Reducing F0 Frame Error of F0 Tracking Algorithms Under Noisy Conditions with an Unvoiced/Voiced Classification Frontend

Wei Chu and Abeer Alwan

Speech Processing and Auditory Perception Laboratory
Department of Electrical Engineering
University of California, Los Angeles

Noise Robust F0 Tracking

Motivation

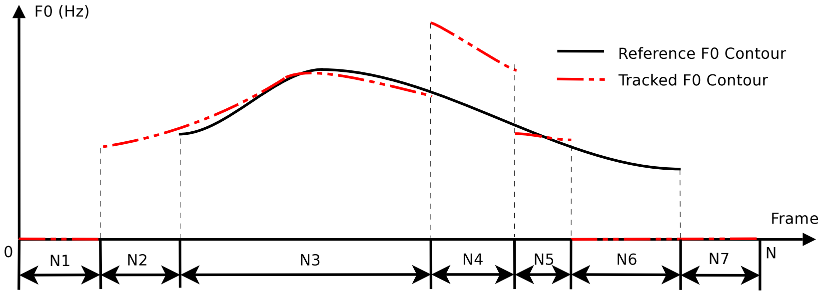
- Develop an error metric that provides a good assessment for F0 tracking algorithms
- Accuratley estimate and track F0 contours under noisy conditions.

Outline

- I. Error Metrics
- II. Statistically-based Unvoiced/Voiced Classifier
- III. Experimental Results and Analysis

I. Error Metrics

An Example of a Tracked and Reference F0 contour



3 possible types of error in any frame *i*

- Unvoiced → Voiced Error;
- Voiced → Unvoiced Error;
- F0 Value Estimation Error.

Current Error Metrics

Two error metrics are currently used:

Voicing Decision Error (VDE)) [NAI08]

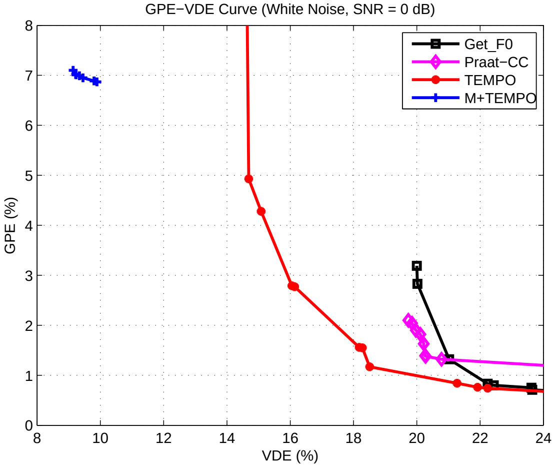
$$VDE = \frac{N_{V \to U} + N_{U \to V}}{N} \times 100\% \quad (1)$$

Gross Pitch Error (GPE) [RCR76]

$$GPE = \frac{N_{F0E}}{N_{VV}} \times 100\% \quad (2)$$

N_{VV}: # of frames which both the F0 tracker and the ground truth consider to be voiced;
N_{F0E}: # of frames for which $\left|\frac{F0_{i,estimated}}{F0_{i,reference}} - 1\right| > 20\%$

GPE-VDE Curve (M+: using U/V classifier output as a mask) in White Noise



A Metric That Combines Two Different Errors

F0 Frame Error (FFE)

$$\begin{aligned} FFE &= \frac{\#\text{ of error frames}}{\#\text{ of total frames}} \times 100\% \\ &= \frac{N_{U\to V} + N_{V\to U} + N_{F0E}}{N} \times 100\%. \end{aligned} \tag{3}$$

FFE is also a combination of GPE and VDE:

$$\begin{aligned} FFE &= \frac{N_{F0E}}{N} \times 100\% + \frac{N_{U\to V} + N_{V\to U}}{N} \times 100\%. \\ &= \frac{N_{VV}}{N} \times GPE + VDE \end{aligned} \tag{4}$$

Therefore, FFE takes both GPE and VDE into consideration.

Why FFE

Look at the Word Error Rate (WER) in ASR:

$$
\begin{aligned}
WER &= \frac{\text{\# of error words}}{\text{\# of total words}} \times 100\% \qquad (5)\\
&= \frac{\text{\# Insertions} + \text{\# Deletions} + \text{\# Substitutions}}{\text{\# All Words}} \times 100\%.
\end{aligned}
$$

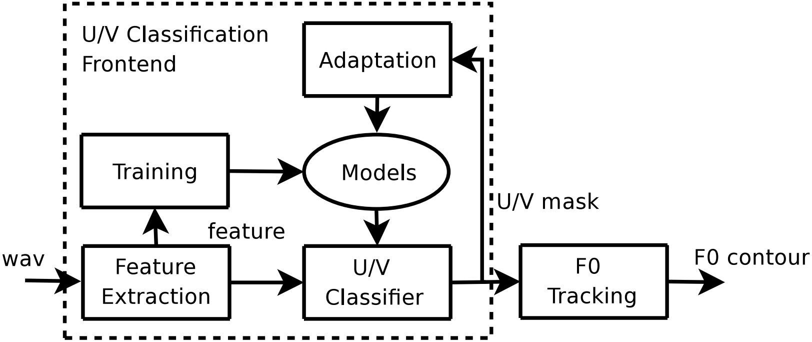
Analogy

- Unvoiced $\rightarrow$ Voiced Error $\Longleftrightarrow$ Insertion Error;
- Voiced $\rightarrow$ Unvoiced Error $\Longleftrightarrow$ Deletion Error;
- F0 Value Estimation Error $\Longleftrightarrow$ Substitution Error.

Thus, FFE in F0 tracking $\Longleftrightarrow$ WER in ASR.

II. Statistically-Based Unvoiced/Voiced Classification Frontend

Figure: 1. The flowchart of our statistically-based U/V classification frontend



Phoneme to Unvoiced/Voiced Dictionary

Table: 1. The mapping from Phonemes to Unvoiced and Voiced

	Stops	Affricates & Fricatives	Nasals & Vowels	Semivowels & Glides	Others
U	p(cl) t(cl) k(cl) bcl dcl gcl q	ch s f th sh	-	hh	epi h pau
V	b d g dx	jh z v zh dh	m n ng em en eng nx iy ih eh ey ae aa aw ay ah ao oy ow uh uw ux er ax ix axr ax-h	l r el w y hv	-

- Phone symbols are used in the TIMIT phone level transcription.
- Two acoustic models were trained: unvoiced(U) and voiced (V).
- The models are left-to-right HMMs

Data Set

For Training the U/V Models: TIMIT corpus

- Only the training data (4 hours) are used.

For Testing the F0 Tracking: KEELE corpus

- A simultaneous recording of speech and laryngograph signals for a phonetically-balanced text.
- The total length: 5 min 37 s, 5 male and 5 female speakers.

White and babble noise are artificially added to training and testing set, SNR = 0 dB

Adaptation to the Speaker Variance

Existing Mismatch

- Only American English corpus (TIMIT) is available for training the U/V models.
- The test set (KEELE) is a British English corpus.

Adaptively learn the distribution of 'Unseen data'!

Maximum Likelihood Linear Regression (MLLR) speaker adaptation [LW95]

A linear transformation $\mathbf{W}_s$ to all the mean vectors of the Gaussians:

$$\mu_s' = \mathbf{W}_s \mu_s \tag{6}$$

III. Experimental Results and Analysis

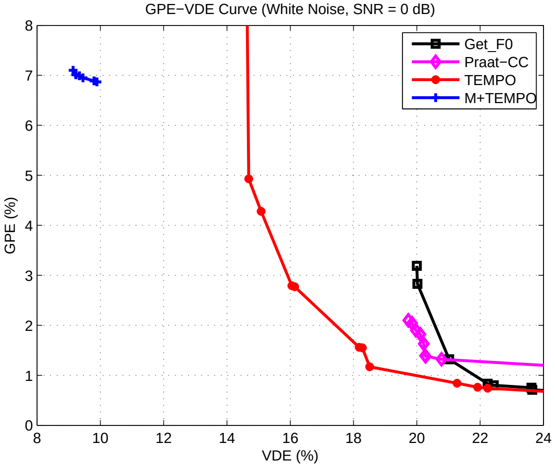
VDE of the U/V Classifier Using the KEELE Corpus

Table: 2. Error rates at SNR = 0 dB, **SI**: speaker independent models, **GSD/RSD**: global style/regression tree style adapted models. (error rates)

VDE	White Noise		Babble Noise	
	MFCC	ETSI	MFCC	ETSI
SI	11.57%	10.84%	30.70%	26.27%
GSD	10.98%	9.81%	27.61%	**22.48%**
RSD	**10.18%**	**9.14%**	**27.23%**	23.54%

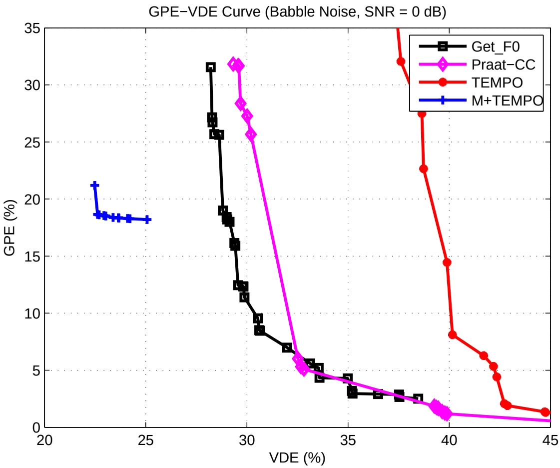
- MFCC: Mel-Frequency Cepstral Coefficients
- ETSI: feature output of the European Telecommunications Standard Institute (ETSI) advanced frontend.
  - before MFCCs extraction: two stage mel-warped Wiener filtering.
  - after MFCCs extraction: blind equalization.

GPE-VDE Curve (M+: using U/V classifier output as a mask) in White Noise



Experiments

GPE-VDE Curve (M+: using U/V classifier output as a mask) in Babble Noise



Analyze the GPE-VDE Curve

For every F0 tracker without the U/V mask, GPE $\searrow$ when VDE $\nearrow$. A possible explanation could be:

- If the VDE $\nearrow$, it may be because the F0 tracker only takes voiced frames with high SNR as voiced.
- Since it is easier to estimate the F0 value over a voiced frame with a higher SNR, the GPE $\searrow$.

Recall: GPE and VDE

$$GPE = \frac{N_{F0E}}{N_{VV}} \times 100\%, \qquad VDE = \frac{N_{V\to U} + N_{U\to V}}{N} \times 100\%$$

GPE, VDE and FFE for the KEELE Corpus Under Default Parameters

Table: 3. Error rates at SNR = 0 dB, **M+**: U/V mask provided by model-based classifier

	White Noise			Babble Noise		
	GPE	VDE	FFE	GPE	VDE	FFE
Get_F0	**0.59**%	35.95%	36.04%	18.89%	30.54%	35.15%
Praat	0.73%	30.77%	30.93%	27.36%	30.99%	38.70%
TEMPO	1.49%	21.92%	22.38%	**8.90%**	47.37%	47.89%
M+TEMPO	6.99%	**9.34%**	**12.64%**	21.19%	**22.48%**	**30.86%**

GPE, VDE and FFE for the KEELE Corpus

Table: 4. SNR = 0 dB, **M+**: U/V mask provided by model-based classifier, **min VDE/FFE**: when VDE/FFE is minimized. (error rates)

		White Noise			Babble Noise		
		GPE	VDE	FFE	GPE	VDE	FFE
Get_F0	min VDE	3.19%	20.00%	21.04%	31.56%	28.21%	37.58%
	min FFE	2.83%	20.02%	20.94%	8.51%	30.65%	32.79%
Praat	min VDE	**2.10%**	19.72%	20.41%	31.82%	29.32%	38.69%
	min FFE	**2.10%**	19.72%	20.41%	**5.31**%	32.67%	33.86%
TEMPO	min VDE	15.87%	14.52%	20.59%	58.05%	36.51%	50.35%
	min FFE	4.93%	14.69%	16.56%	8.11%	40.16%	41.24%
M+TEMPO	min VDE	7.10%	**9.14%**	**12.52%**	18.65%	**22.48%**	**29.86%**
	min FFE	7.10%	**9.14%**	**12.52%**	18.65%	**22.48%**	**29.86%**

Integrating our model-based U/V classifier into an F0-tracking algorithm can improve its FFE and VDE.

Summary

- The F0 Frame Error (FFE) and GPE-VDE curve can be used to evaluate the F0 tracking algorithms in a unified framework.
- The model-based U/V classifier can output robust U/V masks for F0 trackers under both white and babble noise conditions which is helpful for reducing the overall FFE.

Future Work

- Better features for U/V classification to improve VDE.
- Explore noise robust F0 value estimation methods to reduce GPE.

Acknowledgement

The authors would thank Hideki Kawahara for providing the TEMPO package, and Georg Meyer for providing the KEELE corpus.

Thank you!

Q & A?

C. J. Leggetter and P. C. Woodland.
"Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models."
Computer Speech and Language, **9**(2):171–185, 1995.

T. Nakatani, S. Amano, T. Irino, K. Ishizuka, and T. Kondo.
"A method for fundamental frequency estimation and voicing decision: Application to infant utterances recorded in real acoustical environments."
Speech Communication, **50**(3):203–214, 2008.

L. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal.
"A Comparative Performance Study of Several Pitch Detection Algorithms."
IEEE Trans. on Acoustics, Speech, and Signal Processing, **24**(5):399–418, 1976.