, 592, 0)	(2476, 809, -3)	
[a]	1.00	(620, 151, 0)	(1615, 165, -10)	(2695, 104, -12)
[s]	15.2	(994, 386, -0)	(2066, 961, -7)	(3173, 602, -2)

Table 1: BREF80 database - Utterance identification i0mb0858 « La demande de travail dépasse...»

Ph.	I	P1	P2	P3
[dh]	>99	(389, 126, 0)	(1916, 285, -26)	(3103, 187, -25)
[ix]	4.77	(417, 53, 0)	(1807, 238, -28)	(2805, 158, -27)
[m]	0.84	(385, 108, 0)	(2413, 65, -17)	(3703, 182, -33)
[iy]	0.14	(445, 47, 0)	(2410, 102, -14)	(3283, 154, -18)
[dx]	0.16	(484, 137, 0)	(2319, 236, -15)	(3141, 204, -14)
[iy]	0.31	(504, 92, 0)	(2442, 160, -15)	(3151, 187, -16)
[ng]	3.34	(312, 163, 0)	(1101, 196, -18)	(2560, 469, -33)
[ih]	0.86	(566, 304, 0)	(1806, 364, -15)	(2948, 140, -16)
[z]	15.3	(1216, 371, -2)	(2083, 407, -0)	(2972, 603, -4)
[n]	2.22	(691, 188, 0)	(1824, 182, -11)	(3038, 697, -28)
[aw]	2.82	(800, 65, 0)	(1295, 135, -11)	(2837, 733, -40)
[ix]	6.15	(581, 301, 0)	(1324, 76, -0)	(2642, 326, -25)
[dcl]	11.3	(2217, 570, -1)	(2984, 562, -3)	(4487, 506, 0)
[jh]	7.30	(1162,345,-20)	(3305, 186, -1)	(3985, 202, 0)
[er]	1.67	(684, 169, 0)	(1613, 289, -8)	(3135, 278, -22)
[n]	1.66	(1947, 83, 0)	(2903, 186, -13)	(4851, 212, -24)
[dcl]	25.2	(1145, 413, 0)	(2445, 318, -3)	(4044, 1675, -9)
[d]	17.1	(867, 476, -2)	(2036, 445, -1)	(2708, 391, 0)

Table 2: TIMIT database - Utterance identification fcjf0sx307 « The meeting is now adjourned ».

Ph.	Ia	P1	P2	P3
[iy]	1.82	(432, 104, 0)	(2346, 230, -15)	(2620, 136, -10)
[ih]	1.68	(498, 129, 0)	(2131, 204, -15)	(2897, 331, -18)
[ix]	3.61	(538, 185, 0)	(1868, 181, -11)	(2729, 257, -17)
[eh]	1.55	(676, 64, 0)	(1761, 166, -14)	(2612, 247, -20)
[ae]	0.88	(731, 66, 0)	(1724, 131, -8)	(2288, 191, -14)
[aa]	1.31	(785, 89, 0)	(1376, 177, -9)	(2443, 158, -22)
[ao]	2.20	(662, 68, 0)	(1103, 82, -6)	(2294, 179, -28)
[uh]	2.17	(494, 102, 0)	(1381, 115, -13)	(2596, 144, -25)
[uw]	2.39	(470, 106, 0)	(1233, 155, -15)	(3113, 35, -15)
[ux]	1.65	(392, 108, 0)	(1626, 136, -13)	(2318, 140, -16)
[ax]	2.52	(537, 89, 0)	(1424, 122, -12)	(2810, 273, -22)
[axh]	3.34	(1619, 244, 0)	(2599, 486, -7)	(3799, 230, -5)
[ah]	1.72	(704, 131, 0)	(1324, 211, -11)	(2614, 250, -21)
[er]	2.53	(567, 70, 0)	(1535, 166, -10)	(1915, 155, -14)
[ey]	1.28	(514, 113, 0)	(2481, 157, -15)	(3121, 273, -18)
[ay]	1.58	(780, 92, 0)	(1583, 188, -12)	(2948, 226, -20)
[oy]	1.95	(685, 131, 0	(1391, 160, -8)	(2712, 366, -27)
[aw]	1.18	(843, 130, 0)	(1554, 141, -2)	(3009, 189, -24)
[ow]	1.87	(605, 87, 0)	(1203, 114, -10)	(2739, 224, -31)
[w]	4.12	(563, 97, 0)	(982, 96, -8)	(2750, 247, -43)
[y]	2.41	(417, 109, 0)	(2294, 376, -16)	(2505, 288, -15)

Table 3: TIMIT database - Prototypes found by multi-resolution from the 30 first speakers (the 15 first females and males) from the training set of the dialect dr1.

3.3. Speech Database Labeling Analysis

In the previous experiment, a single prototype is given per phoneme. The tracking of the inertia and of the prototypes on the 10 msec MR analysis pose the problem of the speech database labeling. In the neighborhood of steady states (characterized by a low inertia), the analysis of the formants parameters allows us to notice what is a priori a labeling error of contiguous frames. Examples are given in table 4 and table 5 on the two databases.

Ph.	I	P1	P2	P3
[iy]	0.80	(467, 55, 0)	(2441, 193, -18)	(3322, 178, -19)
[iy]	1.15	(474, 60, 0)	(2353, 204, -19)	(3271, 147, -17)
[iy]	3.61	(478, 93, 0)	(2317, 126, -13)	(3235, 111, -12)
[dx]	0.74	(486, 160, 0)	(2314, 149, -11)	(3204, 210, -14)
[dx]	0.16	(484, 137, 0)	(2319, 236, -15)	(3141, 204, -14)

Table 4: TIMIT database - Utterance identification fcjf0sx307. Continuum of speech analyzed by multi-resolution at the samples 5272, 5448, 5608, 5760, 5920.

The table 4 illustrates the well known difficulty of speech labeling. The third frame with the higher inertia is coherent with the change of label of the fourth frame. Nevertheless, the second and the third frames are not stationary parts of the signal as it is shown by their relatively high inertia. Are they well labeled by [iy] ? Are they relevant for any decision process ? Are they relevant in an evaluation of acoustic-phonetic decoding ?

In table 5, it is quite an evidence that the third frame is badly labeled. From the prototype tracking, the accurate label seems to be [ə] instead of [d]. A high inertia (300) on the second frame shows a highly non-stationary part of the signal : the end of [d] and the beginning of [ə]. The tracking of the prototypes doesn't show incoherence with the second label. But this part is surely not a right location for a decision process. The evaluation of acoustic-phonetic decoding should not take this frame into account.

Ph.	I	P1	P2	P3
[d]	88.7	(1311, 797, -4)	(3428, 401, -6)	(4566, 397, -2)
[d]	300	(1503, 377, -4)	(2655, 329, -6)	(3632, 155, 0)
[d]	1.68	(316, 122, 0)	(1488, 189, -20)	(2628, 160, -22)
[ə]	0.99	(312, 115, 0)	(1503, 146, -18)	(2635, 158, -22)
[ə]	0.75	(317, 129, 0)	(1519, 105, -17)	(2633, 176, -23)

Table 5: BREF80 database - Utterance identification i0mb0858 Continuum of speech analyzed by multi-resolution at the samples 14552, 14768, 14880, 15016, 15248.

4. PERSPECTIVES

4.1. Automatic Correction of Database Labeling

An automatic correction of speech database labeling could be developed from multi-resolution analysis. To detect an error labeling and to automatically correct this error a method could be as follows. For any frame which is contiguous to a yet labeled steady state, a low inertia coupled with a low inter-frames distortion measure detects an error, the label of the well identified steady state corrects the other label. The correction processing could be iterative until no label change occurs.

4.2. Hidden Markov Model Integration

There are two ways to use MR analysis for speech recognition in a Hidden Markov Model. The first one is the use of MR analysis instead of the classic 10 ms asynchrounous analysis [6]. The second one is the possibility to compute an a priori best probability emission using a set of probability emissions issued of the set of the analysis. In the HMM decoding step, a future experiment will consist in suppressing the highly unstable continuum.

5. CONCLUSION

Preliminary experiments using multi-resolution analysis are presented. An inertia measure has been introduced. It's a confidence coefficient of the frame analysis. For each phoneme, the average inertia and the prototype are computed for French and English-American languages. The results are given for vowels, diphthongs and semivowels of English-American. At last, the evolution of the inertia could be used to detect and correct the errors of a database labeling.

The multi-resolution method a priori allows a feedback between analysis method and any principle of decision such as dissimilarity measures minimization or output probabilities maximization.

6. REFERENCES

1. Liu, W., Andreou, A.G., and Goldstein, M.H., « An analog cochlear model for multi-resolution speech analysis », *ASA 124th Meeting*, 1992.

2. Gauvain, J.-L., Lamel, L.-F., and Eskénazi, M., « Design Considerations and Text Selection for BREF, a large French read-speech Corpus », *ICSLP* : 2359-2362, 1990.

3. Fisher, W., Zue, V., Bernstein, J., and Pallet, D. , « An Acoustic-Phonetic Data Base », *JASA* : pp. 81, S92, 1986.

4. Caraty, M.-J., Montacié, C., and Barras, C., « Integration of Temporal and Frequential Structurations in a Symbolic Learning System », *ICSLP* : 475-478, 1992.

5. Yé, H., Caraty, M.-J., Boë, J.-L., and Tufelli, D., « Structural (Phonetic) Evaluation of Dissimilarities Functions Used in Speech Recognition », *Eurospeech* : 404-407, 1989.

6. Garner, P.N., and Holmes, W.J., « On the Robust Incorporation of Formant Features into Hidden Markov Models for Automatic Speech Recognition », *IEEE-ICASSP* : 1-4, 1998.