A WAVELET REPRESENTATION EVALUATION FOR STOP–CONSONANTS CLASSIFICATION

Christophe Gerard, Marc Baudry, Alexandrina Rogozan
L.I.U.M., University of Le Mans
Avenue O. Messiaen, B.P. 535, Le Mans 72017 Cedex, France
Tel: +33 4383 32 21; Fax: +33 43 8335 65
E-mail: gerard@lium.univ-lemans.fr

ABSTRACT

Regarding Short Time Fourier Transform based methods, stop-consonants representation could be improved using the wavelet transform. After presenting our framework, we describe the wavelet parameterization and the classification method. Stop consonants are represented with pseudo-cepstral wavelet based parameters computed on a single-burst-neighbourhood-20 ms frame. Non-parametric nearest neighbours method is used. Evaluation is speaker-independent ; 1593 stop-consonants extracted from TIMIT database are evaluated. Results are described and discussed comparatively to MFCC's (Mel Frequency Cepstrum Coefficients). It appears that, in our field of research, wavelet gives equivalent classification percentages. The first thing which was pointed out, is the necessity to build an elaborated-wavelet-based-representation to get significant improvements.

Keywords : wavelets, stop–consonants, statistical analysis

1. INTRODUCTION

In speech recognition, one hard difficulty concerns the signal parameterization. To meet speech signal properties, Short-Time-Fourier- transform (STFT) representations are often used. Beside it, the wavelet transform appears as an attractive tool for speech signal processing.

1.1 Speech Signal Properties

Phonemes, elementary parts of speech have heterogeneous properties both in time and frequency. For example, vowels (e.g. /a/, /e/ or /i/), are considered as quasi-stationary events. In opposite, stop-consonants (e.g. /p/ or /g/) are very short time localised events.

Among the whole phonetic classes of speech, stop-consonants are likely the most difficult to localise and *a fortiori*, to parameter and to identify [4]. Indeed, the time frequency analysis task is to represent both occlusion, burst, friction and formantic transition.

For all these reasons, speech signal analysis would require a Time-Frequency representation with both high time and frequency accuracy [7].

1.2 STFT-Based Representation

For speech processing, the Short Time Fourier Transform (STFT) analysing tool is often used since it achieves a reasonable compromise to get a satisfying time and frequency resolution. Nevertheless, it is well known that such a representation also degrades short time events, presents in stop consonants especially.

1.3 The Wavelet Tool

Regarding the constant-frequency resolution of the STFT, the wavelet transform performs a constant-relative-frequency resolution which enable to step over the previously described STFT limitations. We use Barrat's fast wavelet transform [1]. This algorithm approximates Morlet's wavelet to a infinite impulse response filter. This continuous wavelet transform is also convenient for the Mel bandwidth selection as describe below.

2. OUR WAVELET PARAMETERIZATION

2.1 Working Context

For historical and convenient reasons, the signal analysis step is performed at constant bit-rate and constant time-duration. We have chosen to keep this working context and will exploit the well time localisation property of the wavelet. For stop-consonants parameterization, we represent the 10 ms burst neighbourhood [5] Time duration analysis is then set to 20 ms.

2.2 Psychoacoustic Considerations

Basic psychoacoustic phenomena will be taken in consideration: the frequency analysis is performed on a Mel scale, and thus, to take in count the frequency masking effect of the auditory system [8].

Concerning the temporal masking effect, will be partially integrate it using high or maxima value of spectrum energy.

2.3 Bit Rate Reduction

We previously discussed the frequency distribution: our wavelet tool performs a 24 Mel-filter-band analysis [6]. Concerning the time domain, we represent each 12 ms frame with one wavelet coefficient modulus for each frequency band. Finally, the inverse cosine transform reduce the representation to a set of 10 pseudo-cepstral parameters.

2.4 AMXW and SMXW parameters

AMXW stands for Asynchronous Mel Maximum Wavelet. This parameterization is constituted of one maximum coefficient for each frequency band, and for each frame of signal. The goal of AMXW parameterization is to extract the most significant information present in each frame without any time constraints.

SMXW stands for Synchronous Mel Maximum Wavelet. This second parameterization models each frame of signal with one singular event. The final vector is then constituted of a set of time-synchronous wavelets coefficients.

To get a comparative reference parameterization, MFCC [2] are evaluated too. Such a reference was chosen one the first hand, for its efficiency to represent speech signal, and on the other hand, for its cepstral representation domain, which is near SMXW and AMXW domain.

3. STATISTICAL CLASSIFICATION

Several classification techniques were tested: principal component analysis, standard and non-parametric discriminant analysis. Best results were obtained using the non-parametric nearest neighbours method.

3.1 Nearest Neighbours Method

This non parametric discriminant analysis is based on centroïd notion. Sample classification is determined by the major composition of its k nearest neighbours (k=11). Reseampling method is used, thus turn to turn, each sample belongs now to the test corpus, now to the learning corpus.

3.2 Corpus

To realise this research, we needed a corpus of labelled stop-consonants. We selected TIMIT database because stop consonants are labelled twice: on the first hand, occlusion, and on the second hand, burst, friction, and formantic transition. Our evaluation is based on TIMIT-DR1 subset composed of 1593 stop-consonants as shown Table 4.

3.3 Results

Confusion matrix are shown Table 1 to Table 3. The first column gives the tested stop-consonants and the next one line gives the classification result. Confidence interval is 3 points. Best results among MFCC, AMXW or SMXW methods are highlighted. Values placed on diagonal show good classification percentages.

Theses results show that MFCC and AMXW give equivalent classification rates. SMXW parameterization is significantly less appropriated to stop consonants.

Table 1. MFCC classification percentage

phn	b	d	g	k	p	t
b	**75,1**	3,8	0,00	0,00	19,4	1,6
d	12,7	40,1	1,5	4,1	14,9	26,7
g	12,6	4,8	53,9	19,8	4,2	4,8
k	3,6	2,1	21,2	51,8	11,1	10,1
p	15,6	1,4	0,00	1,4	76,6	5,0
t	2,9	8,6	1,3	3,5	13,3	70,5

Table 2. AMXW classification percentage

phn	b	d	g	k	p	t
b	61,6	8,1	1,6	0,5	25,4	2,7
d	16,8	41,6	1,9	4,3	11,2	24,2
g	10,8	4,2	**56,9**	16,8	5,4	6,0
k	3,4	4,4	21,8	52,3	8,3	9,8
p	13,8	0,9	0,9	0,9	77,5	6,0
t	3,2	7,0	2,5	5,1	14,3	67,9

Table 3. SMXW classification percentage

phn	b	d	g	k	p	t
b	69,7	4,9	0,5	1,1	20,5	3,2
d	18,6	35,7	2,8	4,7	14,6	23,6
g	9,0	4,8	48,5	24,5	9,6	3,6
k	7,8	3,9	21,2	47,7	9,6	9,8
p	22,5	3,2	2,3	2,3	64,2	5,5
t	9,8	8,2	2,5	5,7	15,2	58,4

4. DISCUSSION

We will now discuss our results and give some explanations. We will especially wonder why wavelet are not more efficient and which improvement could be done.

4.1 Why wavelets are not more efficient?

The first think to remind is the generalist property of our wavelet parameterization. The same kind of parameters is used to represent the whole heterogeneous time-frequency properties of speech signals.

4.2 Which improvement could be done?

We notice that SMXW provides less good results than AMXW or MFCC. This show that the event selection is a crucial stage which could be dedicated to stop-consonants-detection. For more detail, refer to MALBOS's researches [5].

In an other way, we could integrate first and second derivative of initial coefficients to add dynamic features. This will probably give a minor improvement, but no more [4].

4.3 What kind of events do we detect ?

Regarding STFT-based representations, the wavelet transform enables to access more accurately to the Time-domain information. And we consequently detect much more transient events whose do not have any signification at the phonetic level. It doesn't mean that wavelets are inefficient for parameterization, but shows that a much more sophisticated process should be used.

5. CONCLUSION

This paper relates our recent researches concerning the use of wavelet representations for speech signals parameterization. We particularly dealt with stop-consonants phonetic class because it is for this class that wavelet transform could give significant improvements.

The main conclusion is that it is no more reasonable to think that a simple exchange between STFT and wavelet modules would be sufficient to get significant improvements [3]. Our personal think is that good wavelet localisation both in time and frequency properties can not and should not be used everywhere. Nevertheless, wavelet stay an extremely powerful tool which must be used whenever its properties are useful, but far from fashioned effects.

ANNEXE

Table 4. Corpus description

phn	b	d	g	p	t	k	total
nb	185	322	167	328	218	315	1593
%	11,6	20,2	10,5	24,2	13,7	19,8	100

REFERENCES

[1] M. BARRAT, O. LEPETIT, "Calcul Rapide de la Transformée en Ondelettes", Traitement du Signal Vol 8, N°1 pp. 43-49, 1991

[2] S. B. DAVIS, P. MERMELSTEIN, " Comparison of parametric representations for monosyllabic word recognition in continuoulsy spoken sentences", IEEE ASSP 28 N°4, pp. 357-366, 1980.

[3] C. GERARD, M. BAUDRY, "Ondelettes et paramétrisation du signal de parole en milieu bruité", XV^e congrès international d'acoustique, Trondheim, NORVEGE, Juin 1995

[4] C. GERARD, "Etude de la paramétrisation du signal de parole à partir de représentations en ondelettes", Thèse de Doctorat, Université de Paris-Sud, centre d'Orsay, Décembre 1995

[5] F. MALBOS "Détection et identification des occlusives à l'aide de la transformée en ondelettes", Thèse de Doctorat, Université de Paris-Sud, centre d'Orsay, Janvier 1995

[6] C. MOKBEL, "Reconnaissance de la parole dans le bruit : bruitage / débruitage", Thèse de Doctorat, ENST Paris, 1992

[7] A. OUAHABI, "Représentations temps-fréquence : une revue orientée vers le traitement de la parole", MCEA, Grenoble, Septembre 1995

[8] E. ZWICKER, R. FELDTKELLER, "Psychoacoustique", CNET-ENST, MASSON, 1981