RESTORATION OF SPEECH SIGNALS CONTAMINATED BY STATIONARY TONES USING AN IMAGE PERSPECTIVE

Blanca I. Andía

Research Associates for Defense Conversion (RADC), Rome, NY 13441

ABSTRACT

Instead of using the traditional approach of attenuating the speech frequencies which are contaminated by tones, the project uses image processing techniques to do an intelligent restoration of the contaminated frequencies. The first step is to obtain the fundamental frequency patterns which are repeated periodically throughout the image spectrogram. These patterns are obtained by using cepstral estimation. The following step is to restore the frequencies affected by the stationary tones making use of the shape and location of frequency patterns found in the previous step. The concept of Markov Random Fields (MRF) used in estimation problems of image processing is used for this restoration phase.

1. INTRODUCTION

Typically, speech that is contaminated by tones goes through a tone removal process that consists of two phases: detection and attenuation. While this process as a whole provides reasonable results and is used widely in the field, there is still room for improvements [1]. These improvements could be made in the attenuation phase of the process. The attenuation phase attenuates the frequency components of the speech signal that correspond to the frequencies of the tonal interference. This operation can either produce a hole in the spectrum or leave some tone in the signal. Both have undesirable effects on the reconstructed speech signal.

The idea behind this research is to replace the attenuation phase of the tone-removal process by a restoration phase, i.e. the regions of the spectrogram of the contaminated speech signal that are contaminated by tones will be restored. While the majority of research and practical efforts in tone removal processes have been directed at tone detection, very little attention has been given to the restoration of the speech affected by tones [1]. The approach that this paper takes to solve the problem comes from an image processing point of view, in which the spectrogram is an image which will be restored in the areas affected by the tone or tones. The idea is that if a spectrogram appears to be closer visually and quantitatively to the original, it will provide a better quality of interference-free speech as an output. Figure 1 and Figure 2 show an outline of the current process and the proposed process respectively.

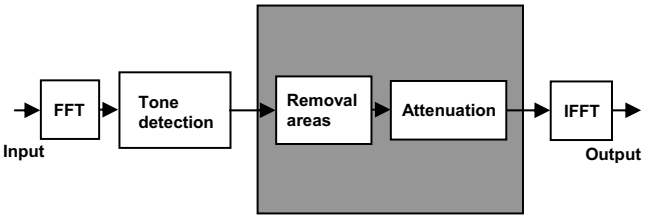

Figure 1. Flow diagram for the current process of tone removal.

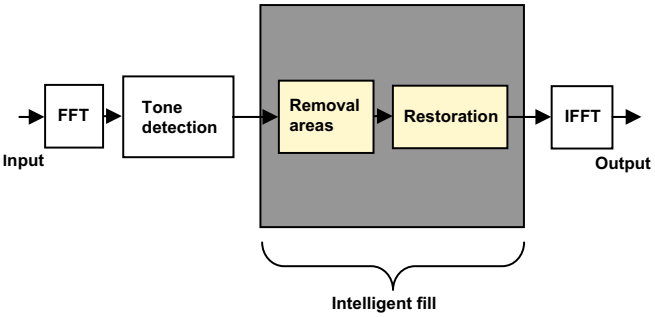

Figure 2. Flow diagram for the proposed process of tone removal.

The restoration process consists of two steps. The first step is the detection of the fundamental frequency patterns that appear throughout the spectrogram [2]. Since the other patterns in the spectrogram are just multiples of the fundamental frequency patterns, it is very easy to reproduce them from the knowledge of the fundamental frequency patterns. The process of detecting the fundamental frequency patterns will be explained in Section 2.

The second step consists of removing the areas of the spectrogram which have been contaminated by a tone and then restoring these affected frequencies by using a Markov Random Field (MRF). The MRF models the missing section by allowing the information about the shapes and locations of the patterns calculated in Section 2 to be

incorporated into the values which are being restored. Section 3 will explain in detail the process of restoring the speech using an MRF which is based on the shapes and values of the frequency patterns in the spectrogram. Finally, Section 4 will provide conclusions and future directions for the image processing solution to the tone removal process that has been outlined in this paper.

2. FINDING PATTERNS IN THE SPECTROGRAM

The spectrogram is an image whose structure contains a repetition of certain patterns through its frequency range. Since speech consists of voiced and unvoiced segments, these patterns are only present in the time frames where voiced speech is present. By contrast, time frames which correspond to unvoiced speech do not contain any such patterns [2].

Voiced speech is characterized by the periodicity in its waveform called the pitch period. In the frequency domain, this pitch period is referred to as the fundamental frequency. The collection of the fundamental frequencies for the voiced time frames creates the fundamental frequency patterns. These patterns and its multiples characterize the structure of the image spectrogram.

To find the fundamental frequency patterns, a pitch estimation technique is used. One should be careful in choosing a proper technique for estimating the pitch, since the speech signals considered for this application contain stationary tones. Therefore, the algorithm can be confused as to whether the period of the tone or the period of the speech is the correct pitch for a particular time frame. In particular, for unvoiced or silence frames, where no speech pitch is present, the algorithm would choose the pitch of the tone as the correct pitch for that particular time frame. In order to overcome this problem, a preliminary step that detects unvoiced speech and silence is required [3]. In this manner, when the algorithm finds an unvoiced or a silence frame, it does not calculate a pitch period for that particular frame.

For this application, a speech frame is classified as unvoiced or silence when its energy is -10dB below the average energy of all the speech frames in an utterance. For all the other speech frames, i.e. the voiced speech frames, the cepstrum technique is used to estimate their pitch [4]. Other techniques for estimating pitch, such as autocorrelation, get confused by the presence of the tone.

For each of those voiced speech frames, the cepstrum is computed. If the speech signal at a determined time frame k is referred to as $s_k(t)$ and its Fourier transform as $S_k(\omega)$, the cepstrum $c_{Sk}(t)$, can be calculated by:

$$c_{Sk}(t) = \Im^{-1}[\log | S_k(\omega) |] \quad (1)$$

where, log corresponds to the logarithm in base 10, | | corresponds to the magnitude and $\Im^{-1}$ to the inverse Fourier transform.

The cepstrum is calculated for each voiced speech frame that forms the utterance. Then, the peak cepstral value is found for each those frames, and the pitch period is determined as the location of the maximum peak, for each of the speech frames [3]. The collection of all the pitch periods for all the time frames forms the pitch pattern. However, in this paper, one is interested in finding the patterns that form the structure of the spectrogram. Therefore, the pitch pattern must be transformed to a fundamental frequency pattern. This is done by computing an inverse of each of the values that form the pitch pattern and multiplying them by the size of the frame.

For our applications, the spectrogram of the tone-contaminated speech utterances will be formed by taking frames of 1024 samples, i.e. if the speech is sampled at 8kHz the frames will be 128ms long. Consecutive frames are taken with a 75% overlap between them, then each frame is multiplied by a Hanning window of 1024 samples. Finally the windowed speech frame is transformed to the frequency domain via the Fourier transform.

Figure 3 shows the spectrogram of a TIMIT utterance generated by the procedure mentioned above. The utterance has been contaminated by a stationary tone which is 10dB above the rms level of the speech and located at 240Hz. This utterance corresponds to a female speaker. From the spectrogram, it is clear that the 240Hz tone erases most of the fundamental frequency pattern. We are interested in recovering this fundamental frequency pattern as well as calculating the intensity values for each "pixel" in the image spectrogram that is contaminated by the tone. Before estimating the pixel values that correspond to the tone frequencies, the tone is removed from the spectrogram. Removal of the tone implies giving a value of zero to the pixels which contained the tone. In the rest of the paper, these pixels will be referred to as the "missing pixels".

To find the fundamental frequency pattern we use the method explained in this section. Figure 4 shows the fundamental frequency patterns and their first three harmonics superimposed to the spectrogram of the utterance. Notice that the cepstrum technique detects the fundamental frequency patterns perfectly in the frames that correspond to voiced speech. The sections in which this estimation errs correspond to some of the unvoiced frames of the speech utterance. In those frames, as predicted, the algorithm chooses the frequency of the tone to be the correct frequency of the spectrogram. However, the pitch or fundamental frequency for those frames should be set to zero, since in theory the concept of pitch is not existent for unvoiced or silence frames.

3. RESTORATION OF CONTAMINATED AREAS

The restoration process of the areas which have been contaminated by a tone will use the knowledge of the

shapes and intensity values of the fundamental frequency patterns and their harmonics. For that matter, a Markov

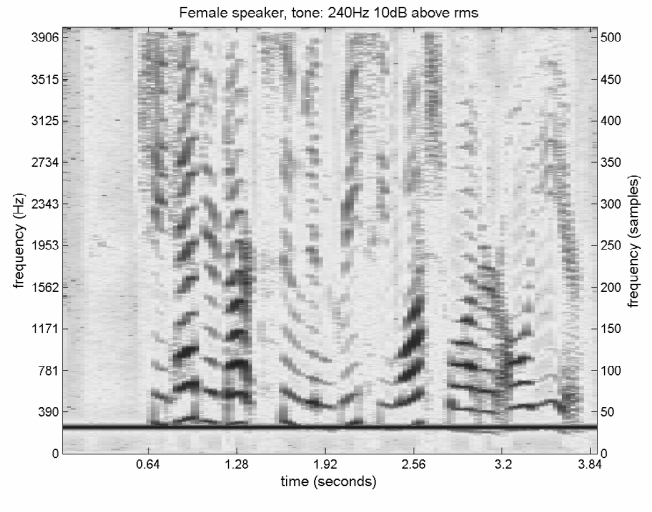

Figure 3. Spectrogram of a TIMIT speech utterance contaminated by a stationary tone at 240 Hz.

Random Field (MRF) is chosen as an appropriate image model for the spectrogram. Its parameters allow it to incorporate certain behaviors among local groups of pixels in the restored image. An MRF is commonly used because of its ability to model successfully both the smooth regions and the discontinuities of the images. By using an MRF whose penalty function is convex, the estimation problem results in a convex optimization problem which can be solved by iterative methods [5].

Let **z** be the image that we want to restore. An MRF is used to describe its prior probability, p(**z**), where,

$$p(\mathbf{z}) \propto \exp[-\lambda \sum_{c \in C} \rho(\mathbf{d}_c^t \mathbf{z})]. \quad (2)$$

In (2), the parameter λ is a regularization parameter, *c* are the indices of local groups of pixels called cliques and *C* is the overall set of those cliques. The $\mathbf{d}_c^t$ terms form spatial activity measures, to allow a degree of similarity or dissimilarity between local groups of pixels in the cliques. The term $\rho(.)$ is the penalty function which controls how heavily the spatial activity measures are penalized. For this application, the convex penalty function $\rho(.)$ is chosen as,

$$\rho(u) = u^2. \quad (3)$$

This paper uses a Bayesian approach for estimating **z**. In particular, the Maximum a Posteriori (MAP) solution to estimating **z** is found by maximizing its posterior distribution given the observations **x**. Namely, the solution is found by maximizing $p(\mathbf{z}|\mathbf{x})=p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$. The observation vector **x** refers only to the interference-free data of the spectrogram. By the nature of the problem, the observations x(k,f) are noise free, i.e. x(k,f)=z(k,f) for all (k,f) that is not an interference region. Then, from this fact $p(\mathbf{x}|\mathbf{z})$ is equal to 1 for all x(k,f)=z(k,f) and is equal to 0 when some x(k,f)≠z(k,f); where k refers to the time frame and f to the frequency bin.

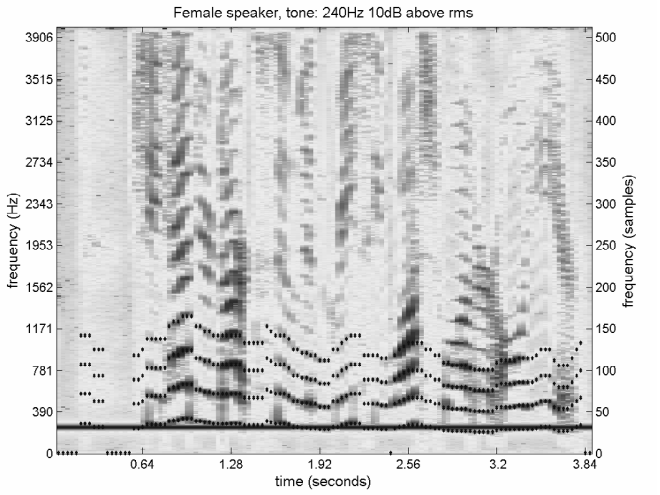

Figure 4. Spectrogram of TIMIT utterance contaminated by stationary tone at 240 Hz. The fundamental frequency pattern and its first three harmonics are superimposed to the spectrogram and shown in dotted lines.

Since we are only interested in the case where $p(\mathbf{x}|\mathbf{z})=1$, the problem is reduced to maximizing only p(**z**) for the values of z(k,f) located in the tone-contaminated regions while maintaining z(k,f)=x(k,f) unchanged for the regions which are interference free. Maximizing the probability p(**z**), is equivalent to minimizing the negative of its natural logarithm. Then, the image estimate is the solution to,

$$\hat{\mathbf{z}} = \arg\min_{\mathbf{z}} [\lambda \sum_{c \in C} \rho(\mathbf{d}_c^t \mathbf{z})] \quad (4)$$

The solution to this convex minimization problem is determined iteratively based on a gradient descent algorithm, which starts with an initial estimate and updates the value of the image, **z**, for every subsequent iteration [5].

Equation (4) is used to estimate the section of the spectrogram which was contaminated by the tone and is currently set to zero. The $\mathbf{d}_c^t \mathbf{z}$ terms and the $\rho(.)$ function of (2) play a very important role in determining the final image result. The terms $\mathbf{d}_c^t \mathbf{z}$ are picked such that they carry the information about the shape and intensity of the frequency patterns to the missing pixels. The collection of vectors $\mathbf{d}_c$ is a set of masks used to establish which pixels are interrelated and in what manner. In this application, for every missing pixel, three masks are used. These masks are shown in Figure 5.

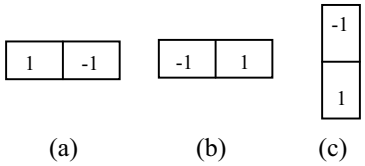

Figure 5. Masks used for the construction of $\mathbf{d}_c^t\mathbf{z}$.

The equations corresponding to these three masks are the following,

$$\mathbf{d}_c^t\mathbf{z} = z(k,f) - z(k+1, f-\Delta_a(k)) \text{ for (a)} \quad (5)$$

$$\mathbf{d}_c^t\mathbf{z} = z(k,f) - z(k-1, f-\Delta_b(k)) \text{ for (b)} \quad (6)$$

$$\mathbf{d}_c^t\mathbf{z} = z(k,f) - z(k, f+F(k)) \text{ for (c)} \quad (7)$$

where $\Delta_a(k)$ is the difference between the fundamental frequency at time frame k and the fundamental frequency at frame k+1, namely F(k)–F(k+1). Analogously, $\Delta_b(k)$ = F(k)–F(k-1). The presence of $\Delta_a(k)$ and $\Delta_b(k)$ in the $\mathbf{d}_c^t\mathbf{z}$ terms in (5) and (6) ensure the preservation of the shape of the fundamental frequency patterns observed in the spectrograms. The $\mathbf{d}_c^t\mathbf{z}$ term in (7), takes in consideration the pixel located one fundamental frequency above the pixel that is being calculated. This ensures that the proper intensity values are estimated for the missing pixels. Notice that an extra mask reciprocal to Figure 5(c) can be added, to consider the pixel one fundamental frequency below the "missing" one. Also the $\rho(.)$ function of (3) penalizes the $\mathbf{d}_c^t\mathbf{z}$ terms quadratically, and it ensures continuity and smoothness between the pixels in the masks of Figure 5.

Figure 6 shows the TIMIT utterance described in Section 2 after it has been restored using an MRF. The section of the spectrogram occupied by the tone is now a restored image which is visually significantly better than that if the tone were just attenuated. Most importantly, when the spectrogram is converted back to a speech waveform via an inverse Fourier transform and an ovelap-add, there is an absence of residual tone and a consequent better quality sounding speech. Even though Figure 6 offers a significant improvement over pure attenuation of the tone, there still is room for improvement. For example, the recovered fundamental frequency pattern could have more defined edges by modifying the $\rho(.)$ function of (3).

4. CONCLUSIONS

This paper demonstrated a novel way of restoring tone-contaminated speech data by using an image processing perspective. The method first detects the shapes of pattern that form the spectrogram structure and then uses them in forming a Markov Random Field as a model for the replacement of the contaminated section of the spectrogram. Visual results of the restored spectrogram are very encouraging as compared to most common tone suppressing algorithms. This in turn translates into a superior quality of the reconstructed speech. Future work involves varying the $\rho(.)$ and $\mathbf{d_c}$ parameters of the MRF to accommodate for sharper edges of the reconstructed patterns as well as situations where multiple stationary tones are present.

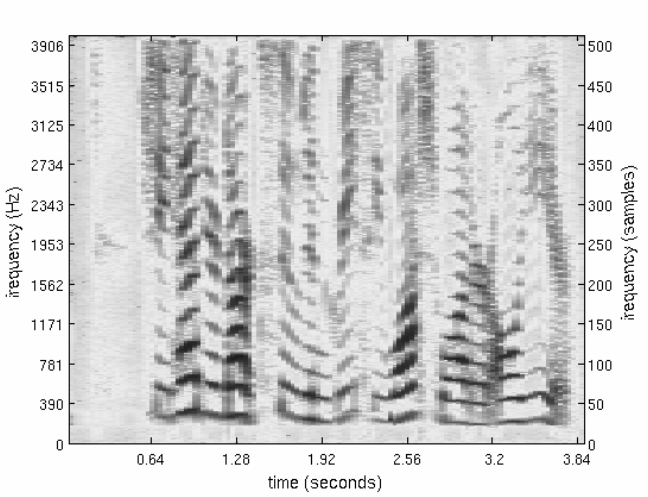

Figure 6. Restored spectrogram of TIMIT utterance in Figure 4. The use of an MRF was instrumental in preserving the pattern shapes which were contaminated by the tone.

5. REFERENCES

[1] Research Associates for Defense Conversion (RADC) Interim report AFRL-IF-RS-TR-2003-100, "Information Extraction for Military and Law Enforcement Applications", pp. 30-38 and 97-99, Rome, NY, May 2003

[2] Lawrence R. Rabiner and Ronald W. Schafer, *Digital Processing of Speech Signals,* Prentice Hall, Englewood Cliffs, New Jersey, 1978

[3] Lawrence R. Rabiner, M.J. Cheng, A.E. Rosenberg and C.A. McGonegal, "A Comparative Performance Study of Several Pitch Detection Algorithms," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-24, No. 5, pp. 399-418, Oct. 1976.

[4] John R. Deller, Jr., J.G. Proakis and J.H.L. Hansen, *Discrete Time Processing of Speech Signals*, Prentice Hall, Upper Saddle River, New Jersey, 1987.

[5] Thomas P. O'Rourke, R.L. Stevenson, "Improved Image Decompression for Reduced Transform Coding Artifacts," *IEEE Transactions on Circuits and Systems for Video Technology,* Vol. 5, No.6, pp. 490-499, December 1995.