ANALYSIS OF PROCESSING DEGRADED SPEECH BY SINGLE FREQUENCY FILTERING

[1]V VIJAYA KUMAR, [2]D KIRTANA, [3]P KIRANMAYEE
[1,2,3]Assistant professor, ECE Department, St.Martin's Engineering College

ABSTRACT

This study proposes new methods for signal processing to highlight some robust speech features in the degraded voice. This takes into account different forms of degradations occurring in action. Single frequency filtering (SFF) of speech signal is the basis for signal processing. At the desired frequency with high-frequency resolution, the SF F output provides the magnitude or cover and the speech signal level. The frequency's SFF output provides some parts with a high signal to noise (SNR) relationship, as the noise power of the single frequency resonator of the bandwidth is very low, while the signal portion would have a high power when it does occur. The high SNR regions are also for different frequencies at different times. This property of the SFF speech analysis is used to extract many robust characteristics from degraded expression, irrespective of the form and scale.

1. INTRODUCTION

SER has many real-life uses. It can be beneficial when natural interaction between humans and computers is required. The detected user emotion will enable the program to respond to the user's inquiry in computer tutorial applications[11]. In the onboard system of a vehicle, the protection of a passenger may be initiated depending on its mental status[33]. It can be used in psychiatric diagnosis by medical professionals as a diagnostic tool[13] It can help effectively express the emotions between two parties in automated translation systems[2]. Speech is the product of the system of dynamic vocal tract, and speech characteristics occur in various ways and with varying amplitudes or levels of energy. The time-different characteristics of the features should therefore be extracted from the time-different signal-to-noise ratio (SNR) of the speech signals, particularly where unknown degradations harm the speech signals. The difficulty in speech processing lies in resolving the specific signal-to-noise ratio of degraded signal (SNR), in particular for the extraction of features. Speech-specific features may suit the various characteristics of speech manufacture (articulatory location and motion), and may also occur at different frequency and time of signal resolution. Robust speech-specific features must be defined and methods must be explored for extracting these features from the speech signals.

The most critical issue for the development of speech systems is inequality of speech and non-discrimination (VAD). The presence of high SNR regions in the SFF outputs at different frequencies is used to differentiate between speaking and non-speaking regions in degraded signals. Usually the speaking regions have differences between samples on a particular frequency and over a certain time sample at a given frequency. The envelope variation across frequency is therefore much higher in spoken regions and lower in noise regions, at a given time and over time, for a given frequency. This spoken property discriminates regions for speech signals compromised by different forms and noise levels. The findings are equal or better than other state-of-the-art VAD, multiple noise and distant speaking methods. The approach is not based on noise sensitivity levels nor does it depend on the form and frequency of the noise.

The weighting functions are used to obtain an improved signal for expression. In order to assess the extent and ease of listening, the improved signal is checked by listening to the degraded signal. It has been shown, in contrast to the deteriorated speech signal, that the proposed improvement approach substantially increases the level of listening comfort.

The most important contributions of this paper are:

(a) The new Single Frequency Filtering (SFF) approach is proposed that provides both time- and frequency-related high signal-to-noise (SNR) regions for speaking with different types of degradations.

(b) The use of high SNR features in the degraded speech SFF outputs is proposed as a new form of speech / no speech detection. The cycle works for each type of degradation and for every other form of degradation without precise tuning.

(c) The high SNR function of the SFF output is further used by use of the data at the rate that gives the most significant SNR in that segment to estimate the fundamental frequency (fo).

(d) For eliminating the location of the major impulse-like excitation during a glottal cycle, the noise correction technique proposed for Voice Aktivity Detection (VAD) is used. The noise compensated envelopes are because the slopes of the spectral variance measured according to time vary distinctly.

2. LITERATURE REVIEW

There have been many attempts to improve VAD efficiency, using speech and noise statistics. Models of discrimination against speech and expression include Artificial Neural Networks (ANNs), Gaussian mixture models (GMMs). One such approach is the VAD based on the statistical model, and suggested refinement. Statistical approaches perform well when labelled language and non-speech training data are available for training models under various noise conditions. Those are referred to as supervised methods of learning. For initialization, the noise model derived from training data is often used. Semi-supervised learning is called such techniques. Universal speech modeling methods are also proposed without any specific noise form being assumed. The method is used to construct a universal speech model through non-negative matrix factorisation (NNMF). In action, a VAD algorithm is preferable to work without instruction, i.e. unattended learning.

Most of the VAD methods are tested by adding noise to data with simulated degradation or by passed the clean signal through a degrading channel. In contrast with known / current methods, this is important to test new methods. Very little has been done to test the performance of VAD algorithms in realistic environments with data collected. Degradations can not match any typical model in these environments. In addition, the task of evaluating VAD methods is difficult to find ground facts.

CNNs are also used for speech emotional recognition (SER). Convolutionary neural networks are used. In such cases, the short-range fourier spectrogram (STFT) is the best option for voice representation, which is fed as a CNN feed. Nevertheless, Fourier's short time ambiguity rules preclude it from simultaneously obtaining time and frequency resolutions. The recently introduced SFF spectrum is, on the other hand, a better option, because it simultaneously measures time and frequency resolutions.

We discuss the SFF continuum as an alternate representation of SER's speech in this study. Our SFF spectrum was adjusted by taking the average of sample amplitudes between two successive glottal closure sites (GCI). Two successive GCI locations provide a pitch which motivates us to refer to the modified SFF spectrogram as an SFF spectrogram with pitch-synchronized data. The GCI sites were found using the filtering method for zero frequencies.

In the glottal closure instant (GCI) impulse-like features of excitation occur as vibrating vocal folds sharply close each glottal process. The GCIs are measured from the arousal portion of the speech signal and the arousal portion is extracted by reverse filters or variants. In this article we suggest a GCI detection approach based on a single frequency (SFF) filtering of the voice signal. In speaking regions, the SFF output has a high signal-to-noise (SNR) quality. The variation of the contours of the SFF output (crosse frequency) indicates rapid shifts around the GCIs, which can be observed even with a decomposition of the speech signal. Through the SFF study, even degraded speech can be extracted from GCI locations. For some instances of speech loss, the robustness of the system is demonstrated.

3. SINGLE FREQUENCY FILTERING METHOD

The Speech Signal has high SNR regions for various frequencies at different times. The chapter emphasizes the importance of single frequency processing speech to capture the high SNR regions of the speech signal at different frequencies. The variable speaking signal SNR properties at various frequencies must be used to extract robust features. The single frequency filtration method (SFF) is suggested in this chapter to extract the speaking signal's time envelopes at the desired frequency in a high resolution. The SFF approach uses a bandwidth resonator near zero and extracts information at a particular frequency with high power from the speech signal (where present). The amplifier envelopes obtained using the SFF method are further processed for the different studies attempted for the thesis to extract speech-specific features.

3.1 Basis for processing speech at single frequencies

Speech signal has time and frequency dependencies. This implies that the signal to noise power ratio is both a time and a frequency measure. The power is divided equally between frequencies for the ideal noise (white noise) of a given total power, whereas the power is not uniformly distributed around the frequency for a signal.

Let

$$\alpha = \int_{f_0}^{f_L} \frac{S^2(f)}{N^2(f)}\, df,$$

$$\beta = \sum_{i=0}^{L-1} \frac{\int_{f_i}^{f_{i+1}} S^2(f)\, df}{\int_{f_i}^{f_{i+1}} N^2(f)\, df},$$

3.2 Single frequency filtering (SFF) method

Single frequency filtering (SFF) method gives amplitude envelopes (ek[n]) at each samplen at the selected frequency fk. with a pole close to the unit circle, and extracts informationat the highest carrier frequency (i.e., half the sampling frequency). Since the same filterat a fixed frequency is used to derive the amplitude envelopes at different frequencies,it avoids the different gain effects, if separate filters were chosen for each frequency toderive amplitude information. The shape of the filter and its gain vary if different filtersare designed to extract information at different frequencies as in the case of filter bankapproaches (eggammatone filter bank method [73]). The SFF method is explained below.The discrete-time speech signal x[n] at the sampling frequency fs is multiplied by acomplex sinusoid of a given normalized frequency ω¯ k to give xk[n]. The time domainoperation is given by

$$x_k[n] = x[n]e^{j\bar{\omega}_k n},$$

$$\bar{\omega}_k = \frac{2\pi \bar{f}_k}{f_s}.$$

3.3 Speech/nonspeech discrimination indegraded speech

Speaking and non-speaking regions are classified by using features derived from specimen and temporal characteristics of degraded voice, supported by statistical models.Statistical training models using Gaussian (GMM), hidden Markov (HMM), deep neural networks (DNN) etc. have gained popularity in recent years. In recent years it has become increasingly popular. For any noise type and at each SNR, models are equipped. It is not possible in growing setting to obtain marked training data. Some methods attempt, assuming it was only non-speaking, to estimate the characteristics of nonspeak from the early seconds of the degraded voice. The starting conditions for non-speech are not always valid.

In addition, degradation characteristics are uncertain in advance and can not therefore be estimated.Several attempts have been made to detect voice activity (VAD), which primarily reports on a small range of simulated degradations. Simulated degradations are degradations observed in a certain environment and subsequently integrated into the voice signal in order to simulate the effect of degradation. VAD method efficiency was focused on the modeling of different degradation characteristics. A sequence of simulated degradations can not reflect the results in realistic environments of multiple degradations.

4. PROPOSED APPROACH

The study we propose uses a time-frequency range for SER, SFF. Amplitude of a signal in increasing frequency is extracted from the SFF spectrogram in function of the time. Therefore, in an emotional voice, the SFF spectrograph captures the temporal variability in each sample. Remember, the discrete Fourier Transform (DFT) can also be collected over a block of data each time the sample is sampled. But, since DFT is performed in each sample, this method is calculationally costly. At each sampling point, the SFF spectroprogram extracts the amplitude envelope and generates a huge matrix of features. Computationally, using such a large feature matrix as a CNN input is unworkable. To resolve this, for all samples between two GCI locations the amplitude envelope has been calculated (referred to as the pitch period). It results in a pitch-synchronous SFF spectrogram. The output representation is

Pitch synchronous SFF spectrogram

The steps to get the spectrogram output from SFF are given below.

1) The speech signal s[n], sampled at frequency fs Hz,is pre-emphasized (p[n]) to remove the low frequency
bias from the speech signal.

$$p[n] = s[n] - s[n-1]$$

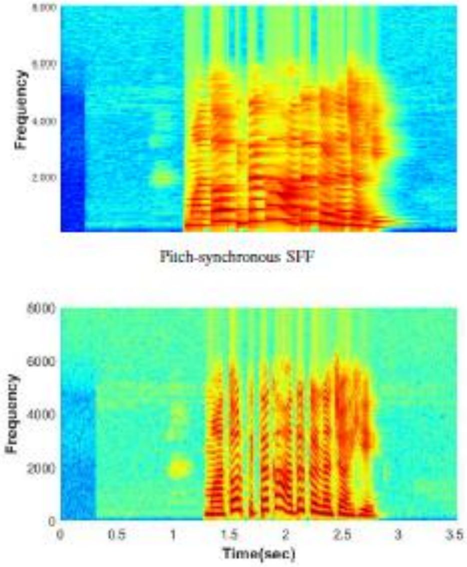
2) The pre-emphasized speech signal p[n] is multiplied bya complex sinusoid of normalized shifted frequency as
follows:

$$p[\bar{k}, n] = p[n]e^{j\bar{\omega}_k n}$$

The output of the filter is given by

$$y[k, n] = -ry[k, n-1] + p[\bar{k}, n]$$

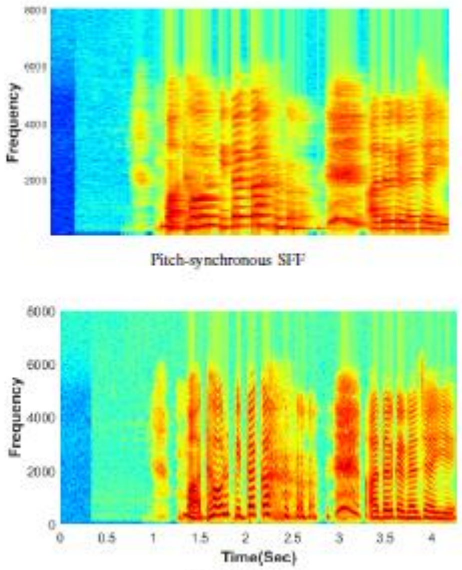
5. EXPERIMENTAL RESULTS



Pitch-synchronous SFF

STFT Spectrogram

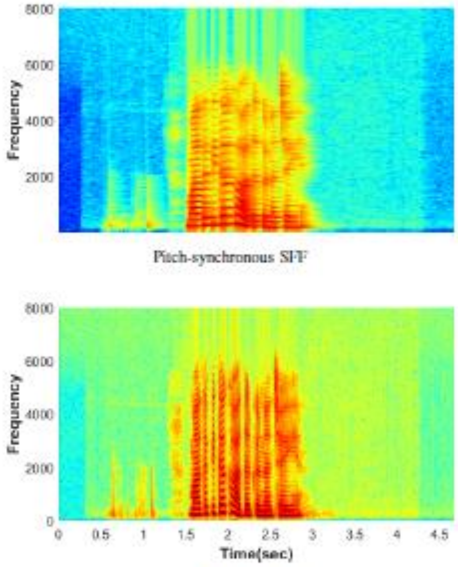
(a) Anger



Pitch-synchronous SFF

STFT Spectrogram

(b) Happy



Pitch-synchronous SFF

STFT Spectrogram

(c) Neutral

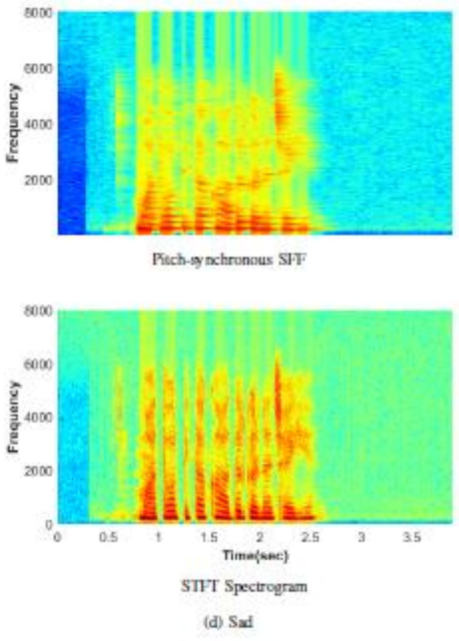


Fig. 1: The pitch-synchronous SFF and STFT spectrograms of the (a) anger, (b) happy, (c) neutral, and (d) sad emotions. Ineach of the sub figures, the top panel shows the corresponding pitch-synchronous SFF spectrogram while the bottom panel shows the corresponding STFT spectrogram.

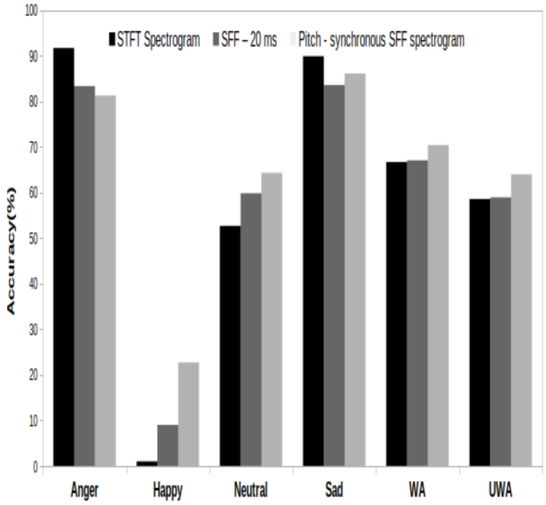


Fig. 2: Emotion classification performance(%) using STFT spectrogram, SFF-20 ms spectrogram and pitch-synchronous SFF
Spectrogram

CONCLUSION

The primary objective of the analysis is to establish resilient to degradational characteristics. Apart from simulated degradations, it is important to target real world degradations. Most approaches deal with simulated degradations in which the voice signal is introduced artificially and are typically segregated by function distribution. Speech devices without pre-data or expertise must be checked under uncertain conditions. A method of signal processing is extracted that highlights important speech signal events. Speech is susceptible to various degradations in real life, and degradations affect the derived features. Individuals are unable, due to their complexity and semantic awareness, to understand and process language for different applications. Nonetheless, robust features must be built for language systems to function through degradations. We also suggested a new SER pitch-synchronous SFF range to address these inconveniences. In deep neural networks, we have tried to solve the SER problem. It is a three CNN building blocks followed by a completely linked layer and an output layer that can be used to measure four emotions (e.g. Rage, Good, Negative, Sad).

REFERENCES

[1] T. Kinnunen, E. Chernenko, M. Tuononen, P. Fr¨anti, and H. Li, “Voice activitydetection using MFCC features and support vector machine,” Proc. Int. Conf. SpeechComput., vol. 2, pp. 556–561, Oct. 2007.

[2] I. McCowan, D. Dean, M. McLaren, R. Vogt, and S.Sridharan, “The delta-phasespectrum with application to voice activity detection and speaker recognition,” IEEETrans. Audio, Speech, Lang. Process., vol. 19, no. 7, pp. 2026–2038, Sep. 2011.

[3] J. Haigh and J. Mason, “Robust voice activity detection using cepstral features,” inProc. IEEE TENCON, 1993, pp. 321–324.

[4] N. Dhananjaya and B. Yegnanarayana, “Voiced/nonvoiced detection based on robustnessof voiced epochs,” IEEE Signal Process. Lett., vol. 17, no. 3, pp. 273–276,Mar. 2010.

[5] T. Pham, M. Stark, and E. Rank, “Performance analysis of wavelet subband basedvoice activity detection in cocktail party environment,” in Proc. Int. Conf. Comput.andCommun. Technologies, Oct. 2010, pp. 85–88.

[6] Z. Song, T. Zhang, D. Zhang, and T. Song, “Voice activity detection using higherorderstatistics in the teager energy domain,” in Proc. Wireless Commun. SignalProcess., Nov. 2009, pp. 1–5.

[7] Y. W. Jitong Chen and D. Wang, “A feature study for classification-based speechseparation at low signal-to-noise ratios,” IEEE Trans. Audio, Speech, Lang. Process.,vol. 22, no. 12, pp. 1993–2002, Dec. 2014.

[8] X.-L. Zhang and D. Wang, “Boosted deep neural networks and multi-resolutioncochleagram features

for voice activity detection," in Proc. Interspeech, Sep. 2014,pp. 1534–1538.

[9] J. Ramrez, J. C. Segura, C. Bentez, A. D. L. Torre, and A. Rubio, "Efficient voice activitydetection algorithms using long-term speech information," Speech Commun.,vol. 42, pp. 3–4, April 2004.

[10] Y. Ma and A. Nishihara, "Efficient voice activity detection algorithm using longtermspectral flatness measure," EURASIP Journal on Audio, Speech, Music Process.,vol. 2015, no. 1, p. 71, 2015.