https://doi.org/10.3176/lu.1998.3.09

UNTO K. LAINE (Espoo)

SPEECH ANALYSIS BY USING A NOVEL, BLOCK RECURSIVE ALGORITHM FOR AUDITORY SPECTROGRAMS (BRASS)

1. Introduction

Many aspects of speech sounds are related in an interesting way to the properties of human hearing. The average sound pressure level of speech locates just in the middle of the dynamic range of the hearing (60 dB). The auditory system has the maximum sensitivity around 1.5 kHz, close to frequencies where the second formant, which is perceptually the most important one, is mainly located. In general the frequency selectivity of hearing follows well the frequency characteristics of speech. In the low-frequency area the hearing is able to follow individual harmonics and in the higher frequencies the formants and formant clusters. Due to the increasing acoustical losses the average formant bandwidths increase towards the higher frequencies where the critical bands are wider, too. These are just examples of a long list of features which clearly indicate that speech is "designed" for auditory perception and auditory perception is "designed" for speech. Their evolution has occurred "hand in hand".

On this basis it is understandable that many attempts have been done to apply auditory modelling and different aspects of human auditory mechanisms to speech analysis (Karjalainen 1985). The most popular line has been the use of auditory filterbanks where the gamma tone filterbank has reached the most wide attention (Patterson 1976). The envelopes of the impulse responses of the gamma tone filters follow the gamma function which AM-modulates the "tone" which in its turn defines the center frequency of the actual auditory filter. This shape of the impulse response is closely related to the reverse correlation analysis of the data gained by the studies of the cochlea of a cat (de Boer 1969). T. Irino (1996) has shown that the gamma chirp filter is superior to the gamma tone when evaluated by its time-frequency selectivity. The channel impulse responses of the block recursive filterbanks used in this study have close to gamma chirp characteristics.

In the gamma tone filterbank the individual channels are designed from the same mathematical equation individually. While a good correspondence to the auditory system is achieved the design does not take into account any properties related to the filterbank as a whole, e.g., how the total filterbank response behaves and, how to get an orthogonal filterbank where the channel responses do not correlate. These properties may be secondary aspects in many applications. However, they may be important when the bank should give relevant time-frequency information with minimal redundancy about the analyzed sounds. It is also desirable that the total magnitude of the filterbank (all channels combined) is flat. When talking about the aspects between the channels we have used the concept of synergy. A well designed filterbank does not only have the required time-frequency properties but also a high synergy between the individual channels. Thus channels do not only work well individually but also together to form an unified, well-organized bank.

Computationally the gamma tone filterbank is quite simple. However, in order to follow the auditory spectrogram the outputs of all channels have to be calculated for each input sample. This is computationally relatively expensive especially if we take into account that the time resolution of human hearing is around one millisecond. In other words, the filterbank outputs are needed only once per millisecond or once per every 48 samples if the sampling rate is 48 kHz. Thus one of the 48 outputs is really used and the rest 47 results just update and prepare the filterbank for the next "true" output.

On more aspect related to all types of filterbanks is the amount of redundant data created. If the filterbank has 14 channels and all these are sampled at every input sample the bank creates 14 times more "information" than what was fed in. The true amount of information cannot increase like that so the redundancy is increased by a factor of 13. In order to avoid this problem filterbanks are created so that so-called critical sampling is achieved. This means that every output channel is critically sampled down by a sampling rate twice as high as the bandwidth of the channel in question. This causes two more problems: in nonuniform resolution filterbank every channel has its own sampling frequency which increases the complexity of the system, and secondly, after down sampling with 14 different frequencies the channels will loose their synchrony and it is very difficult to process them simultaneously so that the necessary time-synchronous inter-channel information is revealed. Especially in speech analysis synchronous processing and visualization of the channels seems to be important.

These four essential aspects of filterbanks, namely, the optimal time-frequency selectivity, the high level of channel synergy, the critical or almost critical down sampling with full inter-channel synchrony, and the computational efficiency guided the way toward a design of a novel block recursive auditory filterbank (Laine 1997). When this algorithm is used to generate Auditory SpectrogramS Block Recursively we talk about BRASS method (or even instrument!).

This paper reviews shortly the background of the BRASS method and shows by examples how the method works in different areas of speech processing.

2. The BRASS algorithm

The BRASS method has evolved through a technique called frequency warping. By applying the frequency warping method the uniform resolution Fourier analysis can be transformed to a nonuniform resolution form. Thus the BRASS method can be seen as a frequency-warped Fourier analyzer (Laine, Härmä 1996). The shape of the frequency-warping function defines the mapping between the uniform Hz scale to the new nonuniform auditory scale. Typically Hz-to-Bark or Hz-to-ERB-rate mappings are used. The BRASS method of this paper applies the ERB-rate scale.

The BRASS algorithm realizes an auditory ERB-rate-scale filterbank and gives a new auditory spectra vector at every Mth signal sample, where the value of M can be freely chosen. The input to the algorithm is a M-vector of signal samples and the output a N-vector of filterbank outputs. A new output vector is computed recursively from the previous one by adding to the recursive part the "novelty" given by the new input M-vector. This can be formulated in the following way:

$$S_{m+1}(\nu) = \mathbf{A}\, S_m(\nu) + \mathbf{B}\, \mathbf{s}_m, \quad (1)$$

where $\mathbf{s}_m$ is the M-vector at the time index m, $S_m(\nu)$ is the filterbank output at the same time index (the auditory spectrum, variable n denotes the channel index), **A** is the recursion matrix and **B** the spectral state control matrix. The whole filterbank design problem is focused on the optimal design of these two matrices. This topic is discussed in more detail in Laine 1997.

All the analyzed signals were sampled at 22.05 kHz rate. Fourteen complex valued channels were realized. The bank is computed after every fourteen samples thus a oversampling of factor two is used. The channel frequencies are: 130, 260, 390, 560, 775, 1030, 1380, 1810, 2370, 3100, 4005, 5210, 6675, and 8610 Hz. They follow closely the ERB-rate scale, however the bandwidths are almost two times broader than in the corresponding gamma tone bank.

The broader bandwidths were applied in this study in order to reach high time resolution. The bandwidths are still narrow enough for the observations of the movements of the two, three lowest formants. With this reduced frequency resolution the bank gives correspondingly higher time resolution than the human ear (around 0.65 ms). This allows to monitor even pitch-synchronous effects and to observe the time-frequency properties of single pitch periods []. A nice feature of the BRASS design is that the number of channels can be freely chosen. The added channel responses always form a flat frequency response and their center frequencies follow closely the desired frequency scale. Thus the user has the full control over the allocation of the time-frequency resolution.

3. A preliminary test

In order to test the BRASS algorithm and to get a preliminary idea of the auditory time-frequency distribution of the glottal excitation five synthetic glottal pulses were analyzed. They were produced using a simple polynomial model of the form g(t) = c $(t^2 - t^3)$ followed by a proper set of zeroes modelling the closed phase. Figure 1 illustrates the auditory spectrogram. An increment in the x-dimension corresponds to 0.64 ms. The sharp glottal closure is the main part of the excitation. Its frequency conteint is distributed over all frequencies. The latency (time delay) between the channels is clearly seen. The auditory spectrogram models closely the delay of the traveling wave in the cochlea. The high frequency part of the cochlea is excited first and the low frequency part later. The maximum latency is about 8 ms.

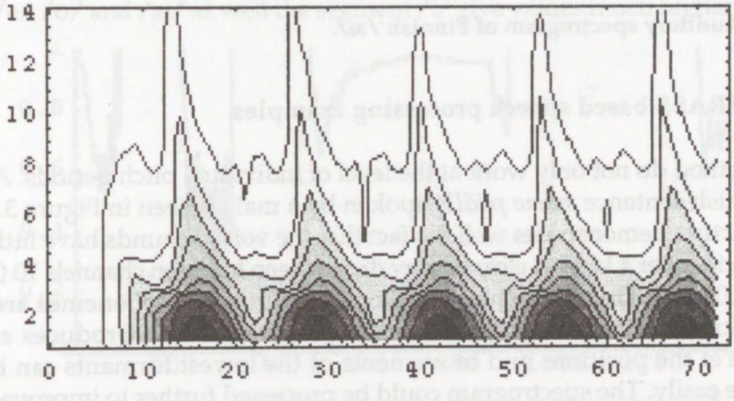

Figure 1. **Auditory spectrogram of five synthetic glottal pulses.**

Another interesting detail in the Figure 1 is the quite large "bubble" in the low frequency area. This corresponds to the open period pulseform which also means that this part ends at the glottal closure.

4. Analysis of Finnish /æ/

Figure 2 illustrates an auditory spectrogram of Finnish vowel /æ/ where five glottal closures and openings are displayed. The high temporal resolution of the chosen BRASS method is well demonstrated in this figure. At each closure especially the second and the third formant are strongly excited (channels 7—10). The first formant at channel 5 (775 Hz) produces a long pattern in time. At about the middle of the closure patterns some hints of the secondary excitation caused by the glottal opening can be seen. At the same instant the second formant is strongly damped. The position and shape of the secondary excitation fluctuates more than the sharp excitation of the closure.

The low frequency "bubble" locates now somewhat higher than in the synthetic case. However, it still precedes the main excitation at the glottal closure. Sometimes this "bubble" is similar to a pattern produced by an excited resonator. This led to a hyphotheses of a possible subglottal resonance. The closer analysis showed that even if some other methods also show some resonator-type of behavior around these frequencies, the main reason to this "bubble" is the shape of the glottal waveform during the open period. A more detailed analysis of this phenomena will be published in Laine 1998.

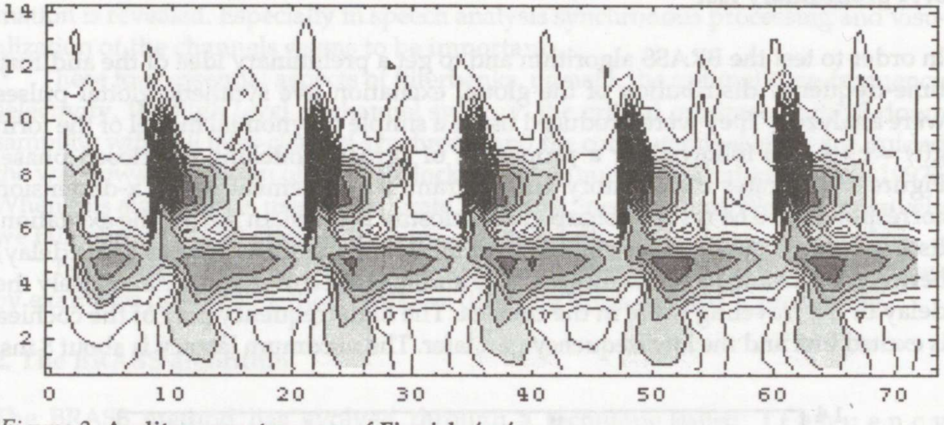

Figure 2. **Auditory spectrogram of Finnish /æ/.**

5. Other BRASS based speech processing examples

BRASS method do not only work at the level of individual pitch periods. An example of Finnish sentence *memo päälle* spoken by a male is seen in Figure 3.

The picture demonstrates well the fact that the voiced sounds have little energy in frequencies over 4 kHz. A clear energy drop is seen between channels 10 (3100 Hz) and 11 (4005 Hz). Due to the high temporal resolution the phonemes are seen as vertical blocks in the spectrogram. Every glottal excitation produces a vertical line, too. Yet the positions and movements of the lowest formants can be monitored quite easily. The spectrogram could be processed further to improve the clarity of formant related information.

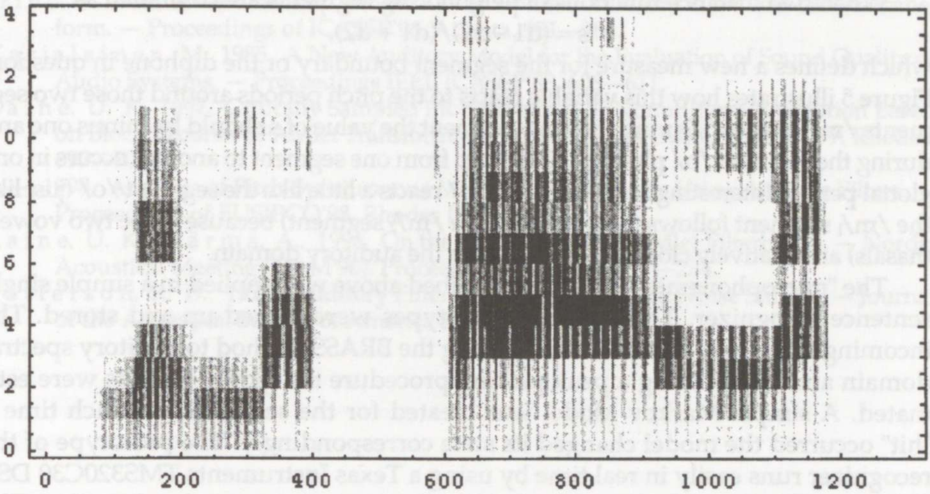

Figure 3. **Auditory spectrogram of the Finnish sentence** ***memo päälle*****.**

5.1. Experiments with BRASS based speech recognition and automatic segmentation

The clearness and the high temporal resolution of the produced auditory spectrogram led to a question if this method could be applied to a simple speech recognizer or to find automatically the phoneme boundaries (automatic segmentation). In the following a couple examples of these applications are given.

One pitch period of the second /m/$_2$ phoneme in the sentence of Figure 3 is picked up to represent the whole phoneme segment. This model period is a 14 × 14 matrix taken from the auditory spectrogram. When this model is compared to all of the other periods in the sentence by computing the Euclidean (time-frequency) distance between the model period and the unknown periods, the distance function of the Figure 4 is created. The function is normalized so that zero corresponds to the zero distance (the two patterns are equal) and one corresponds the comparison to a zero matrix.

Since the model pitch period was picked up from another recording of the same sentence, the distance function does never reach a complete fit (the zero value). However, over the whole sentence it has the lowest values just during the corresponding segment of /m/$_2$. Note that the first /m/$_1$ segment will also give quite low values. The two /m/ segments differ so much that different models are needed for each of them. Vowels /e/, /o/ and /æ/ as well the segment /l/ give values much higher than one.

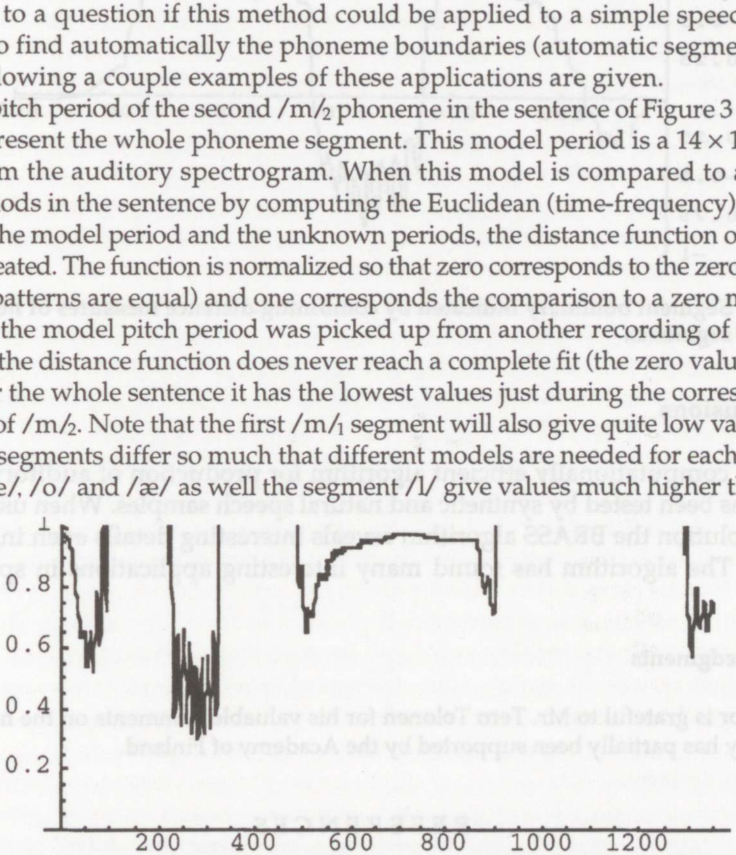

Figure 4. **Normalized Euclidean distance function between the model pitch period and all the other pitch periods (or equivalent subrframes) of the sentence (see text).**

When two distance functions of neighboring segments are combined as

$$s = (d1 - d2)/(d1 + d2),$$

which defines a new measure for the segment boundary or the diphone in question. Figure 5 illustrates how this variable reacts to the pitch periods around those two segments /m/$_2$ and /o/. During the /m/$_2$ segment the value of s should be minus one and during the segment /o/ plus one. The jump from one segment to another occurs in one glottal period. Interestingly, the segment /e/ reacts a little like the segment /o/ (just like the /m/$_1$ segment follows the pattern of the /m/$_2$ segment) because those two vowels (nasals) are relatively close to each others in the auditory domain.

The "microphonemic" approach described above was applied in a simple single sentence recognizer. The segment prototypes were picked up and stored. The incoming signal was transformed by using the BRASS method to auditory spectral domain and by applying a peak picking procedure the glottal periods were estimated. A simple Markov Model was created for the sentence and each time a "hit" occurred the model changed its state correspondingly. The prototype of the recognizer runs easily in real time by using a Texas Instruments TMS320C30 DSP processor. The processor uses about 60% of its capacity to run the algorithm. The prototype has been preliminary tested with promising results.

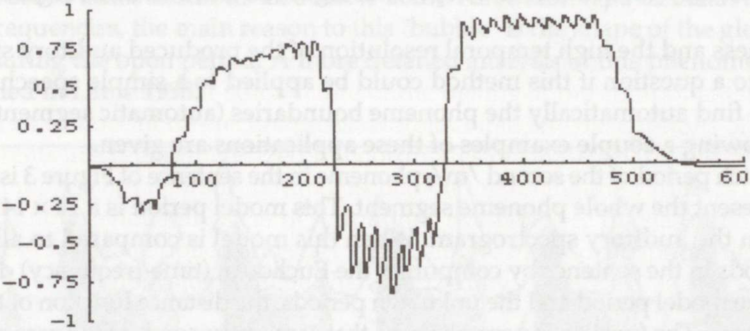

Figure 5. **Segment boundary indicated by combining distance measures of neighboring segments.**

6. Conclusions

A novel, computationally efficient algorithm for production of auditory spectrograms has been tested by synthetic and natural speech samples. When using a high time-resolution the BRASS algorithm reveals interesting details even inside pitch periods. The algorithm has found many interesting applications in speech processing.

Acknowledgments

The author is grateful to Mr. Tero Tolonen for his valuable comments on the manuscript. This study has partially been supported by the Academy of Finland.

REFERENCES

de Boer, E. 1969, Encoding of Frequency Information in the Discharge Pattern of Auditory Nerve Filters. — International Audiology 8, 547—556.

Irino, T. 1996, A Gammachirp Function as an Optimal Auditory Filter with Mellin Transform. — Proceedings of ICASSP'96, Atlanta, 981—984.

Karjalainen, M. 1985, A New Auditory Model for the Evaluation of Sound Quality of Audio Systems. — Proceedings IEEE ICASSP'85, 608—611.

Laine, U. K. 1997, Critically Sampled PR Filterbanks of Nonuniform Resolution Based on Block Recursive FAMlet Transform. — Proceedings of EUROSPEECH'97, Rhodes, 697—700.

—— 1998, Analysis of Pitch-Synchronous Modulation Effects by Using Analytic Filters. — Proceedings of EUSIPCO'98, Rhodes (to be published).

Laine, U. K., Härmä, A., 1996, On the Design of Bark-FAMlet Filterbanks. — Nordic Acoustical Meeting (NAM'96). Proceedings, Helsinki, 277—284.

Patterson, R. D., 1976, Auditory Filter Shapes Derived with Noise Stimuli. — Journal of the Acoustical Society of America 59, 64—654.