# Flood-Filled Mel Cepstrum Coefficients for Vowel Binary Coding

Salam Fraihat and Hervé Glotin

*Abstract*— **We propose a new vowel representation using Flood Fill processing applied to time frequency image MFCC (FFMFCC). We use Allen time algebra to demonstrate that simple binary features from FFMFCC are enough to represent vowels. The results, multispeaker broadcast news, show that these binary features need only 1,2kb/s to give similar vowel classification than usual MFCC (76kb/s), yielding to a strong rate coding compression (factor compression of 60). Our approach yields to a new parsimonious representation of speech. We discuss on its extension to other phoneme coding and application to content based information retrieval.**

*Index Terms*— **Cepstral MFCC Image processing, time-frequency, Quantization, Allen Temporal Algebra, Automatic Speech Recognition.**

## I. Introduction

Most of the acoustic speech analysis systems are based on short-term float spectral features. The most popular one is the Mel Frequency Cepstrum Coefficients (MFCC) [15]. Moreover, from a phonological point of view, speech includes strong singularities, as depicted in the quantal speech theory [14]. These singularities may not be well represented in noisy conditions by a float representation. We then propose that a speech quantization framework may be efficient to generate lighter automatic speech recognizers and a better noise robustness.

In [3], we proposed a phoneme coding method using Allen algebra[1] on a very simple thresholded spectrum image. This method gives an average vowel error rate, but processes only 10% of the vowels, mainly because of the difficulty to set accurate threshold. This difficult issue was arised in spectrogram images processing [5], and more specifically into speech spectrum [6]. However, these papers are not giving solution for automatic speech recognition. In [1] we developed this concept on the subbands voicing level activities. We built an empirical spectrum image processing method that was not matching the correct phoneme dynamics, yielding to poor results.

Here we propose to build the Time Frequency patterns by a systematic and simple recursive algorithm [12, 13], applied on spectrum image of MFCC coefficients. The feature we propose in this paper for speech analysis lies on a basic binarisation of MFCC Time-Frequency (TF), to reveal informative structures speech singularities. Our approach considers the TF plane in a global manner, where each speech spectral pattern is not characterized by an iso-energy level but by a local TF coherency.

The validation experiments are conducted on reference French independent speaker broadcast news called ESTER[2] [8], with a training set of 10 hours, a validation and test set of one hour.

The FloodFill (FF) is presented in next section. The section 3 presents our MFCC binarization process, using the (FF) image processing, yielding to what we call FFMFCC parameters. Section 4 describes the Allen time algebra coding on this FFMFCC parameters. The first, vowel classifications based on this coding defines the optimal FFMFCC parameters: a window size of 16ms for a simple binary representation. This Boolean representation is then successfully experimented in section 5, showing similar results than MFCC, but with a much lower rate coding. The last section concludes and gives perspectives of this promising new approach for speech content information retrieval.

## II. Towards FloodFill Spectrum Segmentation

We propose in this paper to apply the image pattern extraction algorithm "Flood-Fill" [7] (FF) to the speech spectrogram image before any pattern extraction process.

FF is this simple image process:

--Two cells are defined as connected, if both exceed the threshold defined.

--FF is performed recursively on all cells connected to the interest element (e.g. the initial cell in each iteration,

Manuscript received July 01, 2009. This work was supported in part by the Systems and Information Sciences Lab UMR CNRS, University of Sud Toulon Var, Department of Computer Sciences and The French National Found for Research ANR and ANCL project.

Fraihat. S. is with Systems and Information Sciences Lab UMR CNRS 6168. USTV - B.P. 20 132 - 83 957 La Garde, France (phone: +33 (0)4 94 14 28 71; fax: +33 (0)4 94 14 28 97; e-mail: fraihat@ univ-tln.fr).

Glotin. H. is with Systems and Information Sciences Lab UMR CNRS 6168. USTV - B.P. 20 132 - 83 957 La Garde, France (phone: +33 (0)4 94 14 28 24; fax: +33 (0)4 94 14 28 97; e-mail: glotin@ univ-tln.fr).

[1] N.B.: Allen J.B works are related to human speech perception and subband speech analysis, while ALLEN J.F defined a generic time interval algebra.

[2] ESTER: Evaluation campaign of continuous speech broadcast

called ”seed” S0).

Thus FF has 2 parameters: a threshold Th and the initial cell S0.

We compute FF into each spectrum image of Local

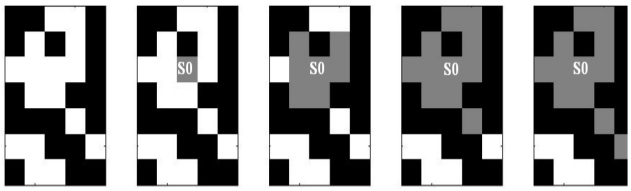


Fig 1. Illustration of the flood-fill procedure. The original image is shown in the left panel. Moving right: a seed S0 is chosen; the neighbors according to given criterion as the cell or pixel intensity to the seed are filled; then the neighbors to the neighbors are filled; so on until all connected points have been added to the stack [13].

Binary Window (LBW) of length optimized on a development set (varying from 16ms to 512ms, half shifted).

We think that FF can effectively eliminate the noise existing in spectrum data, revealing clearly the voicing patterns. The application of the recursive FF algorithm can solve the problem of a local thresholding such as the algorithm finds recursively all segments that are connected to a start node seed S0, in the eight directions, by a path along which all values exceeding the threshold (Fig.1).

In [1] we apply FF to Time Frequency Image of voicing levels (=Harmonic to Noisy Ration [1]). Then, FF allows extracting the patterns (Fig 2 (c, d)) with few parameters. Anyway, the results of recognition are weak with FloodFill on voicing data (see section 5-A).

Thus we propose to change the spectrum image of the voicing levels: we will apply FF on the spectrum image of Mel Frequency Cepstrum Coefficients (MFCC) .

## III. FLOOD-FILLED MFCC

The Mel Frequency Cepstrum Coefficients are one of the most popular speech front ends. It is computed by windowed Fast Fourier Transform in 12 Mel-Frequency critical subbands, defined after physcoacoustical studies [15]. These 12 dimensions float vector is concatenated with the log-energy.

The usual MFCC feature includes the delta and delta-delta of the 12 MFCC statics coefficients, yielding to 39 coefficients. In this paper the MFCC are computed with the toolkit ”SPRO” from IRISA [9], on window of 32ms (4ms shift). First we apply FF with different thresholds on spectrum image of normalized (min max) MFCC coefficients, but we get an unsatisfactory error score. This is because we do not code the harmonic of energy that exists in the high frequency subbands. This energy allows the discrimination of vowel patterns.

Thus we propose here a new extraction pattern algorithm (called UFF), based on FF algorithm.

In UFF algorithm we apply FF algorithm in each subbands in order to code the complete harmonic of energy

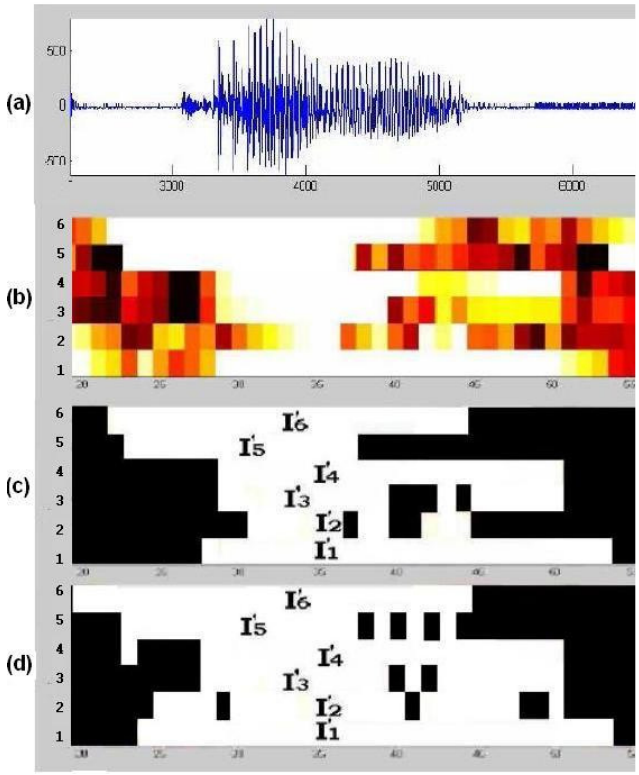


Fig. 2. Pattern extraction by Flood Fill algorithm [1]: (a) Speech signal of one vowel. (b) The voicing level by subands. (c) The pattern detection used Flood Fill algorithm with threshold = 0,7 (d) same as (c) with threshold = 0,4.

exist in speech signal.

The extraction algorithm UFF of our binary pattern (called FFMFCC), proceeds as follow:

1. Select the maximum value of intensity on the LBW: the first maximum is the starting point S0 (S0 like defined in Fig.1).

2. Calculate the initial points $^{(Si)}$ / $i \in [1, 6]$ : for each pair of LBW subands, the initial point is the maximum value that is on the same segment as the first starting point S0.

3. Calculate the threshold for each pair subands $^{Th}i$ : which is the value of the first starting point $^{S}i$ divided by n. In our experiments, we tried different n values, with n = 64 we had the best result.

4. Apply FF algorithm on LBW with $^{(Si, Thi)}$ parameters, in order to generate the six patterns over the LBW.

5. The final pattern FFMFCC is the logical fusion of all the patterns obtained.

Samples of FFMFCC are given in Fig.3, for LBW=128ms and LBW=16ms (half shift).

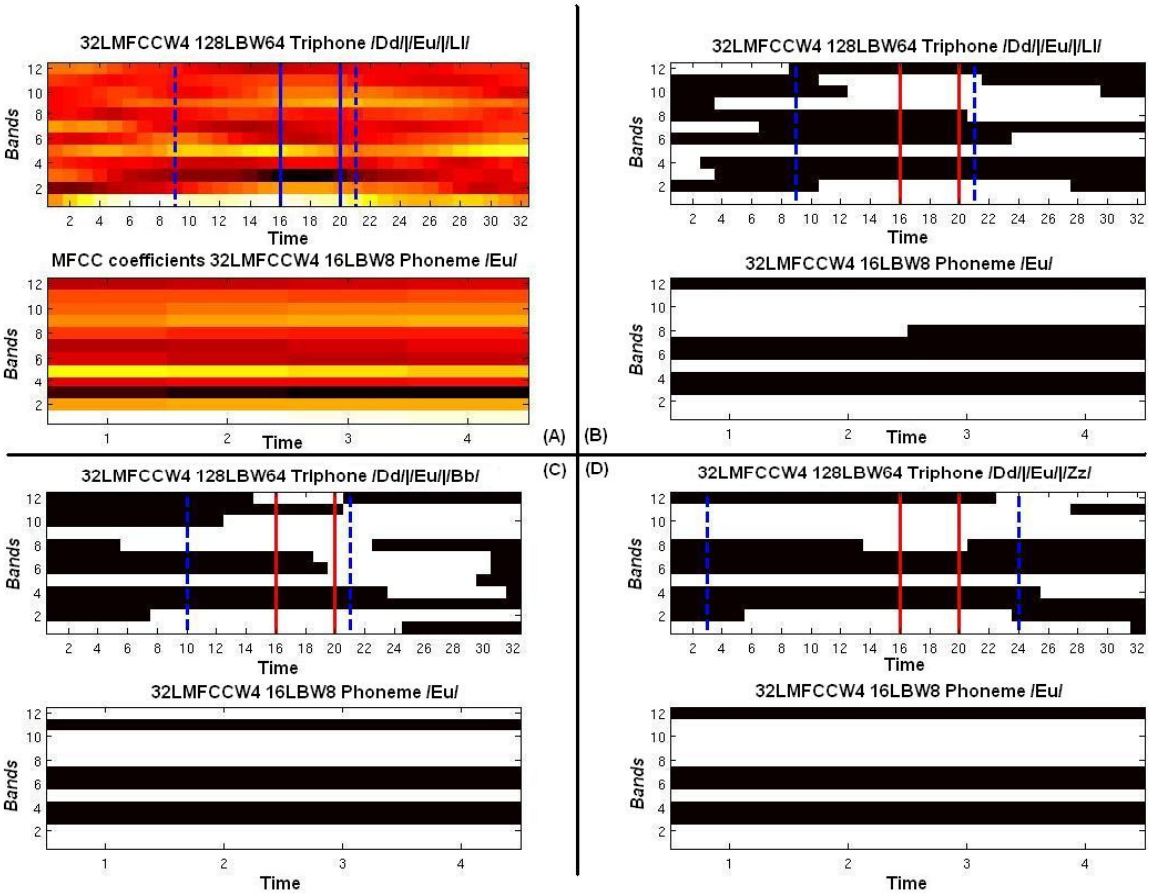


Fig. 3. FFMFCC samples of the phoneme /Eu/, inside three different triphone sequences, for 2 different spectrum image of LBW window size: 128ms (top) or 16ms (bottom). (A): the MFCC values, in which UFF generates the FFMCFC represented in (B). The (C) and (D) represent two other FFMFCC vowel examples. Into the top subfigures, we plot in dashed lignes the forced Viterbi label of the /Eu/ phoneme and in full lines the 16ms LBW window. Each bottom subfigure in B,C,D, represents the FFMFCC computed by UFF inside the 16ms window figured in the upper respective figure. This short time FFMFCC is similar to each others, and represents the centre of the vowel /Eu/.

## IV. ALLEN INTERVAL ALGEBRA VOCALIC CODING

As FFMFCC naturally defines intervals, we code their relations as proposed in [4] using the Allen temporal algebra [2] depicted in Fig.4. We apply Allen's algebra to each couple (k,l) of the 12 FFMFCC subbands, where:

$k \in [1:12]$ and $l \in [k:12]$, $k$ different of $l$. Thus the Allen FFMFCC vector is defined as: [Allen(Sb1, Sb2),Allen(Sb1, Sb3)...Allen(Sb11, Sb12)], resulting in 66 integers. We concatenate it with the log energy and call it AllenMFCC. It feeds a Multi-Layer Perceptron (TORCH [11]). The hyper parameters are optimized on dev. set. Considering LBW=16ms, the MLP MFCC inputs consist in 39 floats * 4, which equals to 156 real, 1248 bits. For the same kind of window, the MLP AllenMFCC inputs are only 66 integers + 1 real (= energy mean), 272 bits. Therefore the coding rate compression is ≈ 5.

| Relation | Symbol | Illustration |
|---|---|---|
| X before Y | 1 | |
| X meets Y | 2 | |
| X overlaps Y | 3 | |
| X starts Y | 4 | |
| X during Y | 5 | |
| X finishes Y | 6 | |
| X equals Y | 7 | |

Fig.4. The Allen algebra with their symbols. There are 6 relations + their inverse (+7) + the "equal" (7) and the "no-relation" (14).

## V. Results on Vowel Classification

Experiments are made on one hour of francophone broadcast news continuous speech (with half women and half men) from ESTER campaign [8].

We use the vowel labels for the training step, which are given from forced Viterbi realignment [8] according to LIA SPEERAL system [10]. Each LBW is labeled with the label that overlaps it the most. In dev. or test phase, we use one hour of ESTER broadcast news. We consider the six most frequent French vowels: /Aa/,/Ai/,/An/,/Ei/,/Eu/,/Ii/.

The expected ER of the random classifier is defined by

$$ER_{rand} = 1 - \sum_{k=1}^{c} (P_k)^2 = 1 - \sum_{k=1}^{c} \left( \frac{card(C_k)}{\sum_{k=1}^{c} card(C)} \right)^2 \quad (1)$$

where c is the number of classes, card(Ck) is the number of elements of the vowel class Ck in the train set.

The Error Rate (ER) of the random classifier is 83% ER.

### A. Vowel recognition from Voicing data

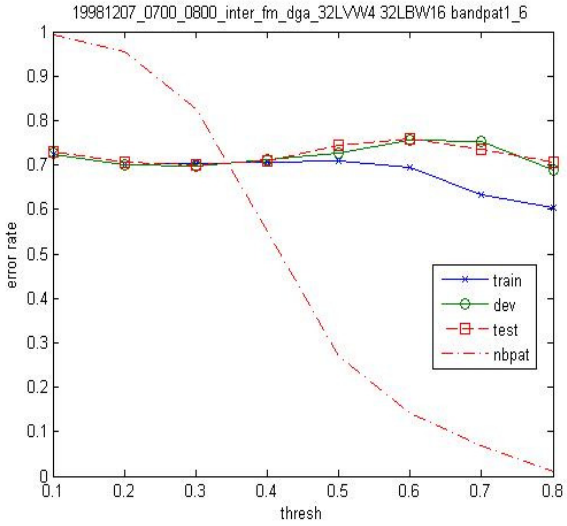


Fig. 5. Train, dev and test error rate and nbpat (nbpat: detected pattern number / exists pattern number) with different threshold for 1hour of INTER_FM radio broadcast with FF on voicing data.

$$\text{Nbpat} = \text{Card(Detec ted Pattern)/C ard(Existi ng Vowel)} \quad (2)$$

We apply FF on spectrum image of voicing levels parameters (Fig.2-(c,d)) (for more details for voicing levels parameters see [1]) . We note that the vowel class error rate is high (~70% error rate, Fig.5) but this model needs few parameters, and detects more enough patterns. Then we change parameters to MFCCs coefficients.

### B. Vowel recognition from MFCCs coefficients

We apply FF with different thresholds on normalized (min max) 12 subbands MFCC images, we get an interesting result (42% error rate) but with a low nbpat (27%). Thus we apply UFF on MFCCs coefficients. The parameters n, and LBW length, are set with the validation set, thus n = 64. We plot in Fig. 6 the class error for various LBW.

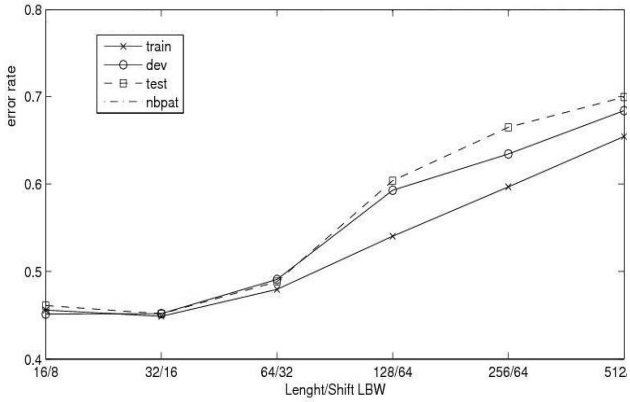


Fig. 6. Error rate curves using the Allen temporal coding on FFMFCC (”AllenMFCC”), on train, dev. and test set, with different LBW window length (half shifted), where is applied the FF on the 12MFCC (with n=4) to produce the FFMFCC. It demonstrates that the optimal LBW=16ms. Nbpat ≈ 98% for all LBW sizes.

The minimal ER (42%) is at LBW=16ms (half shifted). Interestingly, the nbpat (see (2)) nearly equals 100%. Thus we can assume that the correct time scale for vowel AllenMFCC coding is 16ms. Actually, at larger time scale, the pattern is much more complex because of the association before and after the vowel center with other phonemes, as shown in Fig. 3.

The table 1 shows that AllenMFCC is compressing by five the usual 39 MFCC subbands, with a coding rate of 16 kb/s (against 76 kb/s for 39 MFCC subbands). Unfortunately, the vowels recognition is affected, with more than 40% ER against 27% ER for MFCC.

We note that in Fig. 3, there is a difference between the 3 FFMFCC patterns on 128 ms LBW. However, we observe stable patterns for the same 3 examples of /Eu/.

The Fig.7 represents the AllenMFCC relations histogram. We see that the vowels signal generates only two Allen relations:”equal” and”no-relation”. This suggests that simple binary coding of the 16ms AllenMFCC may represent the vocalic center. This is related to the fact that vowel production is synchronous in all MFCC subbands. We investigate this point in the next section.

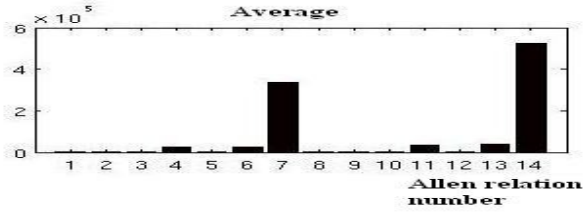


Fig 7. The AllenMFCC relations histogram, averaged on the 6 vowels of 1 hour of broadcast news (ESTER), for optimal parameters n=64 and LBW=16ms. Only 2 relations are present: "Equal" (7) and "Empty" or "NoRelation" (14).

## VI. Results on Boolean Vocalic Center Coding

According to the previous results and the stationary nature of vowel, we quantify FFMFCC patterns, for n=64 and LBW=16ms, with a simple 12 Booleans vector, called BoolMFCC, where in each subbands, any interval longer than 8ms is set to 1, otherwise to 0. Results (Tab. 1) show that BoolMFCC requires only 1.2 kbit/s, without affecting ER as we get similar ER than AllenMFCC (43%).

In Fig.8, we note that the average curves for MFCC and FFMFCC parameters for each vowel are correlated, such that the same subbands represent the same vowel. But with more contrast for FFMFCC, because UFF algorithm reduces noise in MFCC coefficients. It extracts just the structure pattern that represents the vowel. This proves that the UFF algorithm reveals informative structures speech singularities. This structures speech singularities of each vowel, allowing more efficient vowels classification.

Finally, on can assume that ER difference between MFCC and BoolMFCC may be due to the fact that BoolMFCC is only coding the vocalic center, while the automatic segmentation provides large vocalic context. Thus we propose to estimate the whole vowel window by the average of the contiguous MLP outputs, over a 32ms mid-term context. This smooth estimator, called means(BoolMFCC) is more precise than local BoolMFCC, yielding to the same vowels coding performance than MFCC (27% error rate), but with a very interesting rate coding compression (by a factor of 60). Moreover we can see from the confusion matrices (Fig 9) that MFCC and mean(BoolMFCC) produce similar errors.

## VII. Discussions and Conclusion

Three contributions are made in this paper. The first is the investigation of the image based vowels representation by the simple FF algorithm, which is successfully fused with speech processing. Our approach is very simple and thus contrasts with the statistical image process for spectrum analysis proposed in [6], recently applied into speech coder in [5]. Thus it will be interesting to investigate our FFMFCC parameters for speech coder.

The second contribution is the definition, implementation and validation of BoolMFCC feature for vowel classification. We show that, Mean(BoolMFCC) features generate a similar vowel error rate on broadcast news than the state of the art MFCC feature, but with a much lower coding rate (1.2 Kb/sec instead of 23 Kb/sec of 12 MFCCs).

Finally, BoolMFCC has another advantage: it is coding, and thus detects the vowel center inside short-term windows. Thus, BoolMFCC could be used for automatic resynchronization of labels. We see clearly in Fig. 3 the stable FFMFCC pattern of the vowel center (between the two vertical full lines), which represents the intrinsic vowel information, inside the global vowel segmentation produced by the forced Viterbi (between vertical dashed lines). This

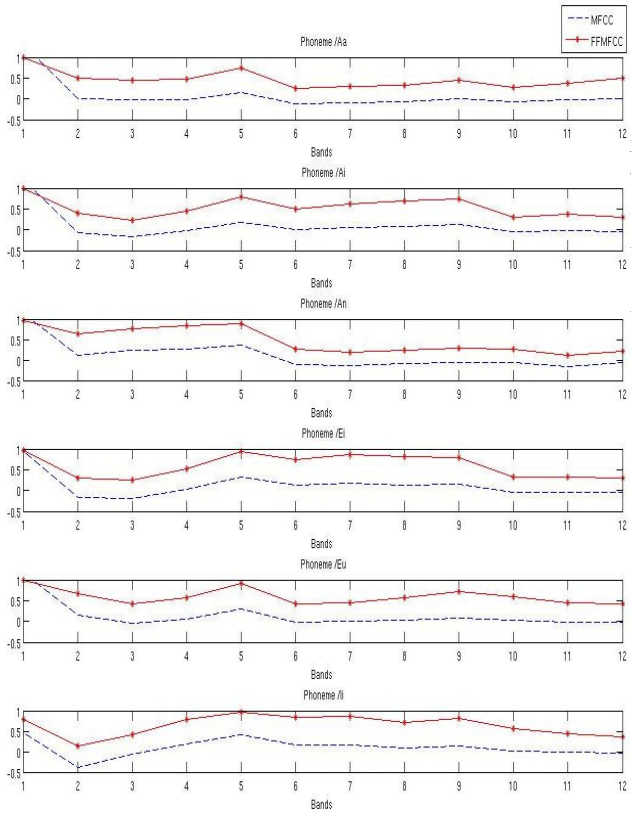


Fig. 8. Average curve of 12 subbands of MFCC and FFMFCC features in LBW lenght =16ms, and n=64, for each vowel.

TABLE I
Class Error Rates of all Systems

| Type of Features | Baud (kb/s) | # dim | CR | class error rate | |
|---|---|---|---|---|---|
| | | | | dev (%) | test (%) |
| full MFCC (39dim) | 76 | 39f*4 | 1 | 24 | 27 |
| static 12 MFCC | 23 | 12f*4 | 3 | 26 | 29 |
| AllenMFCC | 16 | 66i+1f | 5 | 45 | 46 |
| BoolMFCC | 1,2 | 12b+1f | 63 | 42 | 43 |
| Mean(BoolMFCC) | 1,2 | 12b+1f | 63 | - | 28 |

*The vowel class error rate (ER) of the random vowel classifier is 83% (calculated by (1)). #dim : dimension number where (b:boolean, i:integer, f:float). CR: Compression Rate = dimension ratio against the 39MFCC. Mean FFMFCC: vowel estimation by average on 32ms window, integrating 4 contiguous MLP outputs of 16ms windows. "-" we calculated* Mean(FFMFCC) *on the test set only.*

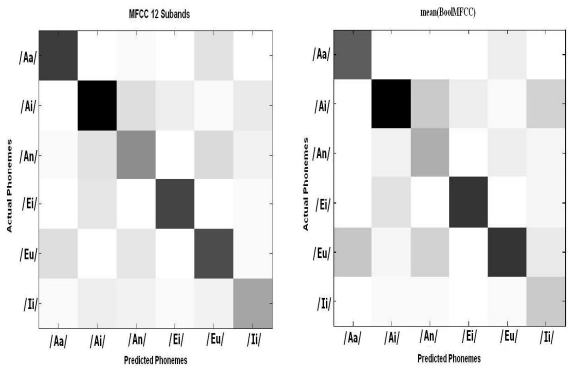


Fig 9. Confusion matrices of MFCC and mean(FFMFCC) systems. The two systems are producing nearly the same kind of confusions, even if the last one is based on speech representation which is 60 times lighter.

Our approach follows also the ULBP framework [12]. We propose in future work to use binary speech dynamics FFMFCC coding for a direct matching of words in a signal audio. This method may be useful for information retrieval in Broadcast news.

## ACKNOWLEDGMENT

We thank G.LINARES at LIA and G.GRAVIER at INRIA for having given the phonetic labels from their ASR system.

## REFERENCES

[1] Fraihat, S. and Glotin, H., "Towards Image Processing Compact Vowels Coding", Int. IEEE Conf. on Sciences of Electronic, Technologies of Information and Telecommunications-SETIT, Tunisia, mars 2009

[2] Allen, J.F.,"An Interval-Based Representation of Temporal Knowledge". In Proceeding 7th IJCAI, 221226, August 1981.

[3] Fraihat, S. and Glotin, H.,"New Time-Frequency quantization for parsimonious speech coding", IEEE Int. Conf. Signal Processing and Multimedia Applications-SIGMAP, Portugal, July 2008.

[4] Glotin, H.,"When Allen J.F. meets Allen J.B.: Quantal Time-Frequency Dynamics for Robust Speech Features", LSIS Tech Rep., janv. 2006.

[5] Jellyman, K-A., Evans, N.W.D., Liu, W.M., and Mason, J.S.D.,"Towards a New Image-Based Spectrogram Segmentation Speech Coder Optimised for Intelligibility", LNCS MMM, 63-73, 2009.

[6] Hory, C., Martin, N., and Chehikian, A.," Spectrogram segmentation by means of statistical features for non-stationary signal interpretation", IEEE Transactions on Signal Processing, vol. 50, no. 12, pp. 2915 2925, December 2002.

[7] Andre, P., "Intelligent Flood Fill or: The Use of Edge Detection in Image Object Extraction" University of Southampton, 2005.

[8] Galliano, S.,Mostefa, D., Choukri, K., Bonastre, J.F. and Gravier,G. "The ESTER phase 2 :Evaluation campaign for the rich transcription of French broadcast news", In: Proceedings of the 9th European Conference on Speech Communication and Technology (InterSpeech 2005), Lisboa, Portugal, September 2005.

[9] Gravier, G., "Speech signal processing toolkit, release 4.0.", in IRISA Research Rapport, France, 2003.

[10] Linares, G., Nocera, P., Matrouf, D., Bechet, F., Massoni, D. and Fredouille, C., "The LIA speech recognition system", in LIA Research Rapport, 2005.

[11] Collobert, R.,Bengio, S. and Mariéthoz, J. "Torch: a modular machine learning software library", IDIAP-RR, 02-46, 2002.

[12] Maenpaa, T., "The local binary pattern approach to texture analysis, extentions and applications", Phd Thesis, Oulu university, Finland, 2003.

[13] Nosal, E-M., " Flood-fill algorithms used for passive acoustics detection and tracking", IEEE Passive Workshop. New Trends in environmental Sciences, France, 2008.

[14] K-N. Stevens, "On the quantal nature of speech", Journal of Phonetics, pp. 3–45, vol 17, 1989.

[15] S.B Davis and P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, journal IEEE Transaction on. ASSR, 357-366, 1980

**S. Fraihat (born 1981),** received the computer engineer degree from the Institut National d'Informatique, Algiers, Algeria, in 2004, and the master's degree in Systems and Information Sciences from University of Marseille in 2005. He is pursuing a Ph.D. degree in computer sciences at CNRS-Lab LSIS, University of Marseille, Marseille, France.
His main research interests are Automatic Speech Recognition, Time-Frequency Quantization, Vowel coding, Patterns recognition and content based Information Retrieval. Currently, he is working on ESTER Evaluation campaign of continuous speech broadcast news rich transcription, supported by AFCP (Francophone Association of Communication Speech) and DGA (The General Delegation for Armaments).

**H. Glotin (born 1970),** PhD 2001, Habilitation Degree (2007), has been assistant professor in the University of Sud-Toulon Var and a member of LSIS CNRS Lab since 2003. He had a master in artificial intelligence from PARIS VI univ. His PhD was on robust automatic audiovisual speech recognition and computational auditory scene analysis, in the Artificial Perceptual Intelligence Inst. IDIAP-EPFL CH & Inst. Nat. Polytec. Grenoble. In 2000 he was invited at the John Hopkins univ. as expert in the human language IBM team developing new speech analyses. In 2001 he was tenure in the Semantics & Syntax CNRS lab. as engineer-researcher. In 2002 he was invited at the NATO advances studies in speech signal dynamics. He is currently carrying researches on robust multimedia content based information retrieval He is qualified in signal processing and computer sciences. He is/was mentoring 7 PhDs. He is leading the Information Dynamics group at LSIS. He organizes recurrent scientific events on multimodal analysis. He is co-author of one hundred international journal or conferences articles.