FREQUENCY OFFSET CORRECTION IN SINGLE SIDEBAND SPEECH FOR SPEAKER VERIFICATION

Hua Xing, Philipos C. Loizou, John H.L. Hansen

Department of Electrical Engineering, University of Texas at Dallas, Richardson, USA
{hxx093020, John.Hansen}@utdallas.edu

ABSTRACT

Communication system mismatch represents a major influence for loss in speaker recognition performance. While microphone and handset differences have been considered in the NIST SRE, nonlinear communication system differences, such as modulation/demodulation (Mod/DeMod) carrier drift, have yet to be considered. In this study, an algorithm for estimating and correcting Mod/DeMod frequency offsets distortion in signal side-band modulation (SSB) speech is formulated based on two processing steps. In the first step, the offset of speech can be roughly scaled to a small frequency interval, which eliminates the ambiguity caused by periodicity of the spectrum. The second step performs fine-tuning within the pre-determined interval. For the first time, a statistical framework is developed for unique interval detection, where an innovative acoustic feature is proposed to represent different offsets and state-of-the-art techniques, the total variety method and PLDA, are applied. Speaker recognition experiments on SSB speech obtained from DAPPA RATS corpus show that a significant performance improvement (up to 50% relative improvement in EER) for speaker verification in SSB speech can be obtained by the proposed estimation and compensation method.

Index Terms— frequency offset, SSB, speaker verification, MFCC, i-Vector, PLDA

1. INTRODUCTION

In radio communication, single side-band (SSB) communication is an important and commonly used contemporary communicative approach. The main reason for its popularity lies in the advantages of power saving and narrow bandwidth introduced by the techniques for suppressing or removing the carrier signal and one sideband, while only leaving a single sideband in the transmitted signal. These advantages are very appealing as the radio-frequency spectrum, once thought to be adequate for all needs, is becoming crowded due to increased data/voice traffic requirement in today's wirelessly connected society.

A disadvantage of SSB transmission is that the received signal is easily distorted by a frequency offset introduced by a mismatch between the carrier frequency of the received signal and the carrier frequency used in demodulation. For speech signals, the distortion of frequency shift makes the speech unpleasant, sounding strange and 'Donald Duck'-like to the listener, and results in poor quality and intelligibility of the speech signal [1]. Moreover, frequency offset causes a problem in automatic speech and speaker recognition because it affects features based on spectral structure such as MFCCs and PLPs. In this study, we explore the problem of automatically estimating frequency offset in SSB speech, in order to help improve speaker recognition in radio communication data.

A number of studies have been reported to detect and correct frequency offset in SSB speech [2,3,4]. Most of these methods are based on the relationship between the estimated pitch f_0 and the observed peak locations corresponding to the harmonics of f_0 in voiced speech. If the signal is distorted by a frequency offset Δf, although linear shifts destroy the expected harmonic relationship between individual components, the spacing between the actual harmonic components remains unchanged. In [2], both f_0 and the positions of several spectral peaks, $p(n)$, are estimated, and Δf is deduced from the linear relationship $\Delta f = p(n) - nf_0$. In [3,4], a comb filter with a spectral period equal to f_0 and a moving phase was fitted to the spectrum of voiced speech. Δf was then estimated as the phase of the best fitting comb filter. The ambiguity is obvious from this framework: if Δf is a possible frequency shift, $\Delta f \pm nf_0$ (n=1,2,3…) are also possible options. [3] overcame this problem by accumulating the maximum value of correlation from frame to frame as a histogram. Δf was estimated as the position that gives the maximum in histogram. The method obtained effective results given sufficiently long speech signals. [5] used a probabilistic estimation method to overcome the above limitation. Recently, a work was developed in [6] to estimate Δf using a modulation spectral analysis.

In this study, we propose a two-step method to estimate the frequency offset in an SSB speech signal. In the first step, the value of Δf is scaled to a unique interval smaller than f_0 by a statistical method; the board ambiguity can be mostly eliminated with this step. Next, a fine-tuning is performed within the predetermined unique interval to estimate Δf without uncertainty. [7] proposed a statistical method to detect any frequency shifts of a speech signal. In this study,

we make modifications on the algorithm in [7] and apply it to the SSB signal in step one.

For validation, a state-of-the-art speaker verification system is established [8-13].

In Sect. 2, an overview of the proposed two-step method is presented. Techniques used in the first stage to improve the result, symmetric partial spectral smoothed MFCC, i-Vector and PLDA are discussed in Sect. 3-Sect. 5, respectively. Experimental results and discussion are presented in Sect. 6.

2. OVERVIEW OF METHOD

The proposed offset estimation method is based on two steps (Fig. 1). In the first step, a unique interval that contains the possible frequency offset without ambiguity caused by the periodic property in frequency is detected, which means that only one value in this interval can be the possible candidate for more-refined estimation results in the second step. This condition requires that the length of each unique interval should be less than the speech pitch frequency. In this study, an interval of 50Hz is considered to satisfy this requirement, given that the pitch values of most speakers are between 60Hz and 200Hz. A statistical method proposed in [7] is used to obtain this unique interval. Once the unique interval is identified, a fine search for the frequency offset within this frequency interval is carried out in the second step.

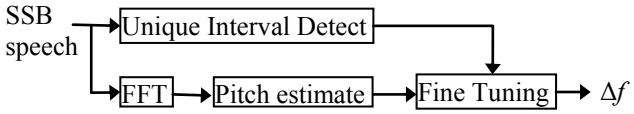

Fig.1: Block diagram of the overall frequency offset estimate

2.1 Unique range detection

A frequency offset range from -300Hz to 300Hz is divided into small bins of 50Hz without overlap, which we call unique intervals. Each utterance is segmented into 40ms frames with a 20ms overlap. Voice activity detection (VAD) is used to choose the frames with the offset information. A feature is needed for each voice active frame representing different frequency offsets. MFCC is a feature successfully used in speech, speaker, and language recognition [12, 13]. MFCC is even better suited for representing frequency offset because the energy in the frequency bands varies with spectrum shifts upward or downward. The energy variation due to frequency offset appears to be constant within a short period of speech for a certain offset frequency but varies when the offset value changes. This simple fact leads us to use a modified MFCC as an acoustic feature representing frequency offsets. The feature works successfully in unique interval detection. Next section describes how we modify the MFCC feature to make it more efficient for this task.

After acoustic features extraction, the total variety method was used to extract i-Vectors for each utterance. Finally, a generative model PLDA was used to decide which unique interval should be labeled for the utterance.

2.2 Fine estimation of frequency offset value

Once the unique interval is detected, value of the frequency offset can be finely searched within the unique interval. A measurement of pitch frequency that is reliable in the presence of both additive noise and the frequency offset is required. 'Complex correlation', suggest by [3] is used in this study to estimate pitch frequency. Mathematically, given $S(0)$, $S(1)\ldots,S(N-1)$, representing the power spectrum on positive frequencies, the complex correlation $C(0),C(1), \ldots,C(N-1)$ is given by the following equation:

$$C(n)=\frac{1}{N}\sum_{k=0}^{N-1}S(k)\exp(\mathrm{j}2\pi kn/N) \quad (1)$$

for n=0,1,…,N-1. It is proven that the pitch period T can be estimated as the index that provides the maximum value of the magnitude of complex correlation, after ruling out the first few indexes. The estimate of pitch frequency f_0 is obtained by the relation:

$$f_0 = f_s / T \quad (2)$$

A quadratic interpolation is used to obtain a more accurate estimate T_c and the corresponding $C(T_c)$ because the real T is not always an integer [3].

With the estimated pitch frequency and corresponding complex correlation value, a set of possible values of frequency offset Δf for a given frame can be calculated from the real and imaginary parts of $C(T)$ according to the following equation [3]:

$$\Delta f_r = \left(\frac{f_0}{2\pi}\tan^{-1}\left(\frac{\mathrm{Im}\,C(T)}{\mathrm{Re}\,C(T)}\right)\right)+rf_0 \quad (3)$$

where r denotes individual possible values of Δf that are f_0 apart. There are several possible values that can be chosen that correspond to different values of r, which is where the unique interval comes into play. We only need to pick the value of Δf from (3) that falls into the pre-determined unique interval. Because there is only one possible value within this interval, the possibility of picking the wrong candidate has been dramatically reduced as long as the correct unique interval was identified. Additionally, compared to other existing methods, a shorter speech is sufficient to make an accurate estimate.

3. SYMMETRIC PARTIAL SMOOTHED SPECTRAL MFCC (SPSSMFCC)

An innovative aspect of the developed method is the step that determines a unique range in which only one potential value exists before finer estimation is attempted. A powerful feature that can represent the characteristics for different frequency offsets is indispensable for the statistical framework. Here, we propose modifying two aspects of the standard MFCC based on the observation of the speech spectrum with frequency offset. First, the frequency shift varies slower than the speech spectrum. The spectrum component in the frequency domain due to the frequency shift remains constant for a relatively long duration compared to a short time variation in the speech spectrum.

In this study, the spectrum component corresponding to the frequency offset is assumed to be unchanged within one utterance. Consequently, instead of pursuing the spectrum variation along time, extracting and enhancing the invariant components in the spectrum over a long-term range is more reasonable. Based on this idea, we propose smoothing the spectrum over adjacent frames to emphasize the constant component while reducing fast varying components in the speech signal. The second observation is that the useful information for a frequency offset is reflected in spectrum variation in low frequencies (positive offsets) or high frequencies (negative offsets), while the spectrum component in the middle frequency range bears little information about frequency offset. In view of this fact, two modifications are made to the mel-frequency filter bank in the MFCC calculation: we propose to use a filter bank where the width of the filters are symmetric to the middle frequency to capture the frequency offset information at both high and low frequency range, and discard the filters in the middle frequency range and use those on the two ends.

Fig. 2 demonstrates the process of this symmetric Partial Smoothed Spectral MFCC (SPSSMFCC) calculation. The overall process is similar to a MFCC calculation except for the spectrum smoothing and filter bank shape. Specifically, after pre-emphasis and segmentation, each speech frame is transformed into the frequency domain by DFT. After that, the spectrum is smoothed by averaging the spectra of each frame with its adjacent frames. The length of the smoothing window can be varied with frequency. Here, we set the window length to 3 for all frequencies for a simple initial attempt. Following smoothing, a vector of spectral power representation is calculated by applying the symmetric partial mel-frequency filter bank described above to the smoothed spectrum. The number of filters in the filter bank was determined empirically. The optimal number is 80 by experience. A discrete cosine transform (DCT) is applied to the logarithm of the power representation. The first 12 DCT coefficients and log of the energy constitute the 13-dimensional SPSSMFCC coefficients, which is concatenated with its first and second differential Δ, ΔΔ to form a 39-dimensional feature vector for each signal frame.

4. I-VECTOR

The total variability approach has become the state-of-the-art technique in speaker recognition field [14-17]. This approach is efficient in reducing the large-dimensional input data to a small-dimensional feature vector named i-Vector, while retaining most relevant information. We believe that useful frequency offset information can be obtained by a similar front-end process. Therefor we applied the total variability approach in frequency offset detection. For a given utterance, the offset and channel variability dependent GMM supervector can be denoted by the following equation:

$$M = m_{UBM} + Tw \tag{4}$$

where m_{UBM} is the UBM supervector, T is total variability space, and the entries of the vector w is the i-Vector. To calculate i-Vector, we need the Baum-Welch statistics:

$$N_c = \sum_{t=1}^{L} P(c \mid y_t, \Omega) \tag{5}$$

$$F_c = \sum_{t=1}^{L} P(c \mid y_t, \Omega) y_t \tag{6}$$

where c=1,…,C is the Gaussian index and $P(c \mid y_t, \Omega)$ is the posterior probability of mixture component c generating the

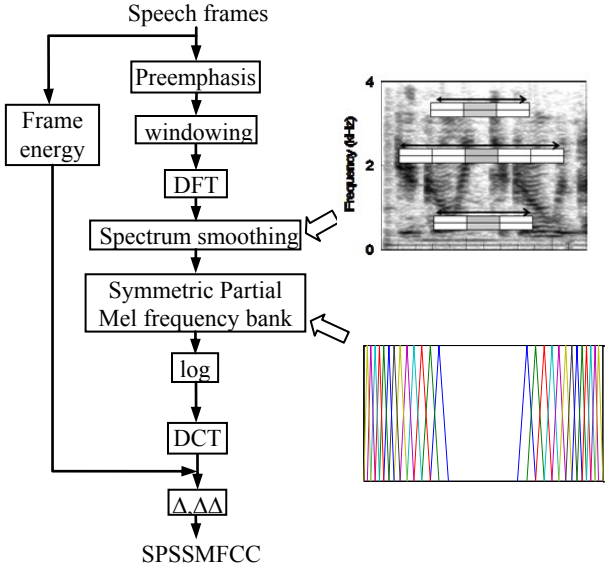

Fig. 2: Diagram of the SPSSMFCC calculation

vector y_t. The centralized first-order Baum-Welch statistics are also needed in i-Vector estimation:

$$\tilde{I} \quad \overline{t=1} \quad c \mid y_t, \Omega)(y_t - m_c) \tag{7}$$

where m_c is the mean of UBM mixture component c. Given the above statistics, the i-Vector for an utterance can be obtained using the following equation:

$$w = (I + T^t \Sigma^{-1} N(u) T)^{-1} T^t \Sigma^{-1} \tilde{I} \tag{8}$$

where N(u) is a diagonal matrix of dimension CF*CF whose diagonal blocks are N_cI (c=1,…,C). $\tilde{I}$ is a vector of dimension CF*1 obtained by concatenating first-order Baum-welch statistics $\tilde{I}$ for a given utterance u. Σ is a diagonal covariance matrix of dimension CF*CF estimated during the factor analysis training that models the residual variability not captured by the total variability matrix T. The total variability matrix T is trained by the same EM algorithm of the total variability space estimation for speaker recognition assuming that each utterance is shifted by a value within different interval. Detailed total variability matrix T estimation algorithm can be found in [16].

5. PLDA

The probabilistic linear discriminant analysis (PLDA) model is a probabilistic generative model that has been used as the backend strategy for fixed-length input vectors

[18,19,20]. We use PLDA in our system as a backend scoring strategy. The generative model can be described as

$$w = m + Vy + Ux + \varepsilon \quad (9)$$

where V represents the offset subspaces and U represents the channel subspace. y is the offset related hidden variable used to evaluate the log-likelihood ratio for the hypothesis test corresponding to the hypothesis that "the two i-vectors were or were not generated by the same class."

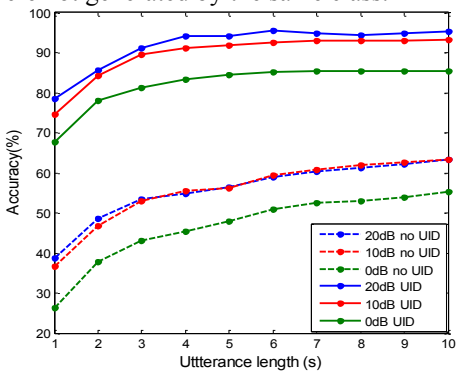

Fig. 3: Results of fine-tuning with and without unique interval detection (UID) using different utterance lengths.

6. EXPERIMENTS AND DISCUSSION

6.1. Frequency offset estimation

The DARPA RATS corpus [21] contains voice communications in various languages transmitted over several adverse radio channels, one of which corresponds to SSB transmission. Performance of the developed system is first evaluated using the data simulated for this channel by artificially applying frequency shifts and additive noise, which is extracted from the SSB channel, to clean source speech. For the training data, shift values are set as the center of each unique interval bin representing the values within the interval; for testing data, shift values are randomly distributed between -300 Hz and 300 Hz. Noise is added at SNR level of 20dB, 10dB and 0dB respectively. Performance is evaluated in terms of the estimation accuracy calculated as the percentage of estimates that fall within 5Hz of the correct value. Fig. 3 demonstrates results of fine-tuning with/without the proposed unique interval detection step with various data lengths. The unique interval detection dramatically improves the estimation accuracy (more than 40% improvement on average). Additionally, the results reach the best value and remain stable after 5s in data length, which is much shorter than other methods which require tens even hundreds seconds of data.

6.2. Application to speaker verification

The proposed frequency shift estimating system is also applied to the SSB channel of the RATS corpora for speaker verification. For each utterance, a 5s segment is used to estimate the frequency offset of the entire speech by the proposed system. Next, the offset is compensated by being shifted with the estimated value in opposite manner by the following equation:

$$y(\mathrm{n}) = \mathrm{Re}\{x(n)\exp(-j2\pi(\Delta f / f_s)n)\} \quad (10)$$

The compensated voice transmissions are processed by speaker verification system using MFCC as acoustic feature.

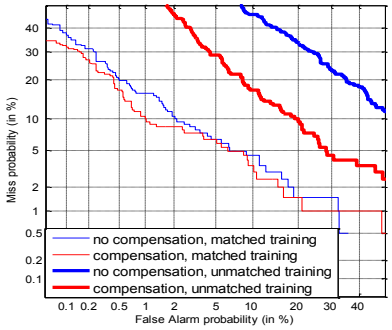

Fig. 4: DET curve under four training and testing conditions

Table 1: EERs (%) with two training conditions

	Unmatched training	Matched training
w/o compensation	26.1	5.8
w compensation	13.1	5.5

A 256-mixture, gender-independent UBM was trained using the training data. The UBM means were used to train a 400-dimensional i-Vector using the development set. The resulting i-Vectors were then used to train a PLDA system, producing a 100-dimensional subspace for the final scoring.

The results are compared according to whether the testing data are compensated using the proposed estimate algorithm under two conditions: (1) training using clean source speech unmatched with distorted/compensated testing data, and (2) training using compensated speech matched with testing data. Fig. 4 plots DET curve of speaker verification results. Table 1 shows the numerical results of EERs. Improvement was observed in either training condition when the testing voice transmissions are compensated, and the improvement is significant when training and testing are unmatched (a relative improvement of +5% for unmatched training and 50% for unmatched one).

7. CONCLUSION

In this study, we proposed a two-step strategy for estimating frequency offsets in a SSB Mod/DeMod communication channel. First, a unique interval is detected for each shifting degraded speech segment in which the estimation ambiguity can be eliminated. Next, fine-tuning is carried out within the unique interval. A statistical method is developed in the first step. We proposed a novel acoustic feature, SPSSMFCC, which can effectively represent different frequency shifts, as proven by our experiments. The total variability method and PLDA techniques are also used in the unique interval detection. The compensation according to the offset estimated by the proposed system is shown to be able to improve the speaker verification performances in both the matched (+5% relative improvement) and the unmatched training condition (50% relative improvement).

8. REFERENCES

[1] 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.

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

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

[4] 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

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

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

[7] H. Xing and P. Loizou, “Frequency shift detection of speech with GMMs and SVMs”, in Signal Processing System (SIPS), workshop on, IEEE, 2012, pp.215-219

[8] G. Liu, et al., "UTD-CRSS systems for NIST language recognition evaluation 2011", NIST 2011 Language Recognition Evaluation Workshop, Atlanta, USA, 6-7 Dec. 2011.

[9] Y. Lei, et al., “The CRSS Systems for the 2010 NIST Speaker Recognition Evaluation,” NIST 2010 Speaker Recognition Evaluation Workshop, Brno, Czech Republic, 24-25 Jun. 2010.

[10] R. Saeidi, et al., "I4U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification", in Proc. INTERSPEECH, Lyon, France, 25-29 Aug.,2013.

[11] J.W. Suh, S. Sadjadi, G. Liu, T. Hasan, K.W. Godin, and J.H.L. Hansen, "Exploring Hilbert envelope based acoustic features in i-vector speaker verification using HT-PLDA", SRE2011 Workshop,Atlanta, USA

[12] G. Liu, Y. Lei, John H.L. Hansen, “Robust feature front-end for speaker identification,” in Proc. ICASSP, Kyoto, Japan, 2012, pp.4233-4236.

[13] G. Liu, John H. L. Hansen. “A systematic strategy for robust automatic dialect identification”, EUSIPCO, Barcelona, Spain, 2011, pp.2138-2141

[14] D. Najim, et al. “Front-end factor analysis for speaker verification,” Audio, Speech, and Language Processing, IEEE Transactions on , vol. 19, no. 4, pp. 788-798, 2011

[15] C. Yu, G. Liu, S. Hahm, and J.H.L. Hansen, "Uncertainty Propagation in Front End Factor Analysis For Noise Robust Speaker Recognition," in Proc. ICASSP, Florence, Italy, May 2014.

[16] K. Patrick, G. Boulianne, and P. Dumouchel. “Eigenvoice modeling with sparse training data,” Speech and Audio Processing, IEEE Transactions on, vol. 13, no. 3, pp. 345-354, 2005

[17] V. Hautamaki, KA. Lee, D. Leeuwen, R. Saeidi, A. Larcher, T. Kinnunen, T. Hasan, SO. Sadjadi, G. Liu, H. Boril, John H.L. Hansen and B. Fauve, “Automatic regularization of cross-entropy cost for speaker recognition fusion”, in Proc. INTERSPEECH, Lyon, France, 25-29 Aug., 2013.

[18] I. Sergey. “Probabilistic linear discriminant analysis,” Computer Vision–ECCV 2006. Springer Berlin Heidelberg, 2006, pp. 531-542

[19] G. Liu, T. Hasan, H. Boril, J.H.L. Hansen, “An investigation on back-end for speaker recognition in multi-session enrollment,” in Proc. ICASSP, Vancouver, Canada, May 25-31, 2013. pp. 7755-7759.

[20] T. Hasan, SO. Sadjadi, G. Liu, N. Shokouhi, H. Boril and J. H.L. Hansen, “CRSS systems for 2012 nist speaker recognition evaluation,” in Proc. ICASSP, Vancouver, Canada, 2013, pp. 6783-6787.

[21] K. Walker and S. Strassel, “The RATS radio traffic collection system,” Odyssey: The Speaker and Language Recognition Workshop, 2012.