A BIOLOGICAL APPROACH TO SPEECH SPECTROGRAPHY

Huawei Colin Li and Willy Wong
Sensory Communications Lab
University of Toronto

1. INTRODUCTION

A spectrogram is a display of the frequency content of a signal drawn so that the energy content in each frequency region and time is displayed on a coloured scale. In the case of speech signal, the spectrogram has been widely used to investigate the features of speech in the time-frequency plane that are perceptually important. For instance, more than fifty years ago, the spectrogram was already being used to determine the average frequencies of the first three formants of the vowels of American English for men, women and children [1]. The spectrogram is also widely used in speech analysis for developing speech recognition systems [2].

The spectrogram today can be generated digitally using the Short Time Fourier Transform (STFT). However, spectrograms implemented using STFT suffer from the unavoidable tradeoff between time and frequency resolution, also known as the uncertainty principle of signal analysis [3]. The emerging recognition of the existence of frequency and amplitude modulations in speech on the order of a pitch cycle, which is referred to as “fine structure”, has motivated many researchers to develop new spectrogram representations to obtain better time-frequency resolution. A filter-bank method called fine structure spectrogram (FSS) introduced in [4] is shown to be able to detect the fine structure in the modulation patterns of speech, not seen by other methods before.

Although FSS is successful in utilizing some of the functional components of a real auditory system, there are other properties, which have not yet been implemented in its approach. In this paper, the fine structure spectrogram is first presented, modifications to build a more biological FSS are then introduced, and finally the results from both original and modified versions of FSS are compared to each other and in relation to the real output as processed by an actual auditory modeling system.

Manuscript received in February 2007. Huawei Colin Li is a Masters candidate with the Sensory Communications Laboratory in the Institute of Biomaterials and Biomedical Engineering and Department of Electrical and Computer Engineering, University of Toronto, ON, Canada (e-mail: huawei.li@ utoronto.ca).

2. FINE STRUCTURE SPECTROGRAM

Structure

The fine structure spectrogram (FSS) is an enhanced realization of the conventional spectrogram model. Instead of representing the local spectral content of a signal, it tracks the instantaneous frequency of a modulated component of speech using a filter-bank. It can be constructed using numerous overlapping filter/detectors. Each filter/detector consists of a bandpass filter, followed by a rectifier and smoother with a detector of local peaks in the outputs of all the filter/detectors. All peaks at one time instant can be utilized to produce the instantaneous frequency spectrum and then be used to produce FSS over the time duration of the speech signal.

Filter/Detectors

Inspired by the thousands of tuning characteristics of the afferent auditory fibers, thousands of filter/detectors (F/D’s) are used to break the signal into frequency components, and then pick the local peaks to detect the instantaneous frequency at each time instant. There are four stages in each F/D: filtering, rectifying, smoothing, and local peak picking of all F/D’s at the end, as shown in Fig. 1(a). The speech signal at one time instant is first decomposed into its different frequency components through a bandpass filterbank.

The filterbank is shown in Fig. 1(b). It consists of thousands of overlapping wide-band bandpass filters spaced equally apart in frequency with a constant bandwidth. Each frequency component obtained through the filters is then passed through a rectifier to extract its relative energy content. The resulted signals (an example is shown in Fig. 1c) are smoothed out before being sent to the last stage, the peak detector, which picks out local energy peaks to track the instantaneous frequency spectrum of the speech signal.

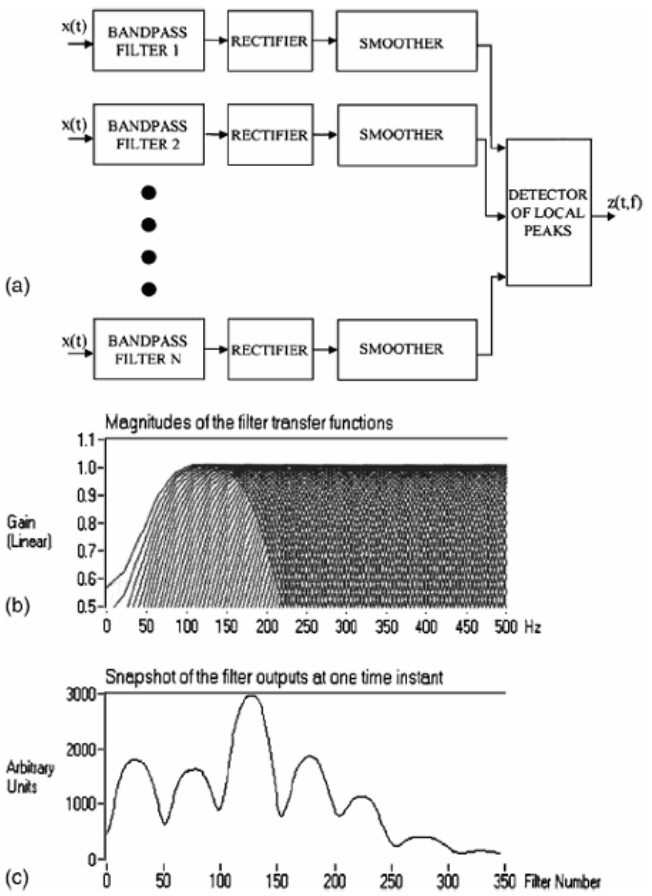

Fig. 1. (a) Basic structure used to generate the fine structure spectrogram (FSS). Hundreds or thousands of F/D's are used. (b) The magnitudes of the filter transfer functions in one implementation where the center frequencies were separated by 5 Hz. (c) A plot of the rectified and smoothed filter outputs, at a given instant in time. The analyzed signal was the syllable /ba/. Adapted from [4].

Spectrogram Implementation

A spectrogram can be implemented from the displaying of instantaneous frequency spectrum throughout the duration of the signal, with the energy content scaled by intensity. FSS detects the fine structure of the modulated patterns in speech signal. It is also able to detect significant time-frequency modulations not seen with other methods before. Following the literature in [4], the model was reconstructed and coded in MATLAB. The attempt in modifying this model is described in the following section.

3. BIOLOGICALLY INSPIRED MODIFICATIONS

Overview

Although FSS shows success in utilizing some properties of the auditory system, it also differs from known properties of a real auditory system in two major ways: (1) bandpass filters have constant bandwidths, and are (2) spaced equally across the frequency range. This is in contrast to the work of gammatone filterbanks by Patterson *et al.* [5] as applied in [6][7]. A Gammatone filterbank is a standard model of cochlear filtering [8]. It provides non-uniform bandwidths and non-uniform spacing of the center frequencies following the characteristics of the auditory system. Our attempt to give FSS a more biological design follows a similar approach.

Filters Spacing

In the FSS model, filter spacing is kept uniform across all frequencies, while the nature of auditory system suggests otherwise. Non-uniform spacing is observed in turning curve characteristics as shown in [10]. In fact, the turning curves are distributed quasi-logarithmically across frequency. In correspondence, the spacing of bandpass filters should be distributed quasi-logarithmically across frequency. Here, an approximated logarithmically spaced model extracted from [11] is used to observe the effect on FSS with minimal invasiveness. The center frequency F_c of each bandpass filter is determined as:

$$F_c(n) = 237.5 \cdot 1.029^n - 144.4$$

n represents the n^{th} filter of the filterbank distributed from 100Hz to 4000 Hz.

Filter Bandwidths

We also note that the bandwidths of the filters are not uniform in the auditory system. The turning curves along the basilar membrane, which corresponds to the bandpass filters in the filterbanks, actually behaves such that the bandwidths increase with frequency [9]. In another sense, the auditory system loses frequency resolution towards higher frequencies, but gains time resolution in return. A similar approach to the Equivalent Rectangular Bandwidth (ERB) given in [5] is adopted here, given the bandwidth B(f) in Hz:

$$B(f) = a \cdot f + 24.7$$

a is 0.1039 in [5], it is doubled to 0.2078 here to provide a more apparent effect on FSS that is relatively easy to observe.

COMPARISON AND DISCUSSION

Overview

To observe the effect upon the modifications in the filterbank, various versions of the spectrogram were generated with a test signal for one time frame (one time instant). 0^{th} order and 1^{st} order time-frequency relations are sufficient as an approximation to provide a general consensus on the modification effects. Therefore, the testing signal used is the sum of single tone sinusoid (0^{th} order) and a chirp tone (1^{st} order).

$$input = \sin(\omega \cdot t) + \sin((k \cdot t) \cdot t),$$
$$\omega = 2\pi \cdot 1000\, Rad/s, k = 2\pi \cdot 1000\, Rad/s^2$$

While speech, a significantly more complex signal, can be seen in part as a combination of 0^{th} order (vowels) and 1^{st} order (glides) tones.

In our implemented version, FSS is simplified with non-overlapping windows, and the number of filters is reduced to 100. The change introduces negligible differences at our level of comparison.

Original FSS

The oversimplified FSS provides reasonably accurate and precise information on the test signal as shown in Fig. 2(a). A smearing, however, does occur at the crossing of the chirp and the single tone. The smearing is caused by the interference between the adjacent overlapping bandpass filters close to the cross point frequency and is commonly found even with other standard techniques. The intensity of the smearing can however be significantly reduced with multiple time frames, although this is not a matter that concerns this paper.

Non-uniform filter spacing

When the filters are distributed logarithmically but the bandwidths are set to be uniform across frequency, the resulted spectrogram on the testing signal is shown in Fig. 2(b). As the separation of filters increases with frequency, a "staircase" figure is formed due to quatization of the filters in the frequency space. The "staircase" observation agrees with the frequency discriminations behavior in the auditory system. In [12], it stated that the value of threshold in frequency difference required to discriminate a change in frequency increases as frequency increases. Moreover, the variation of the widths of the adjacent stairs can give clues to the auditory perceiver as to the slope of the rising chirp.

Non-uniform filter bandwidths

While a "staircase" can be seen in the neural response of the auditory system, one does not hear such "staircase" in real life. The reason lies in the non-uniformity of the bandwidths. The filters are more separated as frequency increase, at the same time the bandwidth of the filter increases. The effects of increasing bandwidth can be seen in Fig. 2(c). The increased bandwidth at high frequencies has the effect of improving temporal resolution. The decrease in spectral resolution behaves in essence as a frequency smoother.

Non-uniform filter spacing and bandwidths

The modified FSS with both non-uniform spacing and bandwidths is shown in Fig. 2(d). As frequency increases, the spectral resolution is sacrificed in gaining temporal resolution. This induces a smearing in frequency, which acts as a smoother to reduce the "staircase" effect. Due to the picking of local energy peaks, this smoothing is not as apparent, but a smearing range, which widens with increase in frequency can be observed.

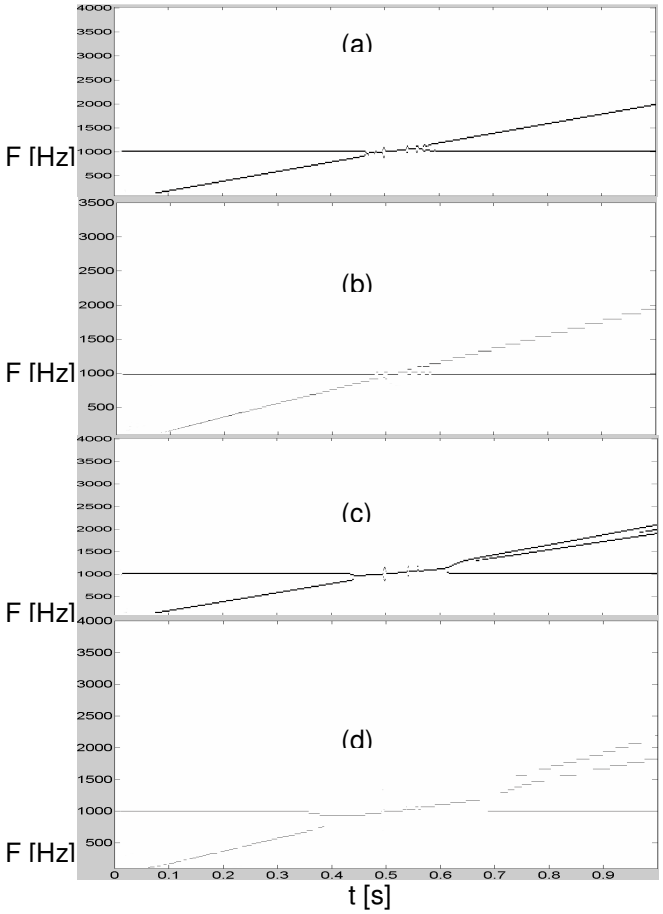

Fig. 2. Outputs of the original FSS and modified FSSl. (a) Original FSS output. (b) Non-uniform spacing (c) Non-uniform bandwidths (d) Non-uniform spacing and bandwidths.

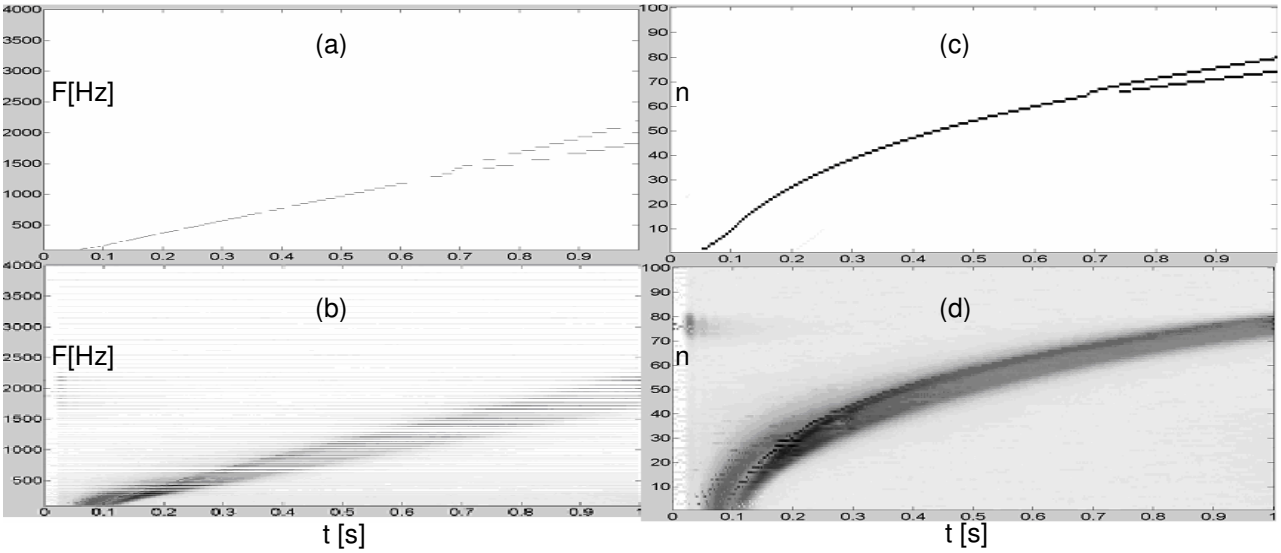

Fig. 3. Outputs of modified FSS and auditory nerve fibers for a rising chirp input signal. (a) Modified FSS output (b) Spike count of the auditory nerve fiber response. (c) Modified FSS output with frequency channel vs. time (c) Spike count of the auditory nerve fiber response with frequency vs. time. Note: The frequency channels separation increase logarithmically. n represents the n^{th} channel.

Development System for Auditory Modeling (DSAM)

We are also in a position to compare the output of the modified FSS with the output of a real auditory system. We use the well-known Development System for Auditory Modeling (DSAM) [13]. The Development System for Auditory Modeling (DSAM) is a computational library designed specifically for producing simulations of the auditory system. It brings together many established auditory models, produced by various research groups, under a flexible programming platform. The particular application of interest is the Auditory Modeling System (AMS), which produces responses at different levels of the biological auditory system.

Since the single tone sinusoid provides limited information in the difference between the original and modified FSS, it is removed from the testing case. We use only the chirp portion of the test signal:

$$input = \sin\left((k \cdot t) \cdot t\right), k = 2\pi \cdot 1000\, {Rad}/{s^2}$$

Showing in Fig. 3(a) is the output from the modified FSS, with non-uniformity for both bandwidths and spacing. As a comparison, the envelope of the spike count in the auditory nerve fiber response of AMS is extracted and shown in Fig. 3(b). "Staircase" and smoothing effects can be observed in both figures with a little discrepancy due to the peak picking of local energy contents for the modified FSS. The two figures are plotted again in terms of the filters in Fig. 3(c)(d). Since the filter spacing is non-uniform, the resulted outputs appear in logarithmic fashion instead of as a straight line, but the great similarity between the two plots can still be easily observed.

CONCLUSION

Although modified FSS creates more complexity in comparison with the original, it is more biological plausible and opens up new ways to study signal spectrography in a manner previously not possible. The human ear is a remarkably versatile and sensitive spectrum analyzer and hence new signal processing techniques may benefit from further study of biology.

REFERENCES

[1] G.E. Peterson, H. L. Barney, "Control methods used in a study of the vowels," J. Acoust. Soc. Am. Vol. 24, pp. 175-184, 1952

[2] B.E.D Kingsbury, N. Morgan, S. Greenberg, "Robust speech recognition using the modulation spectrogram" *Speech Communication.* Vol. 25 No. 1, pp. 117-132, 1998.

[3] J.W. Pitton, K. Wang, B. H. Juang, "Time-frequency analysis and auditory modeling for automatic recognition of speech" *Proc. IEEE* Vol. 84 No. 9, pp. 1199-1215, 1996

[4] H. Dajani, W. Wong, H. Kunov "Fine structure spectrogram and its application in speech" J. *Acoust Soc. Am.* Vol. 117, No. 6, pp. 3902-3918, 2005

[5] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C.Zhang, and M. H. Allerhand, "Complex sounds and auditory images," In *Auditory Physiology andPerception*, (Eds.) Y Cazals, L. Demany, K.Horner, Pergamon, Oxford, 1992, pp. 429-446.

[6] G. J. Brown and M. P. Cooke, "Computational auditory scene analysis," *Comput. Speech Language*, vol. 8, pp. 297–336, 1994.

[7] G. Hu and D.L. Wang, "Monaural speech segregation based on pitch tracking and amplitudemodulation," *IEEE Trans. Neural Net.*, Vol. 15, pp. 1135-1150, 2004.

[8] E. de Boer and H. R. de Jongh, "On cochlear encoding: potentialities and limitations of the reverse-correlation techniques," *J. Acoust. Soc. Amer.*, vol. 63, pp. 115–135, 1978.

[9] W. A. Yost, “The Neural Response and the Auditory Code” *FUNDAMENTALS OF HEARING* Chapter 9, pp. 131, 2000
[10] S. Coren, L. M. Ward, J. T. Enns, “The Auditory System” *SENSATION AND PERCEPTION* Chapter 6, pp. 164. 1999
[11] Lopez-Poveda, E.A., and Meddis, R. "A human non-linear cochlear filterbank" *JASA* 110, 3107-3118, (2001)
[12] W. A. Yost, “Auditory Sensitivity” *FUNDAMENTALS OF HEARING* Chapter 10, pp. 156, 2000
[13] Development System for Auditory Modeling (DSAM) http://www.essex.ac.uk/psychology/hearinglab/dsam/