WAVELET-BASED ENERGY BINNING CEPSTRAL FEATURES FOR AUTOMATIC SPEECH RECOGNITION

Sankar Basu and Stéphane H. Maes

IBM T.J. Watson Research Center,
Yorktown Heights, NY 10598, USA

Abstract

Speech production models, coding methods as well as text to speech technology often lead to the introduction of modulation models *to represent speech signals with primary components which are amplitude-and-phase-modulated sine functions. Parallelisms between properties of the wavelet transform of primary components and algorithmic representations of speech signals derived from auditory nerve models like the* EIH *lead to the introduction of* synchrosqueezing measures. *On the other hand, in automatic speech (and speaker) recognition, cepstral feature have imposed themselves quasi-universally as acoustic characteristic of speech utterances.*

This paper analyses cepstral representation in the context of the synchrosqueezed representation - *wastrum. It discusses energy accumulation derived wastra as opposed to classical* MEL *and* LPC *derived cepstra. In the former method the* primary components *and formants play a primary role. Recognition results are presented on the Wall Street Journal database using IBM continuous decoder.*

1 Introduction

A new method for processing speech signals that uses the wavelet transform as a fundamental tool has recently been introduced in [4, 15, 16, 17, 12]. While the primary emphasis of the initial study reported in these papers had been in speaker identification, in the present work we prepare the extension of the same methodology with special attention to automatic recognition of continuous speech. The underlying method essentially involves 'treating' the wavelet transform of the speech signal in a very specific way, which is called synchrosqueezing. This method of processing includes physiologically motivated auditory nerve models, the ensemble interval histogram (*EIH*) model, and the so called *AM-FM* modulation model of speech production, but now all synthesized together within the more concrete framework of wavelet transform.

2 Synchrosqueezed representations

Two key steps are involved in the method. The first is the computation of the wavelet transform [3, 24], and the second is the the process of synchrosqueezing, which is necessitated by the somewhat de-focussed character of the wavelet transform of speech signals in the time-frequency plane. The wavelet transform is implemented with a quasi-continuous wavelet transform algorithm [13, 18]. While one can think of using the wavelet transform directly for recognition purposes, the synchrosqueezed wavelet transform, among other things, provides us with an alternative to the traditional spectrogram. The latter can also be used for recognition, after further processing via more conventional means such as the computation of the (wavelet based) cepstra - the wastrum [16].

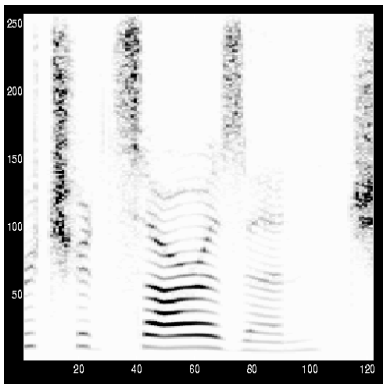
Figure 1 compares the time-frequency representation obtain by wavelet-based synchrosqueezing and *FFT* spectrograms for a segment of speech from the Wall-Street-Journal data base. It is apparent that besides the role of the window sizes, the synchrosqueezed approach extracts coherent structures within the signal, while the *FFT* method represent the harmonics independently of the mutual interferences. For this reason, *primary components* [15] and formants can be efficiently and robustly tracked.

3 Auditory nerve representations

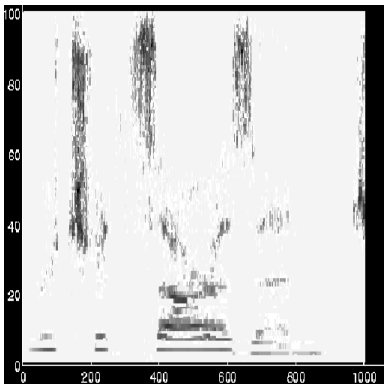
Detailed descriptions of the human peripheral auditory system can be found in [22, 8, 6, 2, 10, 1, 11]. The *EIH* representation results from an attempt to exploit the insynchrony phenomena observed in neuron firing patterns which contain all the information processed by the higher auditory system stages[1]. In general, auditory nerve representations can be modeled as filter banks followed by a dominant frequency extractor. The latter is used to accumulate information from the different subbands along the frequency axis at a given instant of time. The wavelet-based *synchrosqueezed representation* naturally formalizes these models. The cochlear filter bank is approximated by the *QCWT* [18] and the second stage is obtained with the time-derivative of the phase of the wavelet transform as the dominat frequency estimator.[2]

[1] Similar models have been proposed earlier on: the instantaneous-frequency distribution (*IFD*) [7] and the insynchrony bands spectrum (*SBS*) [9].

[2] The wavelet transform is complex is this context. The generating analysis wavelet is a Morlet wavelet [13].



(a)



(b)

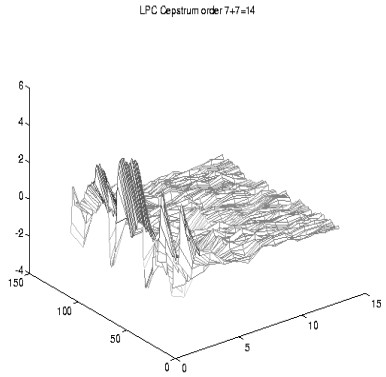
Figure 1: Time-frequency plane for *Richard Sarazen....* (a) presents the *FFT* spectrogram with shift of 10 ms and Hamming windows of 25 ms. (b) illustrates the corresponding synchrosqueezed plane obtained with the method cited in the text.

4 Cepstrum and *ASR*

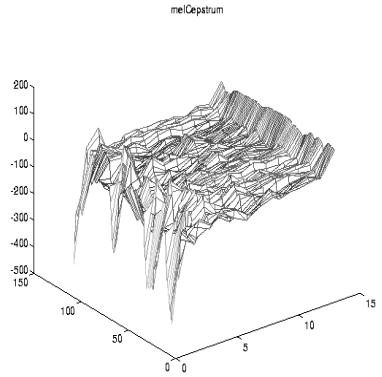
Cepstral parameters are, at present, widley used for efficient speech and speaker recognition. Basic details and justifications can be found in [5, 8, 22, 21]. Originally introduced to separate the pitch contribution from the rest of the vocal cord and vocal tract spectrum [20], the cepstrum has the additional advantage of approximating the Karhunen-Loéve transform of speech signal. This property is highly desirable for recognition and classification [14]. Furthermore, as discussed in [16, 12], the cepstrum can be seen as explicit functions of the formants and other *primary components* of the modulation model. Two main classes of cepstrum extraction have been intensively used:

- *LPC*-derived cepstrum.
- *FFT* cepstrum.

In *ASR* the second approach has become dominant usually with *Mel*-binning. Figure 2 compares *LPC* derived ceptra to *Mel* cepstra for the same segnent of speech as used in Figure 1.



(a)



(b)

Figure 2: The sentence is *Richard Sarazen....* (a) illustrates the *LPC*-derived cepstrum. (b) presents the corresponding *Mel* cepstrum.

5 Wavelet-derived cepstra and wastra

The wavelet transform can be used in different ways to extract cepstral features.

5.1 Energy accumulation-derived cepstra

- **Wavelet binning:**
  Mel frequency binning of the pseudo-frequency and amplitude estimated from the raw wavelet transformed and resulting cepstra are used as features for recognition.
- **Energy binning in synchrosqueezed plane:**
  Same as the above, but now instead of using the estimates from raw wavelet transform, we use the data from the synchrosqueezed time-frequency plane.

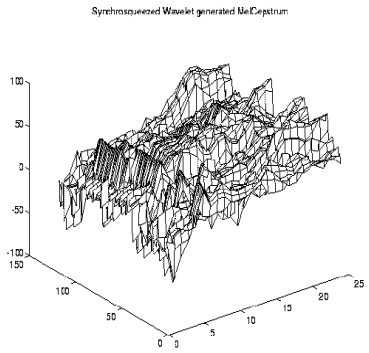


Figure 3: The sentence is *Richard Sarazen....* It illustrates the resulting *MEL* energy binning wastrum.

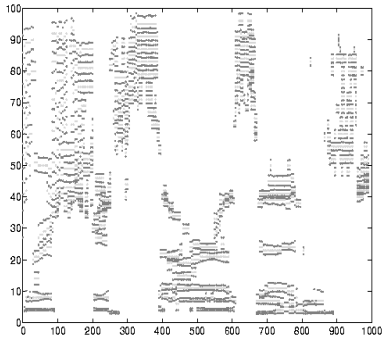
5.2 Time-frequency derived cepstra (wastra)

The wastra (Wavelet based Cepstrum, for short) is introduced in [12, 16] as the cepstral feature obtained by applying Schroeder formula [23] on generalized poles obtained by tracking the formants or *primary components* in the synchrosqueeezed plane. It may be remarked that the improved robustness to noise of synchrosqueeezed wavelet transform for speaker identification has been reported in [15, 16, 4]. Another outcome of the new technique is that the synchrosqueeezed wavelet transform is more amenable to tracking of formants or, more generally, the components of the speech signal. Different methods can be envisioned for tracking of the components. In [16, 19], a *MLE* is described to track formants and primary components. The algorithm is extremely robust but time-consuming.

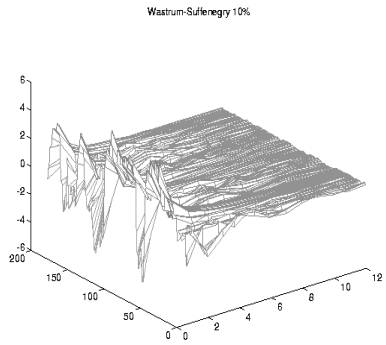
Alternatively, proposed simpler and computationally tractable schemes, which has the flavor of carrying out (K-means) clustering of the synchrosqueeezed spectrum dynamically in time can also be proposed.

- **K-means Wastrum**
  The components are dynamically tracked via a K-means clustering algorithm from the synchrosqueeezed plane. The amplitude, frequency and bandwidth of each of the components are, thus, extracted. The cepstrum generated from this information alone is referred to as the K-mean Wastrum. Figure 4 shows the resulting center frequencies and bandwidths and the resulting cepstrum.

- **Formant based wastrum**
  The K-mean clustering is post processed to limit the set of primary components to formants. Formants are interpolated in unvoiced regions and the contribution of unvoiced turbulant part of the spectrum are added. This method requires adequate formant tracking. The resulting robust formant extraction has numbers of applications in speech processing and analysis.



(a)



(b)

Figure 4: The sentence is *Richard Sarazen....* (a) illustrates the components extracted by the K-mean approach. (b) presents the resulting K-mean wastrum.

6 Speech recognition experiment

To demonstrate the efficacy of the wavelet based synchrosqeezed technique in automatic recognition of speech we consider 20 hours of read speech sampled at the rate of 16 Khz from the wall-street-journal database. We computed the energy binning syncrhrosqeezed wavelet cepstrum (described in Section 5.1) corresponding to a frame rate of 10 ms and a frame size of 25 ms. The cepstrum was then used for decosing the 40 test sentences from the wall-street-journal database. To test the performance of the algorithm in presence of noise, we mixed the clean test signal with cafeteria noise a noise levels from very noisy (10db SNR) to relatively clean (60 db SNR). The results are tabulated in Table 6. The drop of recognition rate with increase in noise level is also diagramatically shown in Figure 5. q Note that training was performed on clean uncorrupted signal for the purpose of these experiments. An obvious way to further improve these results is to train on noise corrupted training data at an appropriate SNR level. Further tuning of the parameters such as the window size and frame rate appropriate for this spe-

Table 1: Word Error rate (WER) as a function of SNR

WER	10	12.5	15	20
SNR	57.04	41.99	27.7	18.2
WER	25	35	40	60
SNR	12.6	10.84	10.43	10.08

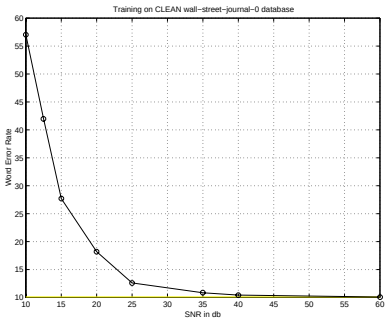


Figure 5: Word recognition error plotted against SNR. Training on clean WSJ database. Test data was contaminated with cafeteria noise.

cific front end processing are also necessary for improved performance. In our experiments these parameters were chosen to be the same as the best known values for FFT based cepstra. In view of these, the present study can only be considered to be preliminary. Since these prelminary results reported in Table 6 and Figure 5, seem to be encouraging, further work is warranted for drawing definitive conclusions on the robustness of wavelet based synchrosqeezed cepstrum.

7. REFERENCES

[1] J. B. Allen. Cochlear modeling. *IEEE ASSP Magazine*, 2(1):pp. 3–29, January 1985.

[2] P. Calliope. *La parole et son traitement automatique.* Masson, 1989.

[3] I. Daubechies. *Ten lectures on wavelets.* Number 61 in CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia, PA, 1990.

[4] I. Daubechies and S. Maes. Wavelets in medicine and biology. In A. Alroubi and M. Unser, editors, *A nonlinear squeezing of the continuous wavelet analysis based on auditory nerve models.* CRC Press, April 1996.

[5] J. R. Deller, J. G. Proakis, and J. H. L. Hansen. *Discrete-time processing of speech signals.* Macmillan, New York, NY, 1993.

[6] J. L. Flanagan. *Speech analysis synthesis and perception.* Springer-Verlag, 1972.

[7] D. H. Friedman. Instantaneous-frequency distribution vs. time: an interpretation of the phase structure of speech. In *IEEE Proc. ICASSP*, pages 1121–1124, 1985.

[8] S. Furui. *Digital speech processing, synthesis and recognition.* Marcel Dekker, New York, NY, 1989.

[9] O. Ghitza. Auditory nerve representation criteria for speech analysis/synthesis. *IEEE Trans. ASSP*, 6(35):pp. 736–740, June 1987.

[10] O. Ghitza. Auditory models and human performances in tasks related to speech coding and speech recognition. *IEEE Trans. Speech and Audio Proc.*, 2(1):pp. 115–132, January 1994.

[11] S. Greenberg, editor. Representation of speech in the auditory periphery. *Journal of Phonetics*, 16(1), January 1988.

[12] S. Maes. *The wavelet tranform in signal processing, with application to the extraction of the speech modulation model features.* PhD thesis, Université Catholique de Louvain, Louvain-la-Neuve, Belgium, June 1994.

[13] S. Maes. Fast quasi-continuous wavelet algorithms for analysis and synthesis of 1-D signals. *preprint to appear in SIAM J. Applied Math.*, April 1995.

[14] S. Maes. LPC Techniques and some new ear-model-based front-ends. Lecture Notes of internal course on Speech Recognition. IBM T. J. Watson Research Center, Human Language Technology, Hawthorne, NY, October 1995.

[15] S. Maes. The wavelet-derived synchrosqueezed plane representation yields a new time-frequency analysis of 1-D signals, with application to speech. *preprint submitted to ACHA*, April 1995.

[16] S. Maes. The wavelet-derived synchrosqueezed plane representation yields new front-ends for automatic speech recognition. *preprint submitted to IEEE Trans. Speech and Audio Processing*, April 1995.

[17] S. Maes. Robust speech and speaker recognition using instantaneous frequencies and amplitudes obtained with wavelet-derived synchrosqueezing measures. In *Program on Spline Functions and the Theory of Wavelets*, Montreal, Canada, March 1996. Centre de Recherches Mathématiques, Université de Montréal. invited paper.

[18] S. Maes. Signal analysis and synthesis with 1-D quasi-continuous wavelet transform. In *Proc. 12th. International Conference on analysis and optimization of systems*, Paris, June 1996. IRSIA.

[19] S. Maes and T. Hastie. The maximum-likelihood-estimation-based living cubic spline extractor and its application to saliency grouping in the time-frequency plane. *preprint to be submitted*, April 1996.

[20] A. M. Noll. Short-time spectrum and 'cepstrum' techniques for vocal-pitch detection. *J. Acoustic Soc. Amer.*, 36(2):pp. 296–302, 1964.

[21] A. V. Oppenheim and S. W. Schaffer. *Digital signal processing.* Prentice-Hall, Englewood Cliffs, NJ, 1975.

[22] L. Rabiner and B-H. Juang. *Fundamentals of speech recognition.* Prentice Hall, Englewoods Cliffs, NJ, 1993.

[23] M. R. Schroeder. Direct(nonrecursive) relations between cepstrum and predictor coefficients. *IEEE Trans. ASSP*, 29:pp. 297–301, 1981.

[24] M. Vetterli and J. Kovačević. *Wavelets and subband coding.* Prentice Hall, Englewood Cliffs, NJ, 1995.