Enhancing Frequency Shifted Speech Signals in Single Side-band Communication

Sriram Ganapathy *Member, IEEE* and Jason Pelecanos, *Member, IEEE*

***Abstract*—The spectral quality of speech signals communicated over high-frequency single side band (HF-SSB) radio channels is affected by acoustic artifacts like linear frequency transpositions. In this paper, we propose an approach to automatically estimate and correct for the frequency shift given the degraded signal at the SSB receiver. The proposed method utilizes the harmonic nature of the speech signal in the voiced regions. The fundamental harmonic frequency, obtained from an autoregressive model of the spectrum, is used to estimate the offset value for the current frame. The offset values from the adjacent frames are pooled together to provide the most likely estimate for the received signal. Various experiments are performed on frequency shifted degraded speech signals received from the HF-SSB channel. The enhanced speech signal is also applied for a language identification task (LID) where the models are trained on speech data without any frequency shift. In these experiments, the proposed algorithm provides significant improvements over other baseline offset estimation methods in terms of accuracy of offset estimation as well as the LID system performance (with relative improvements of about 10-25%).**

***Index Terms*—HF-SSB Radio Communication, Frequency Offset Estimation, Speech Enhancement, Language Identification**

I. Introduction

THe single side-band (SSB) approach to high frequency (HF) radio communication continues to remain popular due to the ability to cover long distances with low power and bandwidth requirements [1]. Here, the receiver uses the heterodyning procedure to convert the modulated signal back to the baseband. If the carrier frequency at the receiver (f_r) is not synchronized with the one used at the transmitter (f_t), the SSB receiver causes a linear frequency shift in the received speech signal with the amount of shift equal to the difference $f_t - f_r$. For speech communication, the resulting frequency shift modifies the naturalness of the speech (speaker sounding robotic) and affects the perception of gender information [2], [3]. The frequency shift artifact causes spectro-temporal degradation which affects the performance of automatic speech systems like speaker and language identification.

The development of speech systems operating on degraded speech obtained from HF communication channels has received renewed interest owing to the DARPA Robust Automatic Transcription of Speech (RATS) program [4]. In the data distributed under the RATS program, there is lack of synchronization of the carrier frequency for HF-SSB transmission (for channels D and H) resulting in frequency shift [4]. Various approaches have been proposed in the past to enhance the frequency shifted signal. The use of the speech cepstrum to determine the frequency shift is investigated in [5]. A statistical approach to the frequency shift estimation is developed in [6] where a histogram matching algorithm determines the offset value. A spectral linear prediction (SLP) method for the estimation of frequency offset has been explored in [7] and the use of the modulation spectrum for offset estimation has been studied in [8]. Except for the study in [8], most of these methods were focused on clean speech and the performance degrades when the low frequency region of the speech spectrum is corrupted [5], which is typically the case in HF-SSB transmission [4].

In this paper, we propose an autoregressive (AR) model based method for frequency offset estimation which exploits the harmonic properties of the speech signal. When a speech signal is frequency shifted, the fundamental harmonics in the voiced regions are also shifted linearly. However, the separation between subsequent harmonics is unchanged and this spectral distance can be used to estimate the fundamental frequency. If the spectral peaks in the frequency shifted signal are identified, the difference between the actual location of these peaks and the expected location in the baseband signal provide likely candidates for the frequency shift estimate. Experiments are performed using HF-SSB signals present in the development portion of the RATS corpus [4]. We compare the proposed method with the previous approach in [7] and with the shift estimates obtained using the fundamental frequency values given by the YIN method [9]. In these experiments, the proposed AR model based shift estimation procedure provides significant robustness compared to other methods in terms of shift estimation accuracy (with relative improvements of about 25%). The signal quality, objectively measured using the perceptual evaluation of speech quality (PESQ) [10], is also shown to improve over baseline methods. Furthermore, the enhanced signal is used to improve the performance of an automatic language identification (LID) system.

The remainder of the paper is organized as follows. In Sec. II, we describe the proposed approach. The experimental setup used in evaluating the proposed approach is described in Sec. III. The results comparing various frequency shift estimation methods are reported in Sec. IV along with the LID results using the enhanced speech signals. Sec. V concludes with a summary of the proposed approach.

This work was supported in part by Contract No. D11PC20192 DOI/NBC under the RATS program. The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. The authors would like to acknowledge Sibel Yaman, Kyu Han, Mohamed Omar and Weizhong Zhu for help in implementing the language identification system.

S. Ganapathy and J. Pelecanos are with the IBM T.J. Watson Research Center, Yorktown Heights, NY, USA. (phone: +1-(914) 945-1326; fax +1-(914)-945-4490 ; e-mail: {ganapath,jwpeleca}@us.ibm.com).

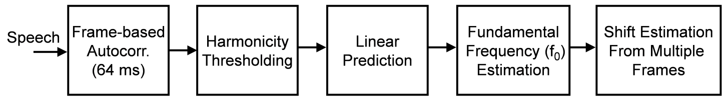
Speech
Frame-based Autocorr. (64 ms)
Harmonicity Thresholding
Linear Prediction
Fundamental Frequency (f_0) Estimation
Shift Estimation From Multiple Frames

Fig. 3. Implementation of the proposed frequency shift estimation algorithm.

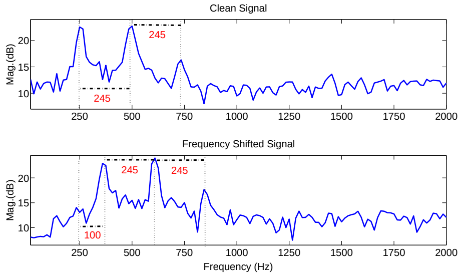

Fig. 1. Magnitude spectrum of a voiced speech frame for the clean signal (baseband) and the frequency shifted signal.

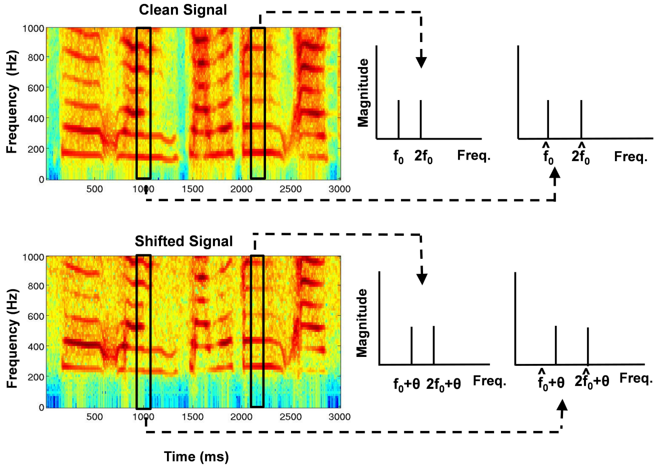

Fig. 2. Overview of the frequency shift phenomenon and its effect on two voiced speech frames.

II. Frequency Shift Estimation Using AR Models

A. Overview

When a speech signal is frequency offset due to the lack of synchrony between the SSB transmitter and receiver, the fundamental harmonic frequencies present in the voiced regions are linearly shifted. However, the separation between the fundamental harmonics is not modified by the SSB communication process as shown in Fig. 1. We propose to use this property to estimate the offset value (θ) given the degraded signal at the SSB receiver.

In particular, if the baseband signal contains the fundamental harmonic frequencies appearing at $f_0, 2f_0, 3f_0, ...$, the frequency shifted signal will contain the same fundamental harmonics appearing at $f_0+\theta, 2f_0+\theta, 3f_0+\theta, ...$, where θ denotes the unknown frequency shift. The effect of the frequency shift on voiced frames of speech is outlined in Fig. 2. Assuming that the frequency shift is constant over multiple speech frames, the following are the steps involved,

- **Determining the fundamental harmonic frequency** - An algorithm for the estimation of the fundamental harmonic frequency which is robust to frequency shift as well as other distortions in degraded radio channels including the loss of low frequency information.
- **Computing the shift** - Using the fundamental frequency estimate, the likely values for the frequency shift are estimated.
- **Combine the estimates** - As seen in Fig. 2, the fundamental harmonic is time-varying ($f_0 \neq \hat{f}_0$), but the frequency shift (θ) is assumed constant over the speech segment. Given the likely values for frequency shift from multiple speech frames, we accumulate this information over a speech segment to obtain a single shift estimate.

Note that this method of determining the frequency shift suffers from non-uniqueness for individual speech frames, i.e., the values θ, $f_0+\theta$, ... could all be likely candidates for the frequency shift. In the proposed approach, this ambiguity is resolved by combining estimates from multiple voiced frames to provide a single frequency shift for a given speech segment using a frequency binning approach.

B. AR Model Based Frequency Shift Estimation

The block schematic of the proposed approach to shift estimation is shown in Fig. 3. The input speech signal is split into short-term (64ms) frames with 50% overlap. For each frame, normalized windowed autocorrelation estimates are derived [11]. These are defined as,

$$r_{xx}[n,k] = \frac{\sum_{m=0}^{N-1} x_n[m]u[m]x_n[m+k]u[m+k]}{\sum_{m=0}^{N-1} u[m]u[m+k]} \tag{1}$$

where x_n denotes the signal for the nth frame, $u[m]$ denotes the window signal and $r_{xx}[n,k]$ denotes the autocorrelation value for frame n and lag k. The choice of normalized autocorrelation reduces the impact of strong formants in the estimation as well as the tapering effects caused by the shape of the window function [11]. In our implementation, we use a Hanning window.

The autocorrelation values are used for obtaining a harmonicity value for the current frame. This harmonicity measure has been used recently for voicing detection in speech activity detection (SAD) applications [12]. The harmonicity value is defined as,

$$H[n] = \frac{r_{xx}[n,\hat{k}]}{r_{xx}[n,0] - r_{xx}[n,\hat{k}]} \tag{2}$$

$$\hat{k} = \underset{2ms \leq k \leq 16ms}{\arg\max} \; r_{xx}[n,k]$$

The choice of 2ms and 16ms limits the fundamental frequency of interest in the $62.5 - 500$Hz range. The speech frames with harmonicity values above a preset threshold are used for frequency shift estimation.

The autocorrelation values are used for autoregressive (AR) model based spectral estimation [13]. The choice of linear prediction as opposed to FFT for spectral estimation is motivated

Segment Length	Accuracy (%)
32ms	68.6
48ms	98.4
64ms	99.0
80ms	98.4
96ms	79.8

Pole Order	Accuracy (%)
20	22.2
30	88.2
40	98.8
50	99.0
60	95.8

TABLE I
FREQUENCY SHIFT ESTIMATION ACCURACY (%) FOR ARTIFICIALLY SHIFTED SIGNALS AS A FUNCTION OF SEGMENT LENGTH (MODEL ORDER OF 50) AND THE MODEL ORDER (SEGMENT LENGTH OF 64MS).

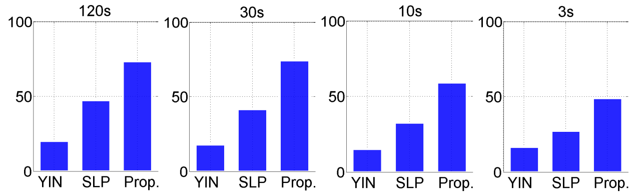

Fig. 4. Comparison of various frequency offset estimation techniques in terms of shift estimation accuracy (using 15Hz bins) for different durations of speech recordings.

by the robustness of the LP spectrum in the presence of noise as well as the parametric nature of the model. The output of the AR modeling process provides the all-pole coefficients $\{a_1, a_2, \ldots, a_p\}$ which characterize the power spectrum given by,

$$\hat{S}_{xx}(w) = \frac{G}{|\sum_{k=0}^{k=p} a_k e^{-i2\pi kw}|^2} \tag{3}$$

where G denotes the gain of the LP model and p denotes the model order. The location of the poles in the AR model relate to the angular positions of the roots of the polynomial [13]. The magnitudes of the polynomial roots indicate the sharpness of the poles. For speech frames with a considerable amount of noise, the peaks in the AR model are less pronounced and the estimated locations of the fundamental harmonics are less accurate. We apply a threshold on the pole sharpness measure for the roots in the 0-1000Hz range to select speech frames for which the harmonics can be precisely located.

The peak locations in the power spectrum are obtained and the distance between consecutive peaks is used as a measure of the fundamental harmonic frequency. In order to predict the locations of harmonics which are not present in the signal (due to low pass filtering and other channel artifacts) the harmonic frequencies are extrapolated in periodic intervals towards the 0Hz frequency value. The peak locations in the extrapolated spectrum $(\theta, f_0 + \theta, 2f_0 + \theta, \ldots)$ form multiple candidates for the frequency shift estimate from the current frame. These values are quantized with a 15Hz bin. This process is repeated for various voiced frames in the speech segment (with different fundamental frequency values) and the most likely shift estimate is chosen as the frequency bin with the highest count for the speech recording. This estimate of the frequency shift is then used in a heterodyning method to shift the speech signal back to the baseband (correcting the effect of frequency shift caused by asynchronous HF-SSB receiver).

III. EXPERIMENTAL SETUP

In this section, we describe the setup used for measuring the robustness of the proposed frequency shift estimation method. We use the development portion of the DARPA RATS database [4]. This database contains multiple parallel recordings of a clean source signal passed through various radio transmission channel configurations. In particular, channel-D corresponds to the SSB transmission where the carrier frequency at the receiver is not synchronized with the transmitter [4].

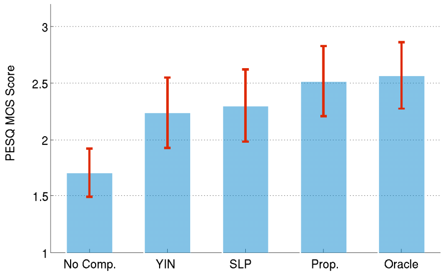

Fig. 5. Comparison of various frequency offset estimation techniques in terms of PESQ score for 120s recordings with the associated standard deviation of PESQ scores.

The availability of parallel recordings for each speech utterance enables us to measure the oracle frequency shift value for each recording. This is done by crosscorrelating the spectrogram from the source recording with the corresponding SSB channel recording. The two recordings are time aligned and the "frequency lag" associated with the maximum cross-correlation is used as the oracle reference. This is used as the target value for measuring the performance of the automatic frequency shift estimation techniques.

The first set of experiments relate to the selection of parameters in the proposed approach. We use artificially shifted versions of the clean source signal for these experiments where 1623 clean source recordings were frequency shifted by 0-250Hz with steps of 15Hz. The proposed method of automatic shift estimation is applied and the performance is measured in terms of accuracy of the binned frequency shift estimation. The parameters of interest are the length of the speech frame used for frequency shift analysis and the model order used for linear prediction. These results are reported in Table I. As seen here, the best performance is obtained with a choice of 64ms frames and a model order of 50 poles per frame. This choice of parameters is used for the rest of the experiments with noisy radio channel data.

IV. RESULTS ON HF-SSB DATA

In the next set of experiments, we compare the shift estimation performance of the proposed approach (denoted as Prop.) with two other baseline techniques,

- Spectral linear prediction (SLP) method for the estimation of frequency offset [7]
- YIN method of the fundamental frequency estimation [9] with the magnitude spectrum.

The estimation accuracy on HF-SSB data for various speech content durations is shown in Fig. 4. For the HF-SSB data, the oracle value is obtained using autocorrelation with the corresponding source recording of 120s duration. As seen here, the estimation accuracy for all methods is notably reduced

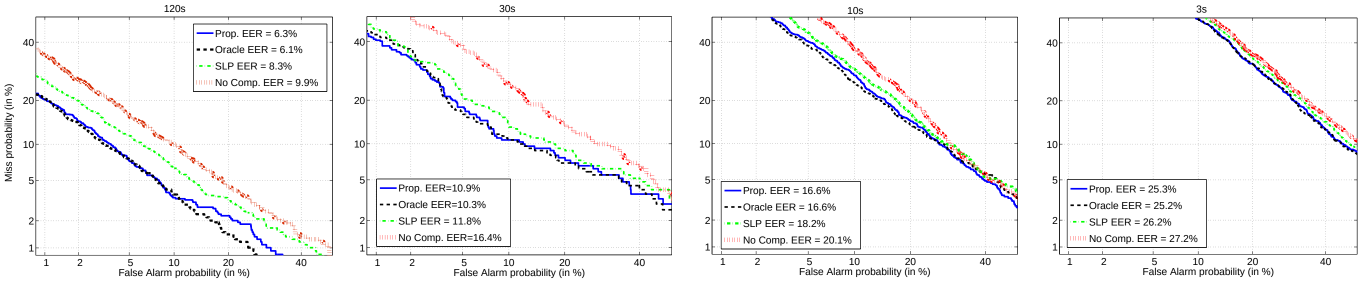

Fig. 6. Detection Error Tradeoff (DET) plots for various frequency offset compensation methods using the 120s, 30s and 10s segments of HF-SSB data.

when the duration is below 30s. The proposed approach provides significant improvements in terms of estimation accuracy (relative improvements of up to 45%) for the frequency shifted recordings obtained from the HF-SSB channel.

In addition to the frequency estimation accuracy, we objectively measure the amount of signal enhancement obtained by addressing the frequency shift problem. Specifically, we compute the Perceptual Evaluation of Speech Quality (PESQ) score between the original source recording and the enhanced signal obtained by heterodyning the HF-SSB signal back to the baseband. These results are shown in Fig. 5. The proposed method provides enhancement quality which is similar to the oracle value based offset correction. On the average, the PESQ score is improved by 0.75 with frequency shift compensation. This shows that considerable speech enhancement can be achieved by shifting the SSB transmitted speech signals back to the baseband.

The final set of experiments use the HF-SSB data for a language identification (LID) task [14]. The system uses dimensionality reduced GMM supervectors obtained by principal component analysis (PCA) with a polynomial kernel based support vector machine (SVM) classifier. This constitutes a modification of the Gaussian SuperVector-SVM (GSV-SVM) system initially proposed for speaker verification [15]. The GMM universal background model (UBM) is trained on radio channel data from the RATS database excluding the channel-D data. The LID system performance is measured on the HF-SSB channel (channel-D) using the equal error rate (EER) metric. The LID results using various frequency offset correction methods are shown in Fig. 6. The proposed approach results in significant improvements of 37% relative compared to the uncompensated baseline and 24% relative compared to other frequency offset estimation techniques for the 120s condition. For the 30s condition, the proposed approach yields 33% and 8% relative improvements compared to the uncompensated and the SLP approach respectively. The results obtained for the proposed method are very similar to those obtained with the oracle values. These results also indicate that correcting frequency offset on HF-SSB data is important for speech applications and results in considerable performance improvements.

V. Summary

In this paper, we have proposed a method for frequency shift estimation and correction of HF-SSB speech data. The proposed method relies on the harmonic properties of the speech signal and uses an AR model based spectral envelope for offset estimation. Various experiments are performed which measure the estimation accuracy, enhancement quality as well as usefulness in a language identification task. The proposed approach to frequency shift estimation provides significant improvements.

References

[1] A. B. Carlson and P. B. Crilly, “Communication systems: an introduction to signals and noise in electrical communication,” vol. 1221, 1975.

[2] P. Assmann, S. Dembling, and T. Nearey, “Effects of frequency shifts on perceived naturalness and gender information in speech,” in *Proceedings of the 9th International Conference on Spoken Language Processing*, 2006, pp. 889–892.

[3] Q. Fu and R. Shannon, “Recognition of spectrally degraded and frequency-shifted vowels in acoustic and electric hearing,” *The Journal of the Acoustical Society of America*, vol. 105, pp. 1889–1900, 1999.

[4] K. Walker and S. Strassel, “The RATS Radio traffic collection system,” in *Odyssey Speaker and Language Recognition Workshop*. ISCA, 2012.

[5] J. Suzuki, T. Shimamura, and H. Yashima, “Estimation of mistuned frequency from received voice signal in suppressed carrier SSB,” in *Global Telecommunications Conference, 1994. GLOBECOM’94. Communications: The Global Bridge., IEEE*, vol. 2, 1994, pp. 1045–1049.

[6] T. Gülzow, U. Heute, and H. Kolb, “SSB-carrier mismatch detection from speech characteristics: Extension beyond the range of uniqueness.” EUSIPCO, 2002.

[7] D. Cole, S. Sridharan, and M. Moody, “Frequency offset correction for HF radio speech reception,” *Industrial Electronics, IEEE Transactions on*, vol. 47, no. 2, pp. 438–443, 2000.

[8] P. Clarke, H. Mallidi, A. Jansen, and H. Hermansky, “Frequency offset correction in speech without detecting pitch,” in *ICASSP*, 2013.

[9] A. De Cheveigné and H. Kawahara, “YIN, a fundamental frequency estimator for speech and music,” *The Journal of the Acoustical Society of America*, vol. 111, p. 1917, 2002.

[10] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-A new method for speech quality assessment of telephone networks and codecs,” in *ICASSP*, vol. 2, 2001, pp. 749–752.

[11] P. Boersma, “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound,” in *Proceedings of the institute of phonetic sciences*, vol. 17, 1993, pp. 97–110.

[12] S. Sadjadi and J. Hansen, “Unsupervised speech activity detection using voicing measures and perceptual spectral flux,” in *IEEE Signal Processing Letters*, vol. 20, 2013, pp. 197–200.

[13] J. Makhoul, “Linear prediction: A tutorial review,” *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[14] S. Yaman, J. Pelecanos, and M. Omar, “On the use of nonlinear polynomial kernel SVMs in language recognition,” in *Proceedings of Interspeech*, 2012.

[15] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, “Support vector machines using GMM supervectors for speaker verification,” *Signal Processing Letters, IEEE*, vol. 13, no. 5, pp. 308–311, 2006.