FREQUENCY OFFSET CORRECTION IN SPEECH WITHOUT DETECTING PITCH

Pascal Clark, Sri Harish Mallidi, Aren Jansen, and Hynek Hermansky

Human Language Technology Center of Excellence, Center for Language and Speech Processing
Johns Hopkins University, Baltimore, Maryland USA

ABSTRACT

Radio-transmitted speech sometimes contains a residual frequency shift or offset, resulting from incorrect demodulation in single-sideband channels. Frequency-shifted speech can mask speaker identity and reduce intelligibility. Therefore, frequency offset will degrade the performance of downstream speech technologies. Existing offset correction methods require a pitch estimate of the speech signal, which is difficult in noisy radio channels. We present a new, automatic algorithm for detecting and correcting frequency offset, based on third-order modulation spectral analysis. Our method is remarkably simple and does not require pitch estimation. We provide derivations, examples, and a pilot study demonstrating how offset correction improves speaker verification for radio-transmitted speech.

Index Terms— Modulation spectrum, speech enhancement, speaker recognition, single-sideband, frequency offset

1. INTRODUCTION

A prevalent form of radio communication is single-sideband (SSB) radio, which is one channel-type in the DARPA RATS challenge [1]. A particular problem in SSB is frequency-shift distortion, described as follows. SSB works by first shifting the speech spectrum to a high-frequency slot in the radio spectrum. In a process called demodulation, the receiver downshifts the signal to audible frequencies [2]. When the receiver fails to synchronize with the transmitter, the received speech contains a residual frequency offset. Defining the speech signal as $y(t)$ with Fourier transform $Y(f)$, the received distorted signal is

$$R(f) = Y(f - \Delta f), \quad 0 \leq f < \infty. \tag{1}$$

Frequency offset distortion is a problem for automatic speech and speaker recognition. It is clear that (1) will affect spectral feature representations such as MFCC and PLP. Perceptually, frequency offset is audible for $\Delta f > 5$ Hz, beyond which it becomes increasingly cartoonish or "chipmunk-like." Experimentally, Başkent and Shannon [3] found that frequency shift significantly degrades vowel recognition by humans. Therefore, automatic frequency offset correction is essential for large-scale processing of radio speech.

In digital communications, it is common to infer Δf from repeated symbols [4, 5], or from a training symbol sequence [6]. This is inapplicable for post-hoc processing, but the general idea is useful. Equation (1) is invertible provided we have an idea of what $Y(f)$ should look like. For instance, Voelcker [7] solved the offset problem with the assumption of a prominent, yet artificial, reference feature in $Y(f)$. Since speech is our signal of interest, the estimate for Δf should maximize the speech-like quality of $Y(f) = R(f + \Delta f)$. For instance, pitch-based methods estimate Δf as the offset that restores harmonicity of voiced speech [8, 9].

Pitch estimation, however, is difficult in adverse channel conditions and noise. Furthermore, pitch unnecessarily abstracts away the communication-theoretic components of the speech signal. We instead demonstrate that frequency shift correction is more simply posed in the modulation-frequency domain. With no need for pitch estimation or training, our proposed method is a lightweight module that can recover previously unusable audio for human or machine consumption.

We first introduce and motivate the modulation perspective in Section 2. Then, we derive the automatic estimator in Section 3. Using an implementation of the estimator, we preprocess RATS data for speaker verification in Section 4. Finally, we conclude in Section 5.

2. A MODULATION SPECTRAL PERSPECTIVE

Although we defined frequency offset in the Fourier domain, the estimation problem is not well-posed in terms of the spectrum $R(f)$. This is due to the high variability of the speech spectrum, depending on what is being said, who is saying it, the channel transfer function, and what noise is present. It is therefore difficult to identify a truly invariant spectral reference for recovering $Y(f)$ from $R(f)$. Harmonicity is one possibility, but this requires pitch detection and a guarantee that $Y(f)$ contains the fundamental. Both requirements are dubious for highly compressed and noisy radio transmissions.

The research presented in this paper was partially funded by the DARPA RATS program under D10PC20015, IARPA BABEL program under W911NF12-C-0013 and the JHU Human Language Technology Center of Excellence. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA or IARPA or the JHU HLTCOE.

Our proposed alternative uses the modulation-frequency spectrum, rather than the acoustic-frequency spectrum, as a reference signal. In contrast to the acoustic spectra $R(f)$ and $Y(f)$, the modulation spectrum is a model-based estimator tuned to the temporal data rate of speech. As a result, the modulation spectrum, also known as the modulation transfer function, is peaked around the syllabic rate of 4 Hz. This observation, demonstrated as early as 1928 [10] and systematized by Houtgast and Steeneken [11], indicates that the temporal bandwidth of speech is actually quite low. Experimental studies have since proven that modulation frequencies up to 16 Hz are necessary for preserving speech intelligibility [12, 13], an indispensable fact in modern speech processing representations (e.g., [14, 15]).

We will use n to denote sample index and f_s the sampling rate in Hertz. The discrete-time complex analytic signal is defined for a general signal $x[n]$ as

$$x_a[n] = \int_0^{f_s/2} X(f) e^{j2\pi(f/f_s)n} df \tag{2}$$

from which we find

$$r_a[n] = y_a[n] e^{j2\pi\Delta f n}. \tag{3}$$

Let us introduce the new, third-order complex "envelope" as

$$d_x[n] = x_a[n] \left| x_a[n] \right|^2. \tag{4}$$

One element of $d_x[n]$ is the second-order envelope $|x_a[n]|^2$. From (3) we see that that $|r_a[n]|^2 = |y_a[n]|^2$, and is therefore offset-invariant. This acts as an implicit harmonic reference in $d_r[n]$, since $d_r[n] = r_a[n]|y_a[n]|^2$.

Taking the Fourier transform yields the new, third-order modulation spectrum

$$D_x(f) = \sum_n d_x[n] e^{-j2\pi(f/f_s)n}. \tag{5}$$

Substituting (3) into (5), we obtain

$$D_r(f) = D_y(f - \Delta f). \tag{6}$$

It would appear that (6) is identical to (1), except for some important properties. As we shall prove in the next section, $D_y(f)$ is peaked at $f = 0$, does not change with pitch, and has narrow bandwidth commensurate with the modulation bandwidth of speech. These properties provide a necessary and stable reference feature for detecting Δf in $D_r(f)$.

Figure 1 demonstrates the advantage of the third-order modulation spectrum over the conventional power spectrum. It shows an example of original speech $D_y(f)$ with its synthetically frequency-offset version $D_r(f)$, to simulate SSB radio reception. The frequency shift is directly visible in $D_r(f)$ as the location of the baseband peak. Such a clear and invariant feature does not exist in the general acoustic spectrum of speech.[1]

[1] A special exception is when the SSB carrier is not suppressed, in which case the acoustic spectrum may contain a peak at Δf Hz.

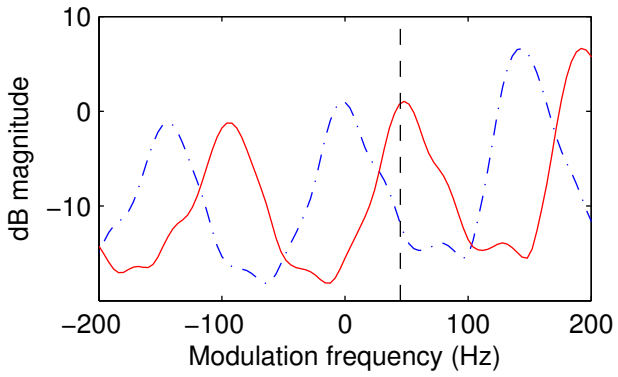

Fig. 1. Overlay of $D_r(f)$ for a synthetically frequency-shifted speech example (red, solid), with $\Delta f = 50$ Hz, compared to the original speech spectrum $D_y(f)$ (blue, dashed).

3. THIRD-ORDER MODULATION SPECTRAL ESTIMATOR

Next, we will derive the third-order modulation spectrum from basic speech properties, and present a means for automatic frequency offset detection.

Rather than the Fourier transform in (5), a spectrogram is more useful for Welch-style spectral estimation. The short-time, third-order modulation power spectrum is defined as

$$D_x[n, f) = \sum_p g[p] d_x[n + p] e^{-j2\pi(f/f_s)p} \tag{7}$$

where $g[n]$ is a short analysis window that dictates the frequency resolution of the modulation spectrogram. As in (6), we still have $D_r[n, f) = D_y[n, f - \Delta f)$.

We are now in a position to state the main result of this paper. Although $D_r[n, f)$ contains many peaks, the offset Δf is the center of pitch-related symmetry in the modulation-frequency axis. Figures 2 and 3 illustrate this property, for speech before and after transmission through an actual SSB radio channel. Our statement is descriptive and requires justification, as well as practical considerations. In the following subsections, we discuss each stage of the estimator in detail.

3.1. Sum-of-Products Signal Model for Speech

The properties of the third-order modulation spectrum are best understood in terms of a modulation synthesis model for speech. Later, we will make predictions about the time-frequency structure of $d_y[n]$ after observing the effects of third-order basebanding on the modulation components of speech.

We assume the sum-of-products model [16], similar to the sinusoidal model [17], given by

$$y_a[n] = \sum_{k=k_1}^{K} m_k[n] \exp(jk\phi_0[n]). \tag{8}$$

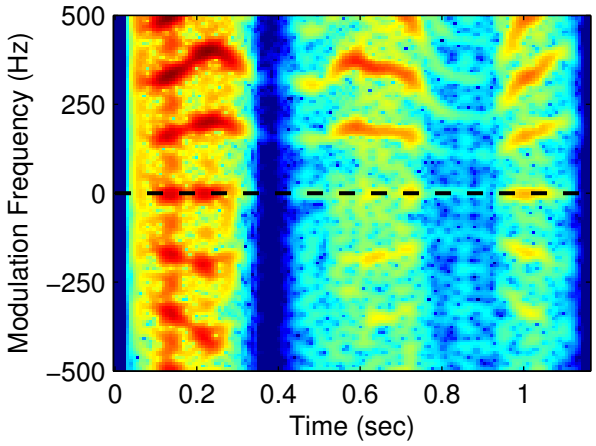

Fig. 2. The modulation spectrogram $|D_y[n, f)|$ for speech, before transmission through an SSB radio channel. The dashed line indicates the center of vertical symmetry about zero Hz.

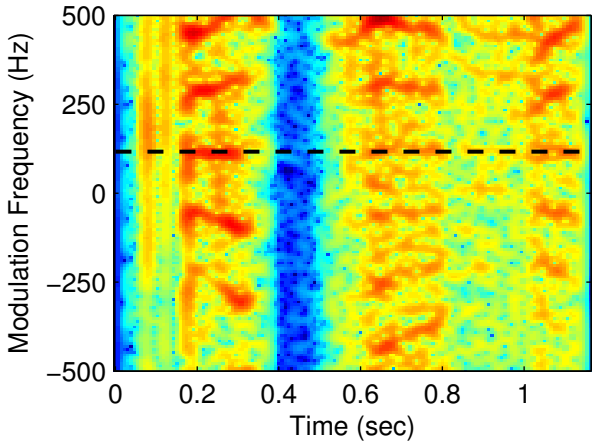

Fig. 3. The modulation spectrogram —$D_r[n, f)|$ after transmission through an actual SSB radio channel. The center of vertical symmetry at 117 Hz, has shifted relative to Figure 2.

where $m_k[n]$ is the kth lowpass modulator, and $\phi_0[n]$ is the phase of the fundamental subband. We indicate that k starts at $k_1 \geq 1$ as a result of possible bandpass filtering in the radio channel.

Based on the empirical modulation studies discussed in Section 2, we assume the modulators are bandlimited on the order of tens of Hertz. Conversely, the harmonics $\exp(jk\phi_0[n])$ are bandpass in frequency, on the order of hundreds of Hertz.

3.2. Third-Order Basebanding

The third-order nonlinearity in (4) should be viewed as an operation that brings speech modulations $m_k[n]$ to baseband, or centered on zero Hz. It is sufficient to prove this for $d_y[n]$, since $d_r[n]$ is identical except shifted by Δf.

In terms of the sum-of-products model (8), $d_y[n] = L_B[n] + L_H[n]$, where $L_B[n]$ is the baseband term

$$L_B[n] = \sum_{k,p \geq k_1} m_k[n] m_p[n] m^*_{k+p}[n] \tag{9}$$

and $L_H[n]$ consists of harmonic sideband terms

$$L_H[n] = \sum_{k \neq q-p} m_k[n] m_p[n] m^*_q[n] e^{j\phi_0[n](k+p-q)}. \tag{10}$$

In Figure 2, $L_B[n]$ is the baseband term centered at zero Hertz. Important facts about the third-order envelope include:

- Due to the slowly-varying nature of speech modulations, $L_B[n]$ is lowpass and disjoint from the harmonic terms in $L_H[n]$.
- $L_B[n]$ is not frequency-modulated, as evidenced by the lack of complex exponentials in (9).
- Due to multiple combinations of cross-terms in $L_H[n]$, it is likely that there exist symmetrically frequency-modulated terms in $L_H[n]$.

Together, these three points imply that Δf can be detected as the center of symmetric-like structure in the Fourier transform of $d_y[n]$, or $D_y(f)$. This property is the basis for the Δf estimator described next.

3.3. Estimator Based on Modulation Symmetry

We are careful to state that $L_H[n]$ has symmetric frequency-modulations, because $D_y[n, f)$ is not truly symmetric in f. That is, the harmonics appear to move symmetrically around zero Hertz in Figure 2, but amplitudes are not symmetric.

To emphasize the harmonic symmetry in $D_r[n, f)$, we propose to use homomorphic filtering [18]. The modified modulation spectrum is

$$D'_r[n, f) = W(f) \circledast \log |D_y[n, f)| \tag{11}$$

where $\circledast$ denotes circular convolution, and $W(f)$ is a highpass edge-detector defined as

$$W(f) = \delta(f) - \sum_{q=-T}^{T} e^{j2\pi(f/f_s)q}. \tag{12}$$

Since $W(f)$ is intended to emphasize harmonics, the highpass cutoff $T = 1/f_{max}$ corresponds to the maximum pitch spacing preserved by the filter. The logarithm serves to compress the vast dynamic range of $D_r[n, f)$, as well as separate the spectral envelope from harmonic ridges in the f dimension.

To find the center of symmetry, we match the modulation spectrum against itself in an auto-convolution, defined by

$$A_r(f) = \sum_n \int_{-f_s/2}^{f_s/2} D'_r[n, 2f - u) D'_r[n, u) du. \tag{13}$$

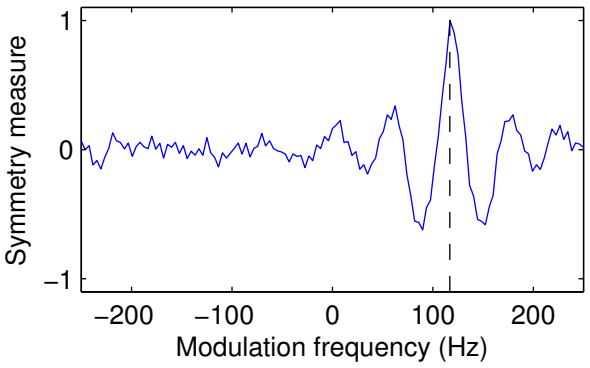

Fig. 4. Plot of $A_r(f)$ for the signal shown in Figure 3. The dashed line indicates the known offset of $\Delta f = 117$ Hz.

We find the offset frequency Δf as

$$\Delta f = \underset{f}{\operatorname{argmax}}\, A_r(f). \tag{14}$$

Due to the circularity of the discrete-time Fourier transform, the auto-convolution estimator is restricted to the range $|\Delta f| < f_s/4$. Figure 4 shows $A_r(f)$ for the received signal in Figure 3, indicating a strong peak at the known offset of 117 Hz.

Finally, the corrected speech signal is

$$y[n] = \mathrm{Re}\{r_a[n] \exp(-j2\pi(\Delta f/f_s)n)\}. \tag{15}$$

4. APPLICATION TO SPEAKER VERIFICATION

The DARPA RATS program [1] deals with conversational telephone speech in various languages transmitted over several radio channels. One channel is known to cause SSB offset, which varies randomly from conversation to conversation. In this section, we describe two speaker verification experiments using frequency-offset correction as a preprocessor module. We conducted both experiments on the LDC2012E40 subset provided by the Linguistic Data Consortium.

In the first experiment, we processed only the files from the SSB channel, labeled as "channel D," to simulate the scenario where data has been flagged for SSB-offset correction. We call this scenario DEMOD1. Our Δf estimates for channel D were approximately Gaussian distributed with mean 41 Hz and standard deviation 25 Hz. Outliers also appeared around 1000 Hz, which we determined were erroneous. In response to outliers, we made a second demodulated corpus called DEMOD2, in which we processed only channel D but this time set outlier estimates to zero Hz, for no demodulation.

Outlier Δf estimates occurred from spurious peaks in the modulation spectrum, usually resulting from tonal noise. However rare, large errors can destroy the identity and intelligibility of a signal. We found that speech activity gating can prevent some outliers. For this reason, we used a robust speech activity detector [19] for both DEMOD1 and DEMOD2, trained specifically for the RATS channels.

After preprocessing, we fed the corrected audio from both DEMOD1 and DEMOD2 into i-vector speaker verification experiments. The features used time-frequency autoregressive modeling [20], referred to as PLP2 features. A 512 mixture, gender-independent GMM-UBM was trained using 42 hours of the data. UBM means were concatenated into supervectors, which were used to train an i-vector factor analysis [21] using 630 hours of data. The resulting i-vectors were then used to train a PLDA system [22], providing a 150 dimensional subspace for final scoring.

The table below shows equal error rates (EER) for the baseline and both demodulation conditions. Although we modified files from only Channel D, this changed the training data used for all channels. For this reason, non-D channels see small differences between the baseline, DEMOD1, and DEMOD2. The reader will see that channel D is the second highest EER in the baseline evaluation, and drops by almost 1.5 and 2 points in the two demodulation conditions. Considering non-D channels, there is on average no difference between DEMOD2 and the baseline.

	Equal Error Rate (%)		
Channel	*Baseline*	*DEMOD1*	*DEMOD2*
A	6.47	6.65	6.47
B	7.63	7.74	7.51
C	5.02	5.19	5.13
D	**11.19**	**9.68**	**9.23**
E	9.02	9.01	8.98
F	6.30	6.34	6.41
G	4.48	4.67	4.54
H	18.01	18.00	17.6
Mean	8.06	7.97	7.82

5. CONCLUSION

Frequency offset is a type of distortion in SSB radio-transmitted speech. Since most speech processing features are based on the Fourier spectrum, this distortion lies outside the usual realm of robust features. Our solution is an automatic estimator of frequency offset using the third-order modulation spectrum. From a sum-of-products model of speech, we showed that frequency offset is the center of pitch-related symmetry in the modulation spectrogram.

In a speaker verification task, our proposed method of frequency offset correction yielded up to 2% absolute reduction in EER. Opportunities for further work include time-varying frequency offset, and a means of removing or mitigating tonal interference in the audio.

6. REFERENCES

[1] K. Walker and S. Strassel, "The RATS radio traffic collection system," in *Odyssey 2012: The Speaker and Language Recognition Workshop*, 2012.

[2] B.P. Lathi, *Signal Processing and Linear Systems*, Berkeley-Cambridge Press, United States of America, 1998.

[3] D. Başkent and R.V. Shannon, "Combined effects of frequency compression-expansion and shift on speech recognition," *Ear and Hearing*, vol. 28, no. 3, pp. 277–289, June 2007.

[4] P.H. Moose, "A technique for orthogonal frequency division multiplexing frequency offset correction," *IEEE Trans. Communications*, vol. 42, no. 10, pp. 2908 – 2914, Oct. 1994.

[5] S.H. Fan, J. Yu, D. Qian, and G.K. Chang, "A fast and efficient frequency offset correction technique for coherent optical orthogonal frequency division multiplexing," *J. Lightwave Tech.*, vol. 29, no. 13, pp. 1997 –2004, July 2011.

[6] M. Johnson, L. Freitag, and M. Stojanovic, "Improved doppler tracking and correction for underwater acoustic communications," in *IEEE ICASSP*, April 1997, vol. 1, pp. 575 –578 vol.1.

[7] H. Voelcker, "Demodulation of single-sideband signals via envelope detection," *IEEE Trans. Communication Technology*, vol. 14, no. 1, pp. 22 –30, Feb. 1966.

[8] D. Cole, S. Sridharan, and M. Moody, "Frequency offset correction for hf radio speech reception," *IEEE Trans. Industrial Electronics*, vol. 47, no. 2, pp. 438 – 443, April 2000.

[9] T. Gülzow, U. Heute, and H.J. Kolb, "SSB-carrier mismatch detection from speech characteristics: Extension beyond the range of uniqueness," in *EUSIPCO*, 2002, vol. I.

[10] R. R. Riesz, "Differential intensity sensitivity of the ear for pure tones," *Phys. Rev.*, vol. 31, pp. 867–875, May 1928.

[11] T. Houtgast and H. J. M. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Am.*, vol. 77, no. 3, pp. 1069–1077, 1985.

[12] R. Drullman, J.M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Am.*, vol. 95, no. 2, pp. 1053–1064, 1994.

[13] T. Arai, M. Pavel, H. Hermansky, and C. Avendano, "Intelligibility of speech with filtered time trajectories of spectral envelopes," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, oct 1996, vol. 4, pp. 2490 –2493 vol.4.

[14] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech, Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct 1994.

[15] S. Ganapathy, S. Thomas, and H. Hermansky, "Temporal envelope compensation for robust phoneme recognition using modulation spectrum," *J. Acoust. Soc. Am.*, vol. 128, no. 6, pp. 3769–3780, 2010.

[16] P. Clark, G. Sell, and L. Atlas, "A novel approach using modulation features for multiphone-based speech recognition," in *Proc. IEEE ICASSP*, Prague, 2011.

[17] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 34, no. 4, pp. 744 – 754, aug 1986.

[18] A.V. Oppenheim, "Superposition in a class of nonlinear systems," *IEEE Int. Conv. Rec.*, vol. 1, pp. 171–177, 1964.

[19] S. Thomas, S. H. Mallidi, T. Janu, H. Hermansky, N. Mesgarani, X. Zhou, S. Shamma, T. Ng, B. Zhang, L. Nguyen, and S. Matsoukas, "Acoustic and data-driven features for robust speech activity detection," in *Proc. InterSpeech*, 2012.

[20] S. Ganapathy, S. Thomas, and H. Hermansky, "Feature extraction using 2-d autoregressive models for speaker recognition," in *ISCA Speaker Odyssey*, June 2012.

[21] N. Dehak, P. Kenny, R. Dehak, P Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verication," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[22] D. Romero and C.Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. Interspeech*, 2011.