RECOGNITION OF SPEECH ENHANCED BY BLIND COMPENSATION FOR ARTIFACTS OF SINGLE-SIDEBAND DEMODULATION

Mark J. Harvilla and Richard M. Stern

Department of Electrical and Computer Engineering
Carnegie Mellon University, Pittsburgh, PA 15213 USA
{mharvill, rms}@cs.cmu.edu

ABSTRACT

This paper concerns the automatic recognition of speech that has been distorted by frequency shifting introduced by a transmitter-receiver frequency mismatch in communications systems using single-sideband (SSB) modulation. The degradation in recognition accuracy depends both on the frequency shift induced by mistuned SSB and additive noise, with a reduction in SNR causing the degradation produced by mistuned SSB to become more profound. We consider the performance of a method for detecting frequency shifts introduced by SSB; the shifts can be corrected easily if identified correctly. The proposed method provides accurate estimates of SSB-induced frequency shifts over a wide range of SNRs if at least approximately 80 seconds of speech is available. The use of the algorithm provides almost-complete amelioration of the effects of mistuned SSB even for utterances shorter than 10 seconds, and signal restoration is expected to improve for utterances of longer duration.

Index Terms— Robust speech recognition, speech enhancement, nonlinear distortion, single-sideband modulation

1. INTRODUCTION

Systems that attempt to transcribe long-range operational point-to-point communications signals, including commercial and military transmissions, must deal with the effects of a wide variety of linear and nonlinear distortions. This paper concerns the automatic recognition of speech that has been distorted by frequency shifting introduced by a mismatch between the carrier frequencies of transmitting and receiving oscillators of a single-sideband (SSB) modulator-demodulator pair. We begin by describing the mathematical origins of the frequency-shift effect and its impact on speech recognition accuracy. We then describe and characterize the performance of a method that has been developed to detect blindly and compensate for the frequency shift, as well as the extent to which these approaches are successful in restoring the speech recognition accuracy obtained from the compensated signal.

Although analog modulation techniques including SSB have been superseded by digital techniques in many communication systems in developed countries, the use of SSB remains common in other parts of the world. Channels that are degraded by SSB frequency shift are included by the DARPA Robust Automatic Transcription of Speech (RATS) program in the simulations of degraded speech signals that are important to military intelligence.

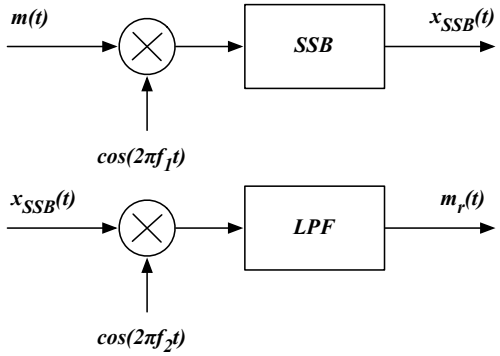

Fig. 1. Single-sideband modulation and demodulation.

2. ORIGINS OF SSB DISTORTION

Single-sideband (SSB) modulation and demodulation has been popular in point-to-point communication systems because it reduces the bandwidth of the modulated signal by a factor of two compared to conventional double-sideband modulation.

While there are multiple ways to implement an SSB signal, the basic procedures for ideal SSB modulation and demodulation are summarized in simplest form in Fig. 1. We assume that the message signal to be transmitted, $m(t)$ is band limited to f_M Hz. The message signal $m(t)$ is first multiplied by an oscillator at frequency f_1 and then passed through an ideal bandpass filter with passband $f_1 - f_M \leq |f| \leq f_1$. As a result, the modulated signal $x_{SSB}(t)$ has frequency components only between $f_1 - f_M$ and f_1 Hz, the passband of the ideal SSB filter. The modulated signal can be expressed as

$$x_{SSB}(t) = m(t)\cos(2\pi f_1 t) + \hat{m}(t)sin(2\pi f_1 t) \qquad (1)$$

where $\hat{m}(t)$ is the Hilbert transform of $m(t)$. Demodulation is easily accomplished by multiplying the modulated signal by a local oscillator of frequency f_2 and passing the signal through an ideal lowpass filter with cutoff frequency f_M Hz. If $f_1 = f_2$ and all the filtering is ideal, the output signal $m_r(t)$ will equal the original message $m(t)$. Unfortunately, maintaining frequency synchrony between the modulator and demodulator oscillators can be difficult in

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. D10PC20024. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of DARPA or its Contracting Agent, the U.S. Department of the Interior, National Business Center, Acquisition & Property Management Division, Southwest Branch. We also thank Dan Ellis for making the RATS renoiser tool available.

practice, and if $f_1 \neq f_2$ the output signal $m_r(t)$ will equal $m(t)$ but with the positive frequency components of $m_r(t)$ shifted upward by $f_\Delta = f_2 - f_1$ and the negative frequency components of $m(t)$ are shifted downward by f_Δ. This produces a characteristic distortion in the audio that becomes more pronounced as $|f_\Delta|$ increases, leading to speech that can sound like the familiar cartoon character Donald Duck. If this frequency shift is detected and estimated correctly, it is easy in most cases to restore the original audio signal by performing a compensatory modulation and demodulation with a deliberate frequency mismatch of $-f_\Delta$. (This fails only when a large negative f_Δ causes the lowest frequency components of $m(t)$ to go through zero and interfere with one another.)

The DARPA RATS Program began in 2011 with the objective of developing methods to provide automatic speech recognition in realistic highly-degraded channels. The data used for system development and evaluation consist of careful simulations of operational audio signals, and some of these channels exhibit frequency shifts. Fig. 2 is a histogram of the estimated values of f_Δ for utterances from one such simulated operational channel.

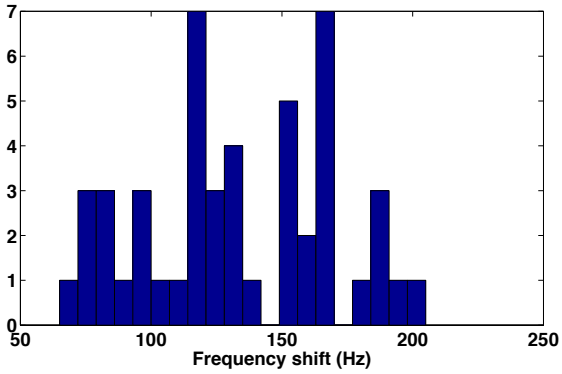

Fig. 2. Distribution of frequency shifts in selected channel of the DARPA RATS data.

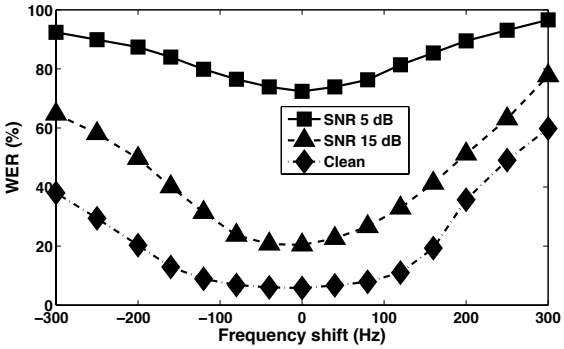

Fig. 3. Effects of SSB-induced frequency shifts on speech recognition accuracy for clean speech and speech degraded by street noise with SNRs of 5 and 15 dB.

2.1. Impact of SSB-based frequency shifts on speech recognition accuracy

As noted above, the distortion produced by frequency shifting increases as f_Δ increases, and interacts with other factors such as additive noise in a nonlinear fashion. Fig. 3 provides some sample error rates obtained using a version of the CMU Sphinx-III system with a weakened language model, applying frequency shifts in the manner described in Section 2 to sentences from the DARPA RM1 database in the manner described in Section 4. Results are plotted for clean speech and speech degraded by digitally-added street noise. We note, unsurprisingly, that the observed word error rate (WER) increases monotonically as the frequency shift increases and the SNR decreases. There is an interaction between the two types of degradation, as decreasing the SNR causes recognition accuracy to become impaired for frequency shifts of increasingly small magnitude.

3. DETECTION AND COMPENSATION FOR FREQUENCY-SHIFTED SPEECH

3.1. Initial detection based on estimated fundamental frequency and spectral peak locations

Over the years a small number of algorithms have been proposed to detect blindly the values of f_Δ for a particular incoming utterance. In general these algorithms work by estimating the fundamental frequency of voiced segments, typically using cepstum-based techniques, and then observing the actual peaks in the log spectrum. The frequency shift f_Δ is inferred by comparing indirectly the locations of the actual spectral peaks to the locations of spectral peaks that would have been produced by a true harmonic series.

One such algorithm was proposed by Suzuki *et al.* [1]. The Suzuki *et al.* approach begins by estimating the fundamental frequency f_0 using cepstral techniques, computing the inverse DCT of the log of the spectrum. The estimate of f_0 is unaffected by the value of f_Δ because the spectrum of a frequency-shifted signal exhibits the same periodocities over frequency regardless of frequency shift. The estimated fundamental frequency $\hat{f}_0$ is obtained by searching for a maximum in the cepstrum at an appropriate quefrency corresponding to the nominal period of the periodic signal. The peaks of the log spectrum of a frequency-shifted signal will appear at frequencies

$$f_k = k f_0 + f_\Delta \tag{2}$$

where f_k represents the spectral peak corresponding to the correct harmonic, k is the harmonic number, and f_0 is the true fundamental frequency. In practice, f_0 can be estimated fairly accurately because this estimate is based on information from all the harmonics of the signal, but the estimated locations of the individual spectral peaks $\hat{f}_k$ tend to be more errorful. Suzuki *et al.* obtained estimates of f_Δ by plotting the estimated peaks of the log spectrum $\hat{f}_k$ as a function of k, fitting a line to these points using linear regression, and producing an estimated frequency shift $\hat{f}_\Delta$ from the intercept of this line with the vertical axis. Suzuki *et al.* [2] also proposed a second approach in which f_Δ is inferred from the degree of symmetry of the cepstrum in each voiced frame.

3.2. Complete estimation of frequency shift independent of harmonic number

A major shortcoming of the approach of Suzuki *et al.* [1] is that the algorithm is critically dependent on a correct match between the putative harmonic number k and the frequency of each peak in the log magnitude spectrum. This can be difficult, especially in telephone channels in the POTS network which typically have a bandpass frequency response with a low-frequency corner of approximately 300 to 350 Hz. We worked on a variety of techniques to overcome this shortcoming, but eventually stumbled on an algorithm, originally

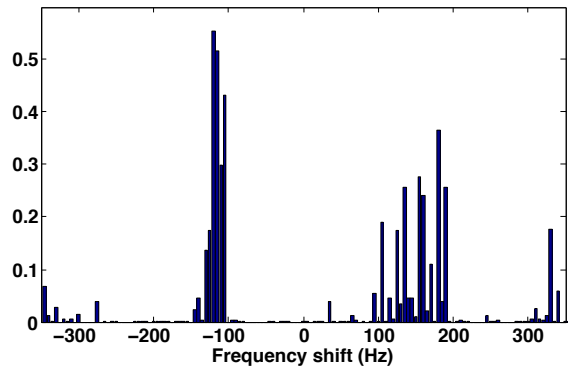

Fig. 4. Histogram of candidate values of $\hat{f}_\Delta$ for an utterance with a true f_Δ of -120 Hz. The peak is located at exactly -120 Hz and the next-largest peaks are closely grouped in frequency.

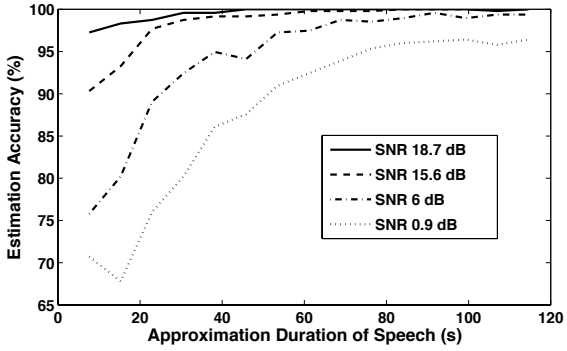

Fig. 5. Percentage of estimates of f_Δ that are within 5 Hz of the correct value as a function of utterance duration.

described by Dick in an obscure technical report in 1980 [3], that requires no *a priori* information about harmonic peak number and which provides a more robust solution. It works by exploiting the usual variations in f_0 during normal, spontaneous speech.

Speech is sampled at 16 kHz and short-time Fourier transforms are computed in the conventional fashion using 80-ms Hamming windows with 50% overlap. For each N-point time frame, the "complex correlation" $C[m]$ is calculated by computing the inverse DFT of the magnitude of the $N/2+1$ positive frequency components of the signal's Fourier transform, padded with $N/2-1$ zeros. The resulting complex-valued signal is referred to by Dick as the complex correlation. (This is similar to the more familiar analytic signal, but is based on the inverse transform of the positive-frequency components of the magnitude of the spectrum, rather than of the original complex spectrum itself.)

An estimate of f_0 is obtained by searching for the peak in the magnitude of $C[m]$ for each time frame, in a similar fashion to obtaining an estimate of f_0 using the magnitude of the complex cepstrum. An estimate of f_0 is obtained by the simple relation

$$\hat{f}_0 = \frac{f_s}{p} \tag{3}$$

where p represents the index in the time domain at which the peak of $C[m]$ is found and f_s is the sampling frequency.

The set of all possible estimates of f_Δ for a given frame is obtained from the real and imaginary parts of $C[p]$ according to the equation

$$\hat{f}_{\Delta_r} = \left(\frac{\hat{f}_0}{2\pi} \arctan \frac{\Im C[p]}{\Re C[p]} \right) + r\hat{f}_0 \tag{4}$$

where the index r denotes one individual possible value of $\hat{f}_\Delta$ within the set. The estimation accuracy of the exact location of the maximum of $C[m]$ is improved using polynomial interpolation, thus allowing the maximum value to lie between two integer values of m; interpolation is similarly used to obtain values of $\Im C[p]$ and $\Re C[p]$ for non-integer values of p.

While Eq. 4 does not have a unique solution, the ambiguity is easily resolved because the correct value of r will lead to an estimate of f_Δ that remains invariant over multiple analysis frames, even though f_0 is constantly changing. In contrast, the other (incorrect) values of r will produce estimates $\hat{f}_{\Delta_r}$ that vary over time. This observation enables us to identify the correct value of r by accumulating a histogram of the multiple frequency estimates $\hat{f}_{\Delta_r}$ on a frame-by-frame basis, each incremented by the square of the value of maximum $C[p]$ for the current frame. Note that in order for Eq. 2 to be meaningful, the signal must be periodic, so we are principally interested only in the voiced frames. Because unvoiced speech is strongly aperiodic, in contrast, the peak of the complex correlation will have a small magnitude in the corresponding unvoiced frames. Thus, incrementing the histogram of $\hat{f}_{\Delta_r}$ by the magnitude of this peak for each time frame is reasonable in that it implicitly takes voicing into account. The final estimate of f_Δ is obtained by simply selecting the largest peak in this histogram of frequency estimates.

Figure 4 shows the histogram of $\hat{f}_\Delta$ estimates obtained over approximately 3.5 seconds of speech for which the true f_Δ is -120 Hz. The (correct) value of -120 Hz is adopted for $\hat{f}_\Delta$ because it is the frequency corresponding to the maximum value of the histogram. As can be seen, other candidate values of $\hat{f}_\Delta$ also appear in the histogram at clusters of frequencies separated by multiples of the fundamental frequency f_0.

Figure 5 describes the sensitivity of the compensation algorithm to utterance duration. The speech signals had been degraded by additive noise and linear filtering that modeled the estimated characteristics of selected channels of the 2011 Development Data from the DARPA RATS Program using the Ellis `renoiser` procedure [4]. The figure shows, as a function of the duration and input SNR, the percentage of utterances for which the estimated $\hat{f}_\Delta$ is within ± 5 Hz of the true frequency shift f_Δ. For these data it is clear that durations of at least 60 seconds are sufficient to provide estimates of the frequency shift that are accurate at least 90% of the time. The shapes of the curves imply an asymptotic rise toward 100% accuracy with increasing input duration.

4. RECOGNITION OF SPEECH WITH COMPENSATED FREQUENCY SHIFTS

The CMU SPHINX-III speech recognition system was used with the DARPA RM1 database to evaluate the impact of blind compensation for SSB mistuning using both clean and degraded speech. For these comparisons we used a bigram language model and a three-state HMM-based acoustic model with mixture densities consisting of 8 Gaussians. A subset of the RM1 database that included 1600 training utterances and 600 test utterances was utilized. The ASR was trained using clean data in all cases. The clean test utterances, as well as the utterances mixed with real-world street noise at SNRs

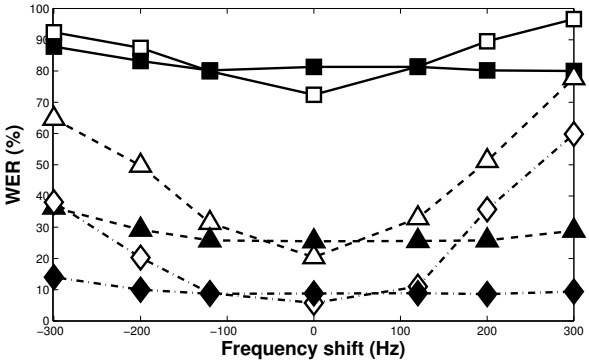

Fig. 6. WER obtained for original and blindly-compensated speech with SSB-induced frequency shifts using short segments of speech. Filled symbols denote WERs obtained for SSB-compensated speech while open symbols denote WERs for the original uncompensated speech. Estimation of frequency shift was performed on an utterance-by-utterance basis for a set of utterances averaging 3.7 s in duration.

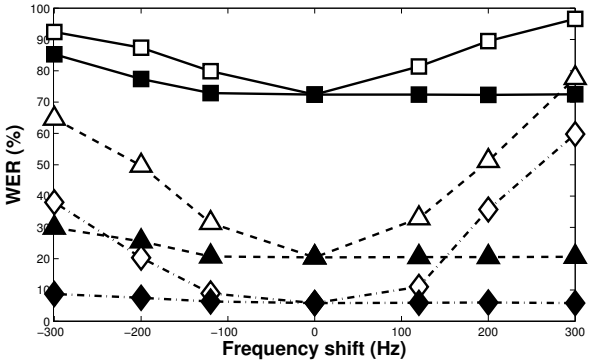

Fig. 7. Ideal WER obtained for original and perfectly-compensated speech with SSB-induced frequency shifts using a long sequence of speech. Filled symbols denote WERs obtained for SSB-compensated speech while open symbols denote WERs for the original uncompensated speech. A single estimate of frequency shift was obtained for the entire 40-minute database resulting in 100% estimation accuracy.

of 15 and 5 dB, were automatically compensated using the SSB mistuning detection algorithm described in Section 3.2. Feature extraction was performed using standard MFCC coefficients with cepstral mean normalization.

Figure 6 describes the WER obtained for a subset of the frequency shifts depicted in Fig. 3, before and after blind compensation, using the compensation method described in Sec. 3.2. For these experiments, compensation was performed on an utterance-by-utterance basis, and the average duration of the utterances was only 3.7 seconds with a standard deviation of 1.25 seconds. It can be seen that in general the compensation procedure is quite effective, and the results are especially dramatic when the SSB-induced frequency shift is large in magnitude. Some performance degradation is observed when no frequency shift is actually present. This is a consequence of the short average duration of the Resource Management utterances, as noted above. With this short duration, the SSB-induced frequency shifts are correctly estimated to within ± 10 Hz for only 84.3% of the clean utterances, so 16% of the utterances are inadvertently degraded by mis-estimation of the frequency shift.

Figure 7, in contrast, shows the WER obtained when estimation accuracy of the frequency shift is 100%. To achieve 100% estimation accuracy, the frequency shift detector was run over the entire database of approximately 40 minutes of test speech with all utterances frequency shifted by the same amount. This type of processing would be appropriate for the case of a communications channel that is known to have fixed characteristics. Comparison of Figs. 6 and 7 underscores the assertion that the algorithm performs quite well even when the durations of the speech segments are relatively brief.

Finally, Fig. 8 describes recognition accuracy obtained for utterances with additive noise and linear filtering that simulated the two channels of the 2011 DARPA RATS development set (Channels D and H) that exhibited substantial frequency shift. We compare results using three types of training procedures: (1) training using clean speech, (2) "multi-style" training using speech samples that had been degraded by all of the simulated RATS devset channels, and (3) "matched" training in which the training data are degraded by the same environmental conditions as the test data. While blind

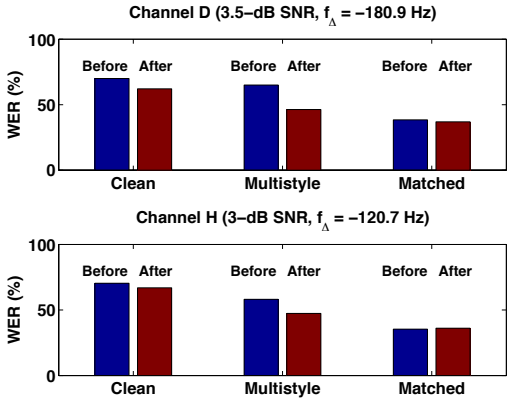

Fig. 8. WER before and after compensation for SSB mistuning for two simulated RATS-like channels using various training styles.

compensation for SSB-induced frequency shifts provided the greatest benefit using multi-style training, some improvement was observed for clean training as well. Unsurprisingly, little improvement is seen for the matched-training condition because the ASR system has incorporated the characteristics of the degraded speech into its acoustic models, including the frequency shift.

5. SUMMARY

We describe an algorithm that provides very accurate estimation of the SSB-induced frequency shifts, over a wide range of SNRs, if about 60 seconds of speech are present. The use of the algorithm provides almost-complete amelioration of the effects of mistuned SSB even for utterances shorter than 10 seconds, and recognition accuracy improves further when longer durations of degraded speech are available to estimate the frequency shifts.

6. REFERENCES

[1] J. Suzuki, T. Shimamura, and H. Yashima, "Estimation of mistuned frequency from received voice signal in suppressed carrier SSB," in *Proc. IEEE Global Telecommunications Conference*, 1994, pp. 1045–1049.

[2] J. Suzuki, Y. Hara, and T. Shimamura, "Improvement in the quality of speech received at suppressed carrier SSB," in *Proc. IEEE Global Telecommunications Conference*, 1995, pp. 1615–1618.

[3] R. J. Dick, "Co-channel interference separation," Tech. Rep. RADC-TR-80-365, Rome Air Development Center, December 1980.

[4] D. P. W. Ellis, "RENOISER – utility to decompose and recompose noisy speech files," http://labrosa.ee.columbia.edu/projects/renoiser/.