F0 ESTIMATION FROM TELEPHONE SPEECH USING DEEP FEATURE LOSS

Supritha M Shetty[†]*, Shraddha Revankar*[⋆]*, Nalini C Iyer*[⋆]*, K T Deepak*[†]

[†]Electronics and Communication Engineering, IIIT Dharwad
[⋆] School of Electronics and Communication Engineering, KLE Technological University
{supritha.shetty,deepak}@iiitdwd.ac.in, {shraddha.revankar, nalinic}@kletech.ac.in

ABSTRACT

Accurate pitch estimation in speech signal plays a vital role in several applications. Robust pitch estimation in telephone speech is still a challenge due to the narrow bandwidth of the signal. Electroglottograph (EGG) signal is a reliable means for pitch estimation, however, it's not practically possible to measure such a signal in many applications. In this work, a method is proposed to synthesize EGG signal from telephone speech using deep feature loss network and subsequently pitch contour is derived from synthesized EGG (SEGG) signal. In order to evaluate the proposed work, CMU-Arctic speech database is used as it contains simultaneous EGG signal recorded. The telephonic speech is derived using International Telecommunications Union ITU-T as specified in the Blizzard Challenge. The robustness of the proposed method is demonstrated under different noisy conditions. The performance of the proposed work is encouraging when compared with other state-of-the-art methods.

Index Terms— Pitch estimation, electroglottograph, telephone speech, deep feature loss.

1. INTRODUCTION

Pitch is an important attribute of speech signal and plays a crucial role in applications as gender identification, speaker verification, speech coding, intonation, speech synthesis, and in speech recognition of tonal languages [1, 2]. It is essentially a measure of frequency at which the vocal chords vibrate to produce voiced sounds, often referred as fundamental frequency ($F0$) [3]. Pitch is attributed more to perceptual phenomena. However, both pitch and fundamental frequency are interchangeably used in the literature. There is pitch variation while speaking, and hence the pitch contour estimation is a challenging task. It is a bigger challenge to estimate the pitch contour in case of telephone speech signal. Due to the telephone speech bandwidth of 300-3400 Hz, most likely the fundamental frequency component is weak or completely lost [4]. Thus achieving an accurate $F0$ is one of the important problems associated with telephone speech signal analysis and processing.

Pitch estimation is relatively a well studied topic and people have worked on this for more than 5 decades. Many approaches have been proposed since. Majority of the proposed methods channelize the speech signal towards a synthetic function of time that helps estimating the pitch contour. This is achieved by processing the speech signal either in temporal or spectral domains. Autocorrelation function (ACF) [5] based approach is one of the earliest methods followed by cepstrum analysis [6], average magnitude difference function (AMDF) [5], normalized cross-correlation function (NCCF), etc. The periodicity property of speech signal is explored for estimating pitch candidates. The authors in [7] proposed a robust algorithm (RAPT) that extract pitch candidates based on NCCF over voiced speech. A relatively simpler algorithm with few parameters that uses the cumulative mean normalized difference function over voiced speech is proposed in YIN [8]. More recently a saw tooth waveform based spectrum is subjected through template matching in Sawtooth Waveform Inspired Pitch Estimator (SWIPE) [9]. The SWIPE fundamentally uses the first and prime harmonics of the signal and thereby reduces the errors which usually affects other methods. However, since the majority of the fundamental frequency is weak or non-existential in case of telephone speech signal, very few methods are present in the literature that specifically addresses the pitch extraction in telephone speech signal. The speech signal is subjected through the non-linear processing to recreate the missing fundamentals in case of Yet Another Algorithm for Pitch Tracking (YAAPT) [10]. In YAAPT pitch estimation is achieved using NCCF similar to RAPT approach. The method is claimed to be highly robust for pitch extraction in telephone speech signal. With the increased computational power, it is possible to use trained neural network models to extract pitch contour [11–14].The authors in [11] explored the shift-invariant property of CNN network which could be useful in pitch estimation. Also in [12], the authors use RNN that captures the temporal dynamics of speech data which is a key point in pitch tracking. Similarly, DNN is used to model the posterior probability of pitch states given the observation in [12].

Instead of speech signals, some approach use excitation signal to determine pitch. Autocorrleation function of excitation signal is explored in SIFT [15] to estimate fundamen-

tal frequency. Also an attempt was made in [16] using the impulse-like nature of the excitation signal i.e GCIs. The time between consecutive GCIs is referred as pitch period and the inverse of it is the fundamental frequency. Similarly, the authors in [17] used spectrum of the LP residual signal to avoid the vocal tract effect.

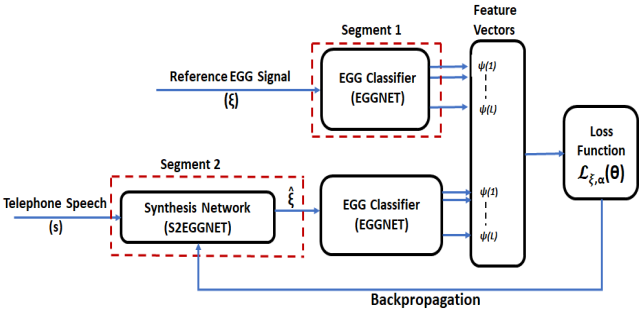

Fig. 1: Block Diagram representation of the training methodology. Mainly consists of Speech to EGG synthesis network, EGG classifier network and Deep Feature loss calculation block

However, methods that are based on harmonics of residual signal highly depend on source-filter decomposition. The source-filter decomposition through inverse filtering techniques fail in high pitch varying datasets such as singing, emotion, etc. Therefore the authors in [18] analyzed the harmonics of impulse like excitation signal derived from modified zero frequency filtering method to estimate the fundamental frequency. It is widely accepted that Electroglottograph (EGG) signal is used for reliable measurement of $F0$ [19]. The EGG signal is essentially a measurement of vocal folds contact area (VFCA). The repetitive cycle due to glottal vibration is the actual measurement of $F0$ contour. Since EGG signal is free from the vocal tract changes, such a signal forms the realistic measurement of $F0$ contour. However, it is always not feasible to obtain EGG signal in most of the practical applications as this requires pair of electrodes strapped around the neck and an EGG equipment. For example, gender identification, speaker verification, speaker identification and speech recognition, etc over telephone network it is not practically possible to record parallel EGG channel and transmit the same. Instead, EGG signal can be synthesized from speech signal. Several research works such as [20–22] have proven that synthesized EGG signal has a close resemblance to the original EGG signal in terms of glottal instants. In this work a preliminary study is reported to synthesize the EGG signal from telephone speech signal using deep feature loss network [23]. The $F0$ contour is extracted from synthesized EGG (SEGG) signal using Autocorrelation function (ACF). It is shown that the SEGG signal can be synthesized directly from the raw telephone speech signal and subsequently used for pitch contour estimation. The experimental results show that the proposed approach gives accurate pitch estimates and provides good generalization ability for unseen speakers.

The rest of this paper is organized as follows: In Section 2, a brief introduction to the proposed approach for F0 estimation is discussed. Also, a brief description about pitch contour estimation from synthesized EGG signal is discussed here. In Section 3, we present the evaluation of the proposed scheme. Finally, the paper is summarized and concluded in Section 4.

2. EGG SIGNAL SYNTHESIS FROM TELEPHONE SPEECH

In this work an attempt is made to synthesize EGG signal from the raw telephone speech signal in both clean and noisy dataset using deep feature loss network. Subsequently, the synthesized EGG signal is subjected through NCCF to obtain the pitch contour of the corresponding telephone speech signal.

2.1. EGG synthesis using Deep Feature Loss network

Deep feature loss architecture consist of 2 convolutional neural networks with 2 different objectives. One being the audio classifier network which is trained to classify different audio events. The second network is the transformation network that maps the speech signal to the target signal [22,23]. Figure 1 shows the block diagram of deep feature loss network. Segment 1 is an EGG classifier network named as EGGNET that classifies normal and singing EGG signals. Segment 2 refers to speech to EGG synthesis network named as S2EGGNET. A loss calculation block at the end of the block diagram computes loss incurred by the reference and synthesized EGG signal at every layer of the Segment 1.

2.1.1. EGG classifier Network (EGGNET)

EGGNET is a binary classifier network whose main purpose is to classify original EGG signal into normal and singing/emotion. The network consists of 16 convolutional layers where the first and last layers form the input and output respectively. Each layer has a kernel size of 3x1 and stride of 2x1. Each layer has a Batch normalization (BN) function and adopts Leaky Rectified Linear Unit (LRELU) as an activation function.

2.1.2. Speech to EGG synthesis Network (S2EGGNET)

S2EGGNET is a context aggregation based convolutional network. The network consists of 16 convolutional layers having a kernal size of 3x1. Each intermediary layer has a dilation factor of 2^0 for layer 2 to 2^{12} for layer 14. As the network layer depth increases, so is the dilation factor. The 15th layer has a factor of 2^0. The last layer is the output layer computed as convolution with kernel size of 1x1. Every intermediary layer has the same number of feature maps = 64. Inorder to maintain every layer length same as that of input signal

length, zero-padding is included during the convolution operation. Also due to padding, the audio samples that are close to the sequence boundary are considered during the training process. The sampling frequency used in this work is 16KHz, thus frame size is set to 2^{14} + 1 samples i.e about 1s of the audio. Also each layer has an adaptive batch normalization function and adopts LRELU for layer activation.

2.1.3. Deep Feature Loss Calculation

Deep feature loss is based on the differences in the internal activations of the pre-trained network (which is EGGNET) applied to the signals being compared. This loss value captures the differences between reference EGG and synthesized EGG signal at different layers and thus helps S2EGGNET map telephone speech to EGG signal better. The feature activations of the reference EGG signal and synthesized EGG signal from the EGGNET is the input to the loss function.

Let ψ^l be the internal activations of m^{th} layer in the classification network. The weighted $\mathcal{L}$ loss function computes the difference between the features induced by the original EGG signal ξ, and the synthesized EGG signal $\hat{\xi}$ when applied to the pre-trained classifier network. The loss equation is as follows.

$$\mathcal{L}_{\xi,\alpha}(\theta) = \sum_{m=1}^{L} w_m \left\| \psi^m(\xi) - \psi^m(G(\alpha;\theta)) \right\| \qquad (1)$$

where, θ are the parameters of the synthesis S2EGGNET network; G is the synthesis network; w_m are the weights of the deep feature loss and is given by the inverse relation as $1/\|\psi^m(\xi) - \psi^m(G(\alpha;\theta))\|$. The weights are updated at the tenth epoch of training, while in the case of the first ten epochs, the default values of the weights are set to 1. Inverse weights are assigned to the w_m of the loss function which helps equally balance the contribution of all the layers to the backpropagation. Note that only first 6 layers of the EGGNET is considered for loss computation.

2.2. Pitch contour estimation from telephone speech using synthesized EGG signal

The CMU-Arctic database has two parallel channels *viz.,* speech and EGG signal recorded simultaneously. The steps are adopted from Blizzard challenge [24] to convert the wide band signal to its corresponding telephone speech signal. The telephone speech signal x is passed through the S2EGGNET to obtain synthesized EGG e. The autocorrelation function $R(\tau)$ of the S2EGGNET output signal $e(n)$ is generally defined as Eqn. (2)

$$R(\tau) = \frac{1}{N} \sum_{n=0}^{N-1} e(n)e(n+\tau) \qquad (2)$$

where N is the length of the underlying speech and τ is the lag number. If $e(n)$ is periodic at pitch period T, $R(\tau)$ exhibits peak at $\tau = iT$, where $i = 0, 1, 2, 3, ...$ As the value of τ increases, $R(\tau)$ tends to decrease which facilitates the use of second peak for estimation of the pitch period.

The Figure 2 illustrates the synthesized EGG signal and the $F0$ contour estimation from telephone speech signal. Figure 2(a) shows a segment of telephone speech signal. Figure 2(b) shows the ground truth EGG waveform of the corresponding speech segment. Figure 2(c) shows the pitch contour of the ground truth EGG extracted using PRAAT [25] software. It can be noticed that at the beginning of the first voiced segment there is an increased pitch and slowly the pitch contour reduces towards the end of voicing. A similar pattern can be observed in the second voiced segment.

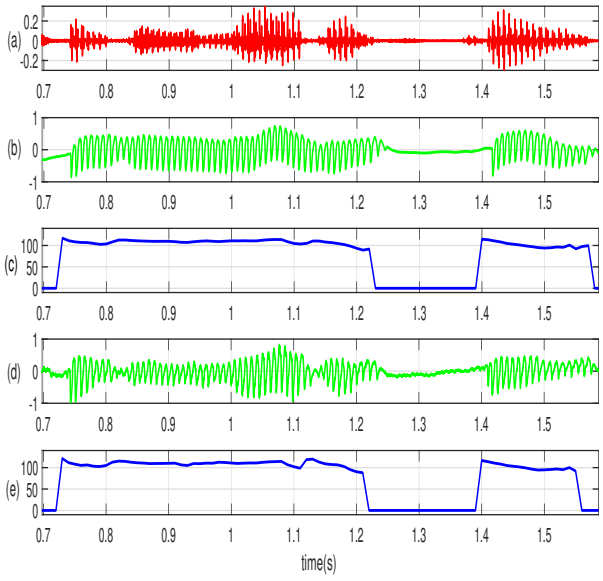

Fig. 2: Illustration of F0 contour of proposed model output. (a) Clean telephone speech signal. (b) Reference EGG signal. (c) F0 contour of reference EGG signal. (d) Synthesized EGG signal. (f) F0 contour of synthesized EGG.

In the Figure 2(d) shows the EGG signal obtained from deep feature loss model. It can be noticed that there is a correspondence between the oscillatory cycles of original EGG signal and the proposed model output. The synthesized EGG signal is subjected to PRAAT to obtain the pitch contour as shown in Figure 2(e). It can be observed from Figure 2(c) and Figure 2(e) that the pitch contours derived from reference EGG and proposed model, respectively are having similar pattern and there is close correspondence between the two. The pitch contour estimated from original EGG signal using PRAAT software is used as the reference. Similarly, synthesized EGG is subjected through PRAAT, to obtain the pitch contour from telephone speech signal. Both are compared and evaluated using Gross Pitch Error (GPE) metric. GPE is a measure of proportion of frames considered as voiced by both reference and estimated for which the relative error

should be higher than the threshold. Further the robustness of the proposed method is evaluated in case of additive white and babble noise.

3. EXPERIMENTAL RESULTS AND DISCUSSIONS

The training and testing of deep feature loss network is done using CMU-Arctic database converted to telephone channel signal. The database consists of 2 channel recordings of speech and EGG signal, respectively. Both channels are recorded at a sampling rate of 32 kHz. For experimentation purpose, the audio files are downsampled to 8KHz. In this work, a total of 4 (3 Male and 1 Female) speakers data is considered for the experiments. The entire dataset has an approximate duration of 3hrs of voiced samples. Note that the EGG signals are time-aligned to compensate the delay between the Electroglottograph and the microphone and an approximate 0.9ms shift is performed on speech data.

It has to be noted that in order to train the deep feature loss network, the telephone speech signal is upsampled from 8kHz to 16kHz. The network is trained in unseen condition wherein both train and test audio files are exclusive to each other. The proposed network is trained using 1 male speaker and the remaining 2 male and 1 female speaker is used for evaluation of the trained model. The learning rate is set to 0.0001 with a batch size of 100 and epochs 400. For comparison of the proposed model result, we chose YAAPT, SWIPE and SRH methods for evaluation. All the above signal processing algorithms are executed in Matlab and tabulated. For neural network comparison, we have considered GAN architecture [21] wherein the GAN network is trained and tested with the same dataset as that of deep feature loss network.

Table 1 shows the results evaluated using clean telephone speech signal for 16 kHz trained model. From the table it is evident that YAAPT performance is better than SWIPE and understandably because, YAAPT is a better method for telephone speech signal. However, SEGAN is slightly better than YAAPT. Also, it can be noticed that the proposed model is significantly better than other state-of-the-art methods. Yet, it can be observed that GPE is higher for SLT. It is understandable that the network performs poorly for cross gender cases.

The robustness of the proposed method is evaluated using a noisy speech dataset which is generated using NOISEX-92. For evaluation of the proposed network, we have chosen white and babble noise with different Signal-to-Noise ratio levels. These noises are added to 1 male speaker with 4 different SNR's (0dB,10dB, 20dB, 30dB) totalling approximately 3.5hrs of noisy speech in each noise type. The performance of the proposed method is shown using the performance plots in Figure 3. From the plots it is observed that the proposed method is comparable with YAAPT algorithm in both white and babble noise for higher SNR levels. However, in case of 0 dB, the proposed model is significantly better than other state-of-the-art approach considered in this work.

Table 1: Comparision table of Gross pitch error

Methods	Speakers	GPE(%)	Average GPE(%)
YAAPT	JMK	5.4	3.43
	SLT	1.8	
	KED	3.1	
SWIPE	JMK	6.4	4.66
	SLT	4.0	
	KED	3.6	
SRH	JMK	22.10	21.2
	SLT	9.8	
	KED	31.7	
SEGAN	JMK	2.5	3.0
	SLT	4.8	
	KED	1.7	
Proposed Model	JMK	1.5	**2.38**
	SLT	2.64	
	KED	0.3	

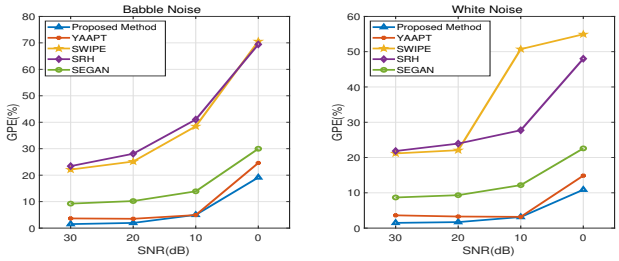

Fig. 3: Performance evaluation of pitch estimation in additive Babble and white noise conditions.

4. SUMMARY AND CONCLUSIONS

A novel method is proposed to estimate the $F0$ contour from telephone speech signal. The method encompasses synthesizing EGG signal using the deep feature loss network. The study has shown the usage of synthesized EGG for estimating pitch contour. The proposed network is evaluated in both clean and noisy speech dataset. However, there are other parameters which can be extracted from synthesized EGG and requires a detailed study. This is a preliminary study in a limited context confined to CMU-Arctic database. But the results are encouraging and it requires a detailed analysis in varying pitch conditions such as emotional speech and pathological conditions.

5. ACKNOWLEDGEMENT

This work is funded by DST to IBITF under the project titled "Speech and Text Analytics for Business Intelligence" under PRAYAS scheme and MeitY under the project titled "National Language Translation Mission (NLTM) : BHASHINI"

6. REFERENCES

[1] Leena Mary and B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition," *Speech Communication*, vol. 50, pp. 782–796, 2008.

[2] Morris S. F. Poon and Manwa L. Ng, "The role of fundamental frequency and formants in voice gender identification," *Speech, Language and Hearing*, vol. 18, 2014.

[3] Randall L. Plant and Ross M. Younger, "The interrelationship of subglottic air pressure, fundamental frequency, and vocal intensity during speech," *Journal of Voice*, vol. 14, pp. 170–177, 2000.

[4] M. Chandni and D. Govind, "Effectiveness of wavelet synchrosqueezed transform for improved epoch estimation from telephonic speech signals using zero frequency filtering," in *IEEE 18th India Council International Conference (INDICON)*, 2021.

[5] Xiao-Dan Mei, Jengshyang Pan, and Sheng-He Sun, "Efficient algorithms for speech pitch estimation," in *Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing.*, 2001.

[6] S. Ahmadi and A.S. Spanias, "Cepstrum-based pitch detection using a new statistical v/uv classification algorithm," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 333–338, 1999.

[7] David Talkin, "A robust algorithm for pitch tracking (rapt)," *Speech coding and synthesis*, 1995.

[8] Alain Cheveigné and Hideki Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, pp. 1917–30, 2002.

[9] Arturo Camacho and John Gregory Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 124, pp. 1638–52, 2008.

[10] Kavita Kasi and Stephen A. Zahorian, "Yet another algorithm for pitch tracking," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002.

[11] Hong Su, Hui Zhang, Xueliang Zhang, and Guanglai Gao, "Convolutional neural network for robust pitch determination," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[12] Kun Han and Deliang Wang, "Neural networks for supervised pitch tracking in noise," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[13] J. Wook Kim, J. Salamon, P. Li, and J. Pablo Bello, "Crepe: A convolutional representation for pitch estimation," *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[14] Atte Virtanen, "Robust f0 estimation of telephony speech using artificial neural networks," in *MSc Thesis*, 2020.

[15] J. D. Markel, "The sift algorithm for fundamental frequency estimation," *IEEE Transactions on Audio and Electroacoustics*, vol. 20, 1972.

[16] B. Yegnanarayana and K. S. R. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Transactions on Audio, Speech Language Processing*, vol. 17, pp. 614–624, 2009.

[17] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," *Interspeech*, 2011.

[18] Sudarsana Reddy Kadiri and B. Yegnanarayana, "Estimation of fundamental frequency from singing voice using harmonics of impulse-like excitation source," *Interspeech*, 2018.

[19] Christian Herbst and Jacob C. Dunn, "Fundamental frequency estimation of low-quality electroglottographic signals," *Journal of Voice*, vol. 33, pp. 401–411, 2018.

[20] Prathosh A. P., V. Srivastava, and M.Mishra, "Adversarial approximate inference for speech to electroglottograph conversion," *IEEE/ACM Trans. Audio Speech and Language Processing*, vol. 27, pp. 2183–2196, 2019.

[21] K. T. Deepak, P. Kulkarni, U. Mudenagudi, and S. R. M. Prasanna, "Glottal instants extraction from speech signal using generative adversarial network," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[22] Supritha M Shetty, Suraj Durgesht, and K T Deepak, "Glottal instants extraction from speech signal using deep feature loss," *IEEE International Conference on Signal Processing and Communications (SPCOM)*, 2022.

[23] François G. Germain, Qifeng Chen, and Vladlen Koltun, "Speech denoising with deep feature losses," in *Interspeech*, 2019.

[24] Marc Schröder, Sathish Pammi, and Oytun Türk, "Multilingual mary tts participation in the blizzard challenge 2009," in *Blizzard Challenge*, 2009.

[25] Yannick Jadoul, Bill Thompson, and Bart de Boer, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics*, pp. 1–15, 2018.