2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

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

Introduction

- F0 in speech signal is the measure of vibration of the vocal cords produced during voiced sound production.
- F0 is one of the crucial parameters in applications like speaker verification, speech synthesis, speech recognition of tonal languages, music information retrieval, etc.
- Telephone-coded speech signal has a bandwidth range from 300Hz-3.4kHz. Most of the F0 information is lost in such signals.
- Therefore robust pitch estimation in telephone speech is still a challenge due to the narrow bandwidth of the signal.
- Most of the traditional F0 estimation methods generate an intermediate signal which is a synthetic function of time or explore the periodicity property of speech signal for estimating the F0 values.
- Due to the learning ability in the neural networks, it is possible to extract F0 contours from raw speech signals [3, 6].
- Instead of the speech signal, it is also possible to extract F0 from the impulse-like structure of the excitation signal [7]. However, methods based on excitation signal largely depends on the source-filter decomposition technique to get an accurate output.
- 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.

Motivation

- Telephone-coded speech signal has little information on F0 due to its narrow channel bandwidth.
- Pitch can be extracted from the excitation signal. However, this approach fails to extract the F0 information for high-pitch varying conditions.
- Electroglottograph (EGG) signal is a reliable means for pitch estimation [4], however, it's not practically possible to measure such a signal in many applications. Therefore, an ideal approach is synthesizing EGG from the raw speech signal and extracting the parameters[1, 5].
- 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.

Main Objectives

- Synthesize an EGG-like signal from the corresponding telephone-coded speech signal.
- Estimate f0 contour from the synthesized EGG signal and evaluate the performance of the proposed model using Gross Pitch Error metric (GPE).

EGG synthesis from Deep Feature Loss Network

- The proposed network consists of 2 blocks [2].
  - EGG classifier network (EGGNET):-
    - A binary classifier network that classifies the normal and singing EGG signals.
    - This network acts as a loss network that provides the loss between ground truth EGG and synthesized EGG across multiple layers during S2EGGNET training.
  - Speech to EGG synthesis network (S2EGGNET):-
    - S2EGGNET is an EGG synthesis network that takes in raw telephone speech signal and generates EGG-like signal.
    - S2EGGNET is a context aggregation based convolution neural network which is trained using deep feature loss.

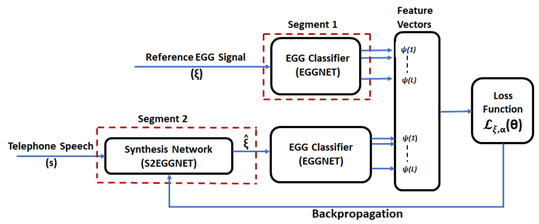

Figure 1: Block Diagram representation of the training methodology. Mainly consists of Speech to EGG synthesis network, EGG classifier network and Deep Feature loss calculation block

Loss Calculation

The overall loss is calculated by applying L1 loss function to the internal activations captured at each layer of the EGGNET model for both ground truth EGG and synthesized EGG signal.

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

Experiments and Results

- The experimental analysis is conducted using CMU-Arctic database.
- The speech signals are converted to telephone channel signal using International Telecommunications Union ITU-T as specified in the Blizzard Challenge. The training and testing of the network is done in the ratio of 1:3 speaker data respectively.
- The performance of the proposed network is evaluated using Gross Pitch Error metric.
- The F0 extracted from the ground truth EGG signal is used as a reference for GPE estimation across all evaluated algorithms.
- Praat software is used to extract the f0 values from EGG signal.

Table 1 shows the overall performance of the proposed network in comparison with other state-of-the-art algorithms.

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

Table 1: Comparision table of Gross pitch error

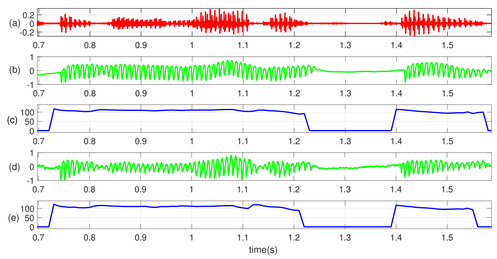

Figure 2: Illustration of F0 contour of proposed model output. (a) Clean telephone speech signal. (b) Reference EGG signal. (c) F0 contour of reference EGG signal. (d) Synthesized EGG signal. (f) F0 contour of synthesized EGG

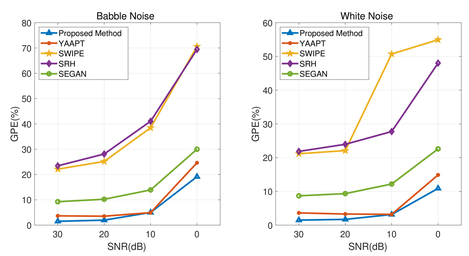

Figure 3: Performance on white and babble noise

The robustness of the proposed work is evaluated by training and testing the network with synthetically generated noise data *viz.,* white and babble noise with 4 different SNRs (0dB, 10dB, 20dB, 30dB).

Conclusion

- The proposed work focuses on synthesizing EGG like signal from telephone speech signal using deep feature loss.
- The analysis of the synthesized EGG-like signal is performed by extracting the fundamental frequency (F0).
- Both clean speech and noisy speech conditions are considered for training and testing the performance of the network.
- Through experimental analysis, it is observed that the performance is comparable to other state-of-the-art methods in terms of GPE (Gross Pitch Error).

Future Work

The results of the experimental study on normal speech conditions is promising, therefore a detailed analysis on varying pitch condition datasets such as emotional speech and pathological conditions would be of great interest.

References

[1] K. T. Deepak, Pavitra Kulkarni, Uma Mudenagudi, and S. R. M. Prasanna. Glottal instants extraction from speech signal using generative adversarial network. *ICASSP*, 2019.

[2] Francois G. Germain, Qifeng Chen, and Vladlen Koltun. Speech denoising with deep feature losses. *Interspeech*, 2019.

[3] Kun Han and Deliang Wang. Neural networks for supervised pitch tracking in noise. *ICASSP*, 2014.

[4] Christian Herbst and Jacob C. Dunn. Fundamental frequency estimation of low-quality electroglottographic signals. *Journal of Voice*, 2018.

[5] Supritha M Shetty, Suraj Durgesht, and K T Deepak. Glottal instants extraction from speech signal using deep feature loss. *SPCOM*, 2022.

[6] 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.

[7] B. Yegnanarayana and K. S. R. Murty. Event-based instantaneous fundamental frequency estimation from speech signals. *IEEE Transactions on Audio, Speech Language Processing*, 2009.