An Investigation into Instantaneous Frequency Estimation Methods for Improved Speech Recognition Features

Shekhar Nayak, Saurabhchand Bhati, K. Sri Rama Murty

Department of Electrical Engineering, Indian Institute of Technology Hyderabad, India

Email: {ee13p1008, ee12b1044, ksrm}@iith.ac.in

Objectives

- To explore different IF estimation methods for improved phase based features for automatic speech recognition (ASR).
- To combine the evidences from magnitude and phase for improving ASR performance.

Motivation

Recent perceptual studies have demonstrated that features from the phase of the speech signal or frequency modulation features from speech significantly enhances the human speech recognition in noise [1],[?]. There is a renewed interest in the analysis of phase spectrum of speech signals [2].

IF estimation methods

IF estimation using zero-crossing method (IF-ZC)

- IF can be estimated from the average number of zero-crossings in a short window.
- Advantage - Simple and computationally efficient
- Limitation -
  - Efficiency of this method depends on the size of the window.
  - Large window violates the local property of IF.
  - A smaller window leads to noisy IF estimates.

IF estimation using least mean squares (LMS) algorithm (IF-LMS)

- IF can be estimated by minimizing the squared instantaneous error between the speech signal and its estimate from a time-varying predictor using LMS algorithm.
- Advantage - Based on adaptive filtering
- Limitation -
  - Its performance is severely affected by the choice of step size involved in gradient descent for updating adaptive filter coefficients.
  - A very small step size cannot track fast varying changes in the IF.
  - A large step size results in noisy IF estimates.

IF estimation using time-varying auto-regressive (TVAR) modelling (IF-TVAR)

- Time-varying predictor coefficients are expressed in terms of basis functions.
- The weights of the basis functions are estimated to compute the predictor coefficients which are used to estimate the IF.
- Advantage - Better modelling of speech signal using time-varying predictor coefficients.
- Limitation -
  - Less number of basis functions fails to track fast variations in IF
  - Higher number of basis functions results in model over-fitting to noise in the data.

IF estimation using Fourier transforms (IF-FT)

- IF can be estimated from analytic phase of speech signals using differentiation property of Fourier transform.
- Advantage - Does not involve any hyper-parameters.
- Limitation - Works well only for synthetic narrowband signals and not for speech-like signals.

IF-FT estimation equations

IF-FT can be estimated in continuous time as -

$$f_i(t) = \phi'(t) = Re\left(\frac{\mathcal{F}^{-1}(j\omega X_a(\omega))}{\mathcal{F}^{-1}(X_a(\omega))}\right) \quad (1)$$

where $Re(.)$ denotes the imaginary part of a complex quantity. IF can be computed in discrete form as

$$f_i(n) = \phi'(n) = \frac{2\pi}{N} Re\left(\frac{IDFT(kX_a(k))}{IDFT(X_a(k))}\right) \quad (2)$$

- IDFT - inverse discrete Fourier transform
- $X_a(k)$ - DFT of analytic signal
- N - length of the signal in samples

Synthetic signal generation with known IF

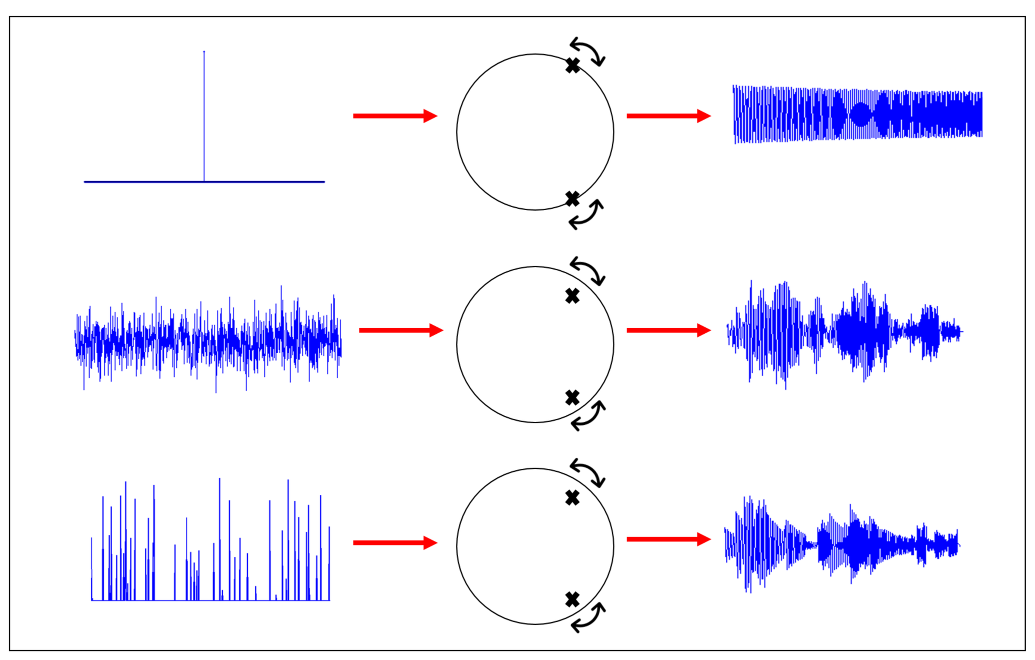

Figure 1: Time varying all-pole excitation system for synthetic signal generation with known IF

- A time-varying all-pole system with a pair of complex conjugate poles at $r[n]e^{\pm j\theta[n]}$ is simulated, whose input $u[n]$ and output $x[n]$ are related by

$$x[n] = 2r[n]\cos(\theta[n])x[n-1] - r^2[n]x[n-2] + u[n] \quad (3)$$

$r[n]$ and $\theta[n]$ control the instantaneous bandwidth and frequency of the output signal $x[n]$.
- Different narrowband signals generated are -

- **Critically damped system with unit sample (CD-US)**
- **Under damped system with random noise (UD-RN)**
- **Under damped system with train of impulses (UD-TI)**

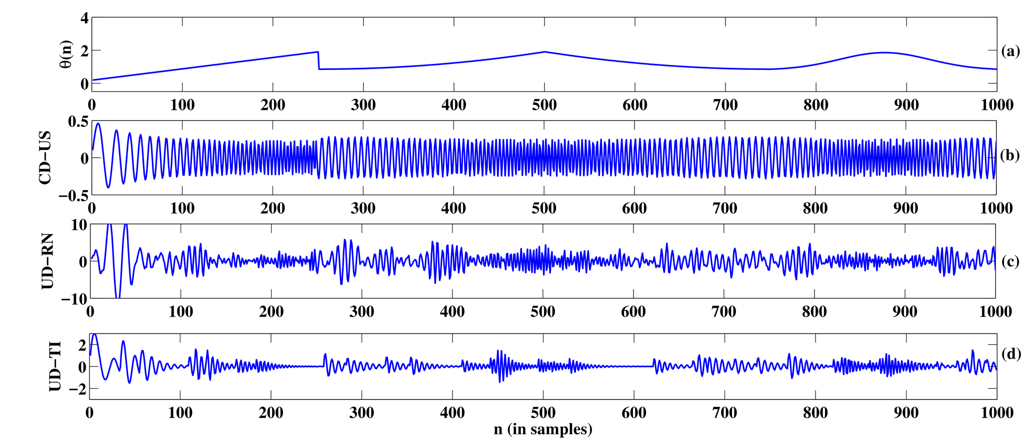

Figure 2: Instantaneous frequency and output of the system for different excitations. (a) IF of the system. System output for - (b) unit impulse (c) random noise (d) train of impulses.

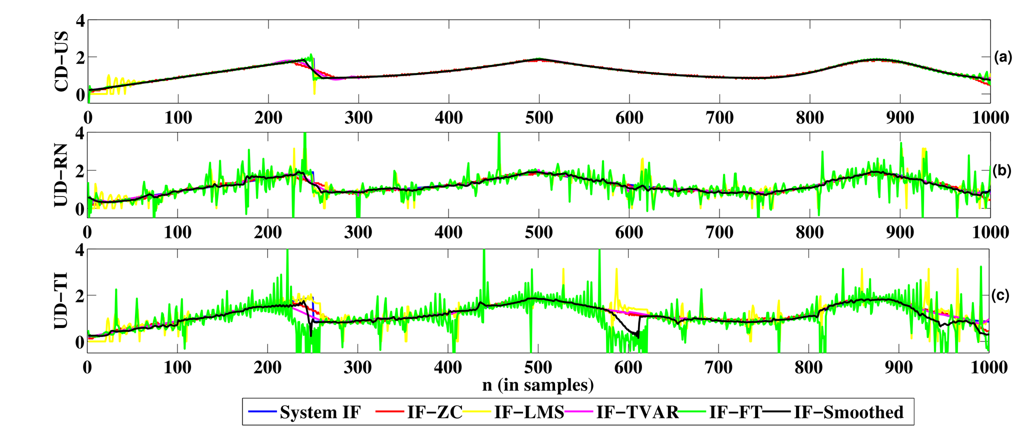

Figure 3: True and estimated IF for Synthetic signal for the three systems. System excited with - (a) unit impulse. (b) random noise. (c) train of impulses.

MSE between true and estimated IF

Features	CD-US	UD-RN	UD-TI
IF-ZC	0.0065	0.0120	0.0085
IF-LMS	0.0043	0.1154	0.0795
IF-TVAR	0.0037	0.0114	0.0239
IF-FT	0.0053	0.3455	0.7411
IF-Smoothed	0.0044	0.0092	0.0306

Acoustic feature extraction from IF

- IF is meaningful for only narrowband signals.
- A narrowband filter-bank is designed to get narrowband components of speech signal.
- IF Pyknogram clearly demonstrates that IF preserves the formant transitions.

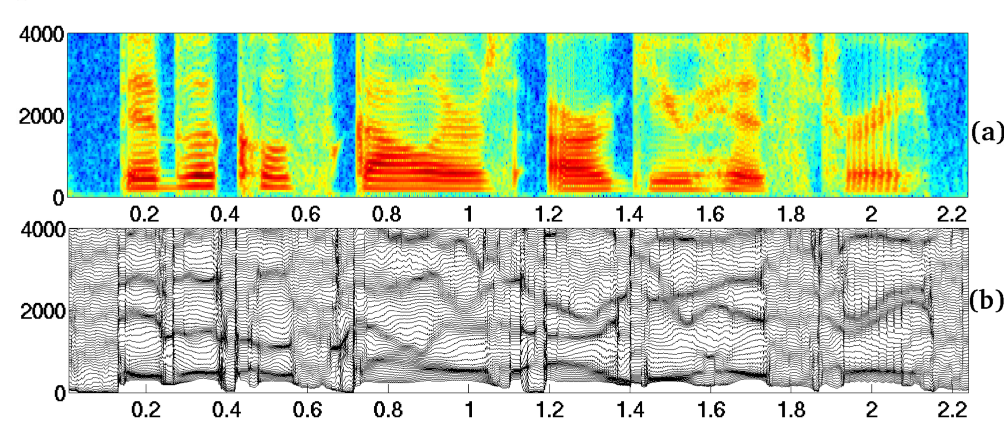

Figure 4: (a) Spectrogram and (b) Pyknogram of IF-Smoothed for a TIMIT sentence, sx42.wav.

- Smoothing is done to suppress the spiky nature of IF-FT.
- Narrowband filtered speech components are used to compute IF for different methods.
- Per-frame averaging is done to obtain IF features.

Development of speech recognizer using IF and magnitude features

- Separate DNN-HMM systems are built using various IF features and MFCCs.
- Performance on TIMIT is evaluated in terms of phone error rate (PER).
- Scores from MFCC and IFCC based posterior lattices are combined using minimum Bayes risk decoding.

Phone error rates on TIMIT

Feature	PER (Dev)	PER (Test)
IFCC-ZC	24.4	26.3
IFCC LMS	23.9	26.4
IFCC-TVAR	21.8	24.0
IFCC - FT	23.9	26.2
IFCC-Smoothed	20.1	21.8
MFCC	17.1	18.4
MFCC+IFCC-Smoothed	15.8	16.8

Conclusions and Future Work

- IF features exhibit significant speech-specific information and provide comparable performance to magnitude based features.
- Smoothed IF features derived from analytic phase of speech signal performs the best among different IF techniques in consideration.
- Score level combination of MFCC and IFCC features deliver state-of-the-art performance for TIMIT phone recognition.
- Exploring IF features for noisy speech recognition as phase becomes significantly important in the presence of noise.
- Exploring IF features for large vocabulary continuous speech recognition.

References

[1] F.-G. Zeng, K. Nie, G. S. Stickney, Y.-Y. Kong, M. Vongphoe, A. Bhargave, C. Wei, and K. Cao, "Speech recognition with amplitude and frequency modulations," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 7, pp. 2293–2298, 2005.

[2] P. Mowlaee, R. Saeidi, and Y. Stylanou, "Interspeech 2014 special session: Phase importance in speech processing applications," in *Proc. Interspeech*, 2014, pp. 1623–1627.

[3] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal. ii. algorithms and applications," *Proceedings of the IEEE*, vol. 80, no. 4, pp. 540–568, 1992.

[4] K. Vijayan, V. Kumar, and K. S. R. Murty, "Feature extraction from analytic phase of speech signals for speaker verification." in *INTERSPEECH*, 2014, pp. 1658–1662.