COMBINED MULTI-RESOLUTION (WIDEBAND/NARROWBAND) SPECTROGRAM

Shiufun Cheung and Jae S. Lim

Massachusetts Institute of Technology
Cambridge, MA 02139

ABSTRACT

The speech spectrogram, used extensively in speech research, is a two-dimensional time-frequency display of a one-dimensional signal. The existing spectrograms - the wideband spectrogram and the narrowband spectrogram - are deficient either in frequency or in time resolution. In this paper, we present a method to combine the two spectrograms by evaluating the geometric mean of their corresponding pixel values. The combined spectrogram appears to preserve the visual features associated with high resolution in both frequency and time.

1. Introduction

Since the advent of the sound spectrograph in the 1940's [1,2], the spectrogram which contains many visual features that can be related to parameters known to be important to human speech perception has been used extensively in many branches of speech research. In most cases, it is used as a tool to visualize and display speech utterances.

In this paper we will present one approach which leads to a visually improved spectrographic display of speech. Previous efforts to obtain improved time-frequency displays of speech include the development of better time-frequency distributions, such as the Wigner distribution, many of which are detailed in [3], and the development of neural spectrograms which use critical filters in imitation of the human auditory system [4]. The main difference between our approach and those mentioned above is that instead of pursuing the general goal of developing a more efficient time-frequency distribution, we attempt only to preserve the visual features already familiar as representatives of speech without significantly altering the traditional form of the spectrogram representation.

Our method takes an image processing approach to combine the two existing kinds of spectrograms - the wideband spectrogram and the narrowband spectrogram - while preserving their important visual features. In section 2 we will discuss the wideband and narrowband spectrograms and describe the relative advantages of each. The problem of improving the spectrograms and our approach towards its solution will be presented in section 3. In section 4 one example of the combined spectrogram will be shown. Section 5 concludes the paper.

2. Wideband and Narrowband Spectrograms

Generally the spectrogram is obtained by displaying the magnitude of the short-time Fourier transform. The short-time Fourier transform, $X(n, j\omega)$, of a discrete-time sequence, $x(n)$, is a two-dimensional function of time and frequency. It is given by

$$X(n, j\omega) = \sum_{m=-\infty}^{\infty} w(n-m)\cdot x(m)\cdot e^{-j\omega m} \qquad (1)$$

where n represents the discrete time axis and ω represents the continuous frequency axis [5]. The sequence $w(n)$ in the equation is known as the analysis window or, alternatively, the analysis filter.

One of the major shortcomings of the spectrographic display is due to the limitation imposed by the Principle of Uncertainty, so that spectrograms are deficient either in frequency resolution or in time resolution. If the analysis window, $w(n)$, is relatively short, or, when the analysis filter is wideband, time resolution is good while frequency resolution is poor. The resulting spectrogram is known as a wideband spectrogram. Visually the wideband spectrogram is characterized by vertical striations which represent the pitch period and dark horizontal bands which are the formant frequencies.

For a longer analysis window, the frequency resolution improves at the expense of the time resolution. When a relatively long analysis window is used, the resulting narrowband spectrogram has poor time resolution but good frequency resolution. In a narrowband spectrogram, the

CH2977-7/91/0000-0457 $1.00 © 1991 IEEE

pitch period is no longer visible along the time axis, instead the corresponding fundamental frequency will appear as horizontal striations.

Because of the different time and frequency resolution properties of the wideband and narrowband spectrograms, they are typically used for different applications. For example, the wideband spectrogram is valued for its quick temporal response and is generally used for word-boundary location. On the other hand, the narrowband spectrogram is preferred for measuring the fundamental frequency.

3. The Problem and Our Approach

Due to the deficiencies of the spectrograms, it is sometimes necessary to display more than one spectrogram to illustrate different aspects of a single speech utterance. We believe that this inconvenience can be eliminated by combining the wideband spectrogram and the narrowband spectrogram in a manner that preserves the important visual features of both. Even though the approach we propose is developed heuristically, it is simple and appears to be effective.

Among the essential features to be preserved are the horizontal striations of the narrowband spectrogram which provide a simple method for measuring the fundamental frequency, and the quick temporal response of the wideband spectrogram which allows accurate word-boundary detection and formant tracking. Also of interest is the vertical striations of the wideband spectrogram which can sometimes be used for marking the pitch period.

A heuristic but simple method which appears to satisfy our goal of preserving the above visual features is to compute the geometric mean of the wideband and narrowband spectrograms. Specifically, if we use the discrete Fourier transform to compute samples of the short-time Fourier transform with a finite-length analysis window, equation (1) can be reduced to

$$X(n, k) = \sum_{m=-N/2}^{N/2-1} w(m)\cdot x(n\cdot s - m)\cdot e^{-\frac{j2\pi km}{N}} \qquad (2)$$

where n still represents the discrete time axis, but k now represents the discrete frequency axis. The parameter N is the discrete Fourier transform size and the parameter s is the window shift. The spectrograms can now be viewed and displayed as a digital image where each data point, X(n, k), represents the value of one pixel (picture element). [6] If we view the two original wideband and narrowband spectrograms as images, we can also view the combined spectrogram as an image. For each (n, k), the value of the combined spectrogram is given by the geometric mean of the corresponding pixel values of the wideband and the narrowband spectrograms. The combined spectrogram $X_{cb}(n, k)$ is given by

$$X_{cb}(n, k) = \left[\,|X_{nb}(n, k)|\cdot|X_{wb}(n, k)|\,\right]^{1/2} \qquad (3)$$

where $X_{nb}(n, k)$ and $X_{wb}(n, k)$ are the narrowband and the wideband spectrograms respectively.

The reason for our choice of the geometric mean is the preservation of the small values of the short-time Fourier transform magnitude or the valleys of each of the two spectrograms. For example, if either of the corresponding pixels in the two spectrograms is zero, the result in the combined spectrogram is also zero. Consequently, both the horizontal striations and the vertical striations should remain clearly visible in the combined spectrogram. Word boundaries should also be preserved because the edges should remain intact without significant blurring.

4. Example

To illustrate the method discussed in section 3, we have chosen the sentence "These shoes were black and brown". by a male speaker. The speech utterance is sampled at 8 kHz. Figure 1(a) is a wideband spectrogram of this utterance. The analysis window used is a 64-point Hamming window and is given by

$$w_{wb}(n) = 0.54 - 0.46\cdot\cos\left(\frac{2\pi(n+32)}{63}\right) \qquad -32 \le n \le 31 \quad (4)$$

The discrete Fourier transform size N is 512 points and the window shift s is 24 points. As expected, the pitch period is clearly visible through vertical striations in the figure.

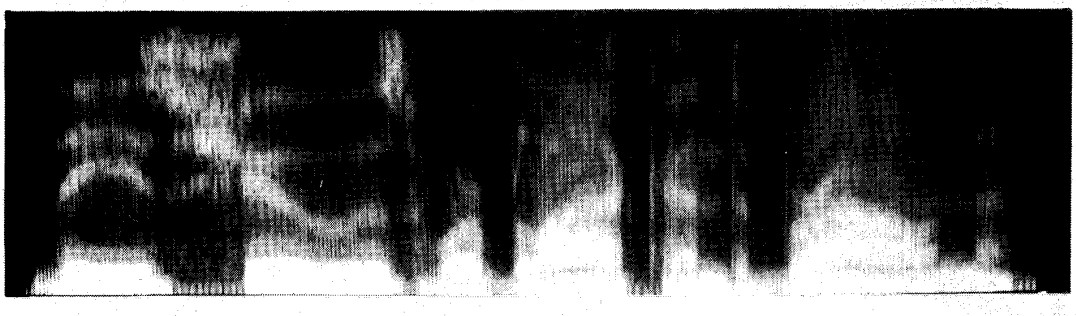
The narrowband spectrogram of the same speech utterance is shown in figure 1(b). A 256-point Hamming window is used as the analysis window and is given by

$$w_{nb}(n) = 0.54 - 0.46\cdot\cos\left(\frac{2\pi(n+128)}{255}\right) \qquad -128 \le n \le 127. \quad (5)$$

The horizontal striations appear in the figure as expected. Notice that the two analysis windows in equations (4) and (5) have been centered around the origin. This ensures that the two spectrograms are synchronized in time before they are combined.

The combined spectrogram obtained in the manner described by equation (3) is shown in figure 1(c). As was described earlier, both the horizontal and vertical striations are clearly visible. The formants are not blurred significantly and word boundaries remain sharp.

(a)



(b)

(c)

Figure 1. (a) A wideband spectrogram of the speech utterance "These shoes were black and brown", using a 64-point Hamming window

(b) A narrowband spectrogram of the same speech utterance using a 256-point Hamming window

(c) A combined spectrogram using the geometric-mean merge on (a) & (b)

The speech utterance is sampled at 8 kHz.

It should be noted that in figures 1(a), (b) and (c), A nonlinearity has been applied to the spectrograms prior to their display. Specifically, the short-time Fourier transform magnitudes are first mapped linearly onto a zero-to-one scale. It is then followed by a nonlinear mapping given by

$$y = x^{1/4} \qquad (6)$$

where x is the input intensity and y is the output intensity. The nonlinear mapping is used to enhance the visibility of the important freatures in the spectrograms The enhanced values are then mapped linearly onto 256 grey levels with lower values corresponding to the darker levels and higher values to the brighter ones. This is different from the convention normally used in spectrograms but is nonetheless consistent with the general practice in image processing. [7] This method is applied to the spectrograms only after the combined spectrogram has been computed from the unenhanced values of the two original spectrograms.

5. Conclusion

In this paper, we discussed a specific method to combine a wideband spectrogram and a narrowband spectrogram. The general approach to view a spectrogram as an image can lead to a number of other methods for combining spectrograms of different time-frequency resolutions. For example, we can combine the spectrograms by first mapping the narrowband spectrogram to one color and then the wideband spectrogram to another color. It is also possible to extend the geometric-mean merge to combine more than two spectrograms with different time-frequency resolutions. These and other related methods to improve the visual appearance of a spectrogram is currently under investigation.

References

[1] R. Koenig, H. K. Dunn, and L. Y. Lacey, "The sound spectrograph,"*J. Acoust. Soc. Am.*, vol.18, pp.19-49, 1946

[2] R. K. Potter, G. A. Kopp, and H. G. Kopp, *Visible Speech*, New York: Dover Publications, 1966

[3] L. Cohen, "Time-frequency distributions - a review," *Proc. of the IEEE*, vol.77, no.7,pp 941-981, July 1989

[4] D. Klatt, "Speech Processing Strategies based on Auditory Models," *in CG*, pp.181-196, 1982

[5] S. H. Hawab and T. F. Quatieri, "Short-time fourier transform," in *Advanced Topics in Signal Processing*, (J. S. Lim and A. V. Oppenheim, eds.), ch6, Prentice Hall, 1988

[6] A. V. Oppenheim, "Speech spectrograms using the fast Fourier transform," *IEEE Spectrum*, vol. 7, pp.57-62, August 1970

[7] J. S. Lim, *Two-dimensional Signal and Image Processing*, ch 8, Prentice Hall, 1990