NEW TIME-FREQUENCY VOWEL QUANTIZATION ENHANCED BY SUBBAND HIERARCHY

Fraihat Salam and Glotin Hervé
Information and System Sciences Lab - UMR 6168, USTV - B.P. 20 132 - 83 957 La Garde, France

Keywords: Speech analysis, Quantization, Time-frequency, Allen Temporal Algebra, Automatic Speech Recognition.

Abstract: Speech dynamics may not well be addressed by the conventional speech processing. We analyse here a new quantization paradigm for vowel coding. It is based on simple Allen temporal interval algebra applied on subband voicing levels, yielding to a compressed speech representation of only 21 integers for a speech window up to 32 ms long. Experiments show that we take advantage of the ranking of the average values of the voicing interval accross the various subbands. Theses new features are evaluated for vowel recognition (1 hour, 6 vowels) on a referenced multispeaker radio broadcast news used during evaluation campaign ESTER. We work on the subset of the most frequent french vowels. We get 62% class error rate adding the ranking information to the Allen's relations, instead of 70% using Allen relations alone, and 57% the set of the raw 48 floats. We then discuss on the advantage of using more subbands, and we finaly propose a strategy to tackle the combinatorial complexity of Allen relations.

1 INTRODUCTION

Most of acoustic speech analysis systems are based on short-term spectral features : Mel Frequency Cepstrum Coefficients (MFCC), PLP, etc... The purpose of this paper is to present and discuss a novel vowel representation. We propose here to use med-term Time Frequency (TF) speech dynamics. It has been established that phonological perception is a subband (SB) process (Fletcher, 1922). That has inspired various algorithms for robust speech recognition (Glotin, 2001), also linked to the TF voicing level (Glotin et al., 2001; Glotin, 2001). Nevertheless, the SB TF dynamics may be more investigated, compared to usual delta and delta-delta coefficients. Thus we propose in this paper a quantization of TF dynamics following some preliminary works (Divenyi et al., 2006; Glotin, 2006). We base our approach on voicing dynamics, composing binary intervals, assuming that they may provide a qualitative framework to generate parsimonious phoneme features using the time events representation proposed by Allen J.F.(Allen, 1981)[1].

In (Fraihat et al., 2008) we made preliminary experiments yielding to 70% of vowel classification. Here we present a method that significantly enhance the model, adding the subband ranking, and we discuss on further works. Experiments are conducted on the most french frequent vowels of one hour of the ESTER broadcast news database[2](Galliano et al., 2005).

In the first section of this paper we recall the Allen temporal Algebra, and the properties of TF voicing index. Then section 3 shows how we binarize and generate our speech parsimonious representation. After a presentation of the vowel coding, we propose different features sets and we show their class error rate results. We then discuss on the strategy that should be conducted for developing robust TFQ features.

2 ALLEN TEMPORAL ALGEBRA

A temporal algebra has been defined in (Allen, 1981; Glotin, 2006), where 14 atomic relations (including the 'no-relation' one) are depicted between two time intervals. These Allen's time relations are defined by one interval sliding another. If one set to 1 the a algebraic distance d between the two nearest intervals,

[1] Note that ALLEN J.B worked on SB speech analysis, but ALLEN J.F on generic time representation, while our model is based on both.

[2] ESTER: Evaluation campaign of continuous speech broadcast news rich transcription

Salam F. and Hervé G. (2008).
NEW TIME-FREQUENCY VOWEL QUANTIZATION ENHANCED BY SUBBAND HIERARCHY.
In *Proceedings of the International Conference on Signal Processing and Multimedia Applications*, pages 189-192
DOI: 10.5220/0001933601890192
Copyright © SciTePress

and increment it as the intervals move away, we define an integer for each relation. Thus the “b” symbol is coded into “1”, “m” into “2”, . . . for the 14 relations that are : before, meets, overlaps, stars, during, finishes, equals, and their symmetric (see (Fraihat et al., 2008) for details). The ’no-relation’ happens between two empty intervals. We propose to use these time representation for coding speech events into a small discrete integer set. In order to define the intervals we use the voicing levels as depicted in the next section.

In order to get the subband voicing activity intervals, we estimate the TF voicing activity interval using the voicing measure R (Glotin, 2001) that is well correlated with SNR and equivalent to the *harmonicity index* (HNR). R is calculated by autocorrelogram of the demodulated signal. In the case of Gaussian noise, the correlogram of a noisy frame is less modulated than a clean one. We first compute the demodulated signal after half wave rectification, followed by pass-band filtering in the pitch domain. Then we autocorrelate each frame of LVW (Local Voicing Window) ms long and we calculate $R = R1/R0$, where $R1$ is the local maximum in time delay segment corresponding to the fundamental frequency ([90 350]Hz), and $R0$ is the window energy. We showed (Glotin, 2001) that R is strongly correlated with SNR in the 5..20dB range as illustrated in fig. 1. The SB are defined as in ALLEN J.B. analysis (Allen, 1994; Glotin, 2001) : [216 778;707 1631;1262 2709;2121 3800;3400 5400;5000 8000] Hz.
We set for vowel recognition LVW=32ms, with a shift of 4ms.

3 BINARIZATION AND REPRESENTATION

In order to generate principal separated time intervals for Allen relations, we threshold the voicing levels : for each band and each window of Local Binary Window (LBW 32ms shift and 64 ms length), we binarize to 1 the T% frame highest quantil, the other to 0.

In order to remove noisy relation, we remove interval that is connected to any window range. Finally we keep window containing at least 4 connected intervals. We then derive their Allen temporal relations (see fig. 1). The vowel labels for the training task are given from forced realignment on standard HMM-MMG model (Galliano et al., 2005).

As we have 6 SB, we have 15 temporal relations (one for each couple), ordered from low to high frequency. In our example (fig. 1), from I’1 to I’5, we get the parameter vector [di di di oi oi d d d d s oi d oi f d], where i is the inverse relation. Then these TFQ features estimated in each LBW window, feed a neural network (any classifier could be used), that we trained for automatic vowel decoding.

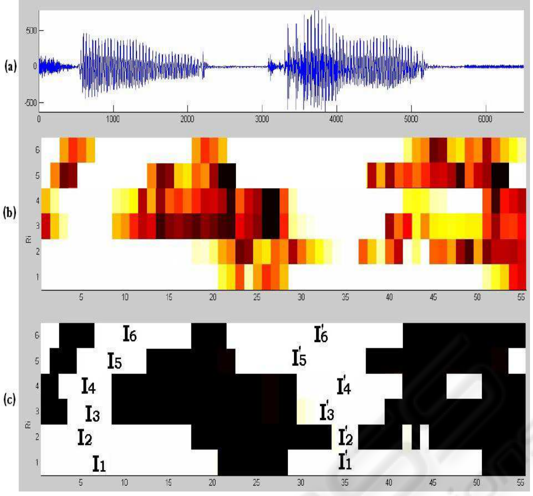

Figure 1: From voicing levels to the Allen’s interval relations: (a) voicing signal (b) the voicing level by subband (c) the binarized voicing levels by subband using mean threshold (From (Glotin, 2001)).

Moreover, in order to confirm that voicing levels and intervals definitions are informative, we build a 6 integer feature, called RANK, ranking the subband of each window using the relative R level of each interval. This information may be correlated to the formant position, that we lose in simple ALLEN relations.

Thus the functions of binarization and extraction should also integrate the hierarchy of SB frequency in ALLEN+RANK concatenated features.

4 DATABASE

Our experiments are made over all the speakers on the six most frequent French vowels: /Aa/,/Ai/,/An/,/Ei/,/Eu/,/Ii/. SB are defined like in previous section. We set the shift of each voicing window LVW to 4ms , and the LVW length to 32ms. We vary the T% parameter in [0.4 0.5 0.6 0.7]. The training windows are labelled with the label which covers at most the window. The features from 1h of continuous speech are used to train an MLP, and we test on other 20 minutes, best results with number of hidden units are given in tab.1.

Table 1: Results of the class error rates of the experiments. The error rate of the random classifier is 83%. T is the proportion of 1 in the window in each SB after binarisation. The Relative Gain is the relative reduction of error rate against the voicing experience. #dim : dimension number of the MLP input. CP: compression ratio of the Parameters. Nhu: Hidden units numbers of the MLP.

Type of Features	T	# dim	Type	# bytes	CP	Nhu	Class_Error Train (%)	Test (%)	Relative Gain (%)
Voicing	-	48	float	384	1	128	49,8	57,2	-
Binairy	0.5	48	bool	48	8	512	75,3	67,5	-18
Allen	0.4	15	int	60	6,4	128	10,1	72,2	-20,7
Allen	0.5	id	id	id	id	512	14,7	70,5	-23,2
Allen	0.6	id	id	id	id	128	10,2	72	-25,8
Allen	0.7	id	id	id	id	128	12,3	70	-23,3
Rank	0.5	6	int	24	16	32	67,4	69	-20,6
Allen+Rank	0.5	15+6	id	84	4,6	512	9,7	62,4	-9,1
Allen+Rank	0.6	id	id	id	id	128	11,7	65,1	-13,8
Allen+Rank	0.7	id	id	id	id	128	4,3	67,7	-18,3

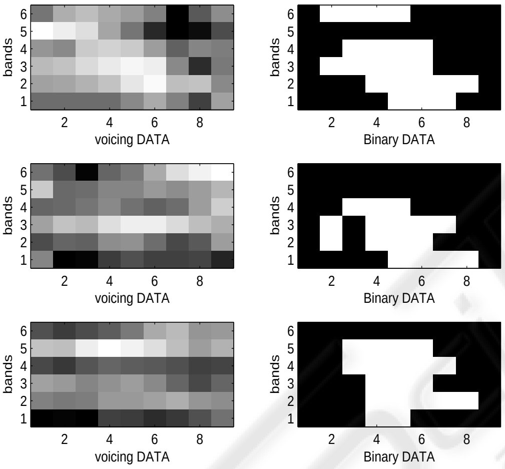

Figure 2: Example of float R voicing values and Binary data of three different sample of vowel /Aa/. The vectors of these examples are: vec1=[d,io,io,no,io,io,io,no,io,if,no,is,no,no,io], vec2=[io,io,io,no,no,s,id,no,no,id,no,no,id,no,no,no], vec3=[s,s,d,d,no,s,io,io,d,d,no,is,no,no].

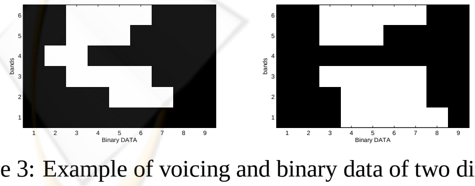

Figure 3: Example of voicing and binary data of two different samples of vowel /Ii/.

5 RESULTS AND DISCUSSION

We notice in the figure 2 that there is a shape similarity between patterns 1 and 2, and differences between 2 and 3. This may be due to different speakers that may negatively influence phoneme recognition (Fraihat et al., 2008). Further studies will have to be conducted on this issue.

The concatenation of subband ranks (RANK) and ALLEN features well improve the score : we have at best 70% Class Error Rate (ER) at best with ALLEN features alones, and 62% with ALLEN+RANK features (see table 1). Moreover it is interesting to note that RANK features alone, with 6 integers give 69% ER, similar to the complementary ALLEN features. This tends to show that the interval construction algorithm we propose extract representative information for vowel coding.

These vowel recognition results, with a feature compression of 4,6 are interesting (=62% ER), compared to the 57% ER given by the raw voicing data (we note that the direct binarization of the voicing data is worst) and compared to the 83% ER given by random classifier (see footnote[3]).

Moreover interval soft coding may enhance classification as revealed by the results of raw vs binary voicing levels. We then could use mean and variance interval length to enhance our classification. In future works will also use more detailed subband representation, like the 36 Mel Filter Cepstral Coefficient. This multiplication of the number of the intervals may explose the ALLEN representation size. A simple way to tackle this combinatorial effect is to generate local ALLEN relations on some frequency domains, and to train local classifier for each domain. Then a

[3] The error rate of the random classifier equals $1-\sum_{k=1}^{c}(P_k)^2 = 1-\sum_{k=1}^{c}\left(\frac{card(C_k)}{\sum_{k=1}^{c}card(C)}\right)^2$, where c is the number of classes, $card(C_k)$ is the number of elements of the class C_k in the train set.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

global classifier can merge the whole information, as depicted in fig 4. This strategy could allow the application of method to usual MFCC delta, delta-delta for example.

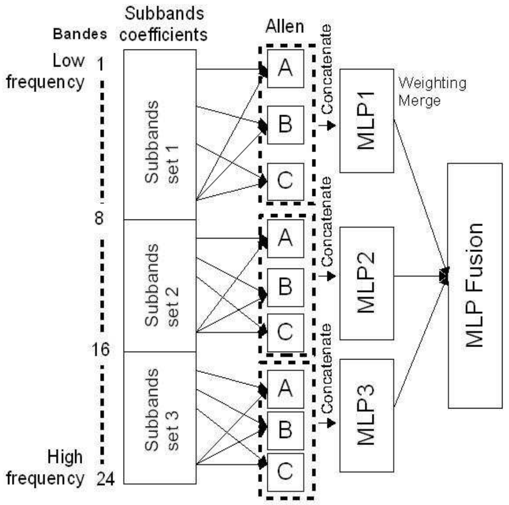

Figure 4: Schema of a further system. One may divide the features into three parts and for each part generate three Allen relations sets from six overlapped subbands, yielding to 3*15 relations that feed a MLP, finally one merge the three MLPs.

Further experiments will also be conducted on consonants, considering for example a zeros interval (ie. between two vowels) or by considering null binarized intervals (i.e. less or unvoiced intervals). A simple classic silence detector (e.g. based on energy thresholding) will avoid confusion between consonant and silence events. Moreover di-phones Consonant Vowel (CV), and triphones sequences (CVC) modeling could be done with simple extension of the same framework, and are expected to contribute to enhance ASR robustness.

ACKNOWLEDGEMENTS

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

REFERENCES

Allen, J. (1981). An interval-based representation of temporal knowledge. In *7th IJCAI*, pages 221–226.

Allen, J. (1994). How do humans process and recognise speech. In *IEEE Trans. on Speech and Signal Processing 2(4)*, pages 567–576.

Divenyi, P., Greenberg, S., and Meyer, G. (2006). *Dynamics of Speech Production and Perception*. IOS Press Inc.

Fletcher, H. (1922). The nature of speech and its interpretation. *J. Franklin Inst.*, 193 6:729–747.

Fraihat, S., Aloui, N., and Glotin, H. (2008). Parsimonious time-frequency quantization for phoneme and speaker classification. In *IEEE Conference on Electrical and Computer Engineering (CCECE)*.

Galliano, S., Geoffrois, E., a. M. D., Choukri, K., Bonastre, J.-F., and Gravier, G. (2005). The ester phase 2 : Evaluation campaign for the rich transcription of french broadcast news. *European Conf. on Speech Communication and Technology*, pages 1149–1152,.

Glotin, H. (2001). Elaboration and comparatives studies of robust adaptive multistream speech recognition using voicing and localisation cues. In *Inst. Nat. Polytech Grenoble & EPF Lausanne IDIAP*.

Glotin, H. (2006). When allen j.b. meets allen j.f.: Quantal time-frequency dynamics for robust speech features. Technical report, Research Report LSIS 2006, Lab Systems and Information Sciences UMR-CNRS.

Glotin, H., Vergyri, D., Neti, C., Potamianos, G., and Luettin, G. (2001). Weighting schemes for audio-visual fusion in speech recognition. In *IEEE int. conf. Acoustics Speech & Signal Process. (ICASSP)*.