Vocal melody extraction in the presence of pitched accompaniment using harmonic matching methods

Vishweshwara Rao, Preeti Rao

Department of Electrical Engineering
Indian Institute of Technology
Bombay

SITUATION

INDIAN CLASSICAL VOCAL PERFORMANCE

Drone (*Tanpura*)

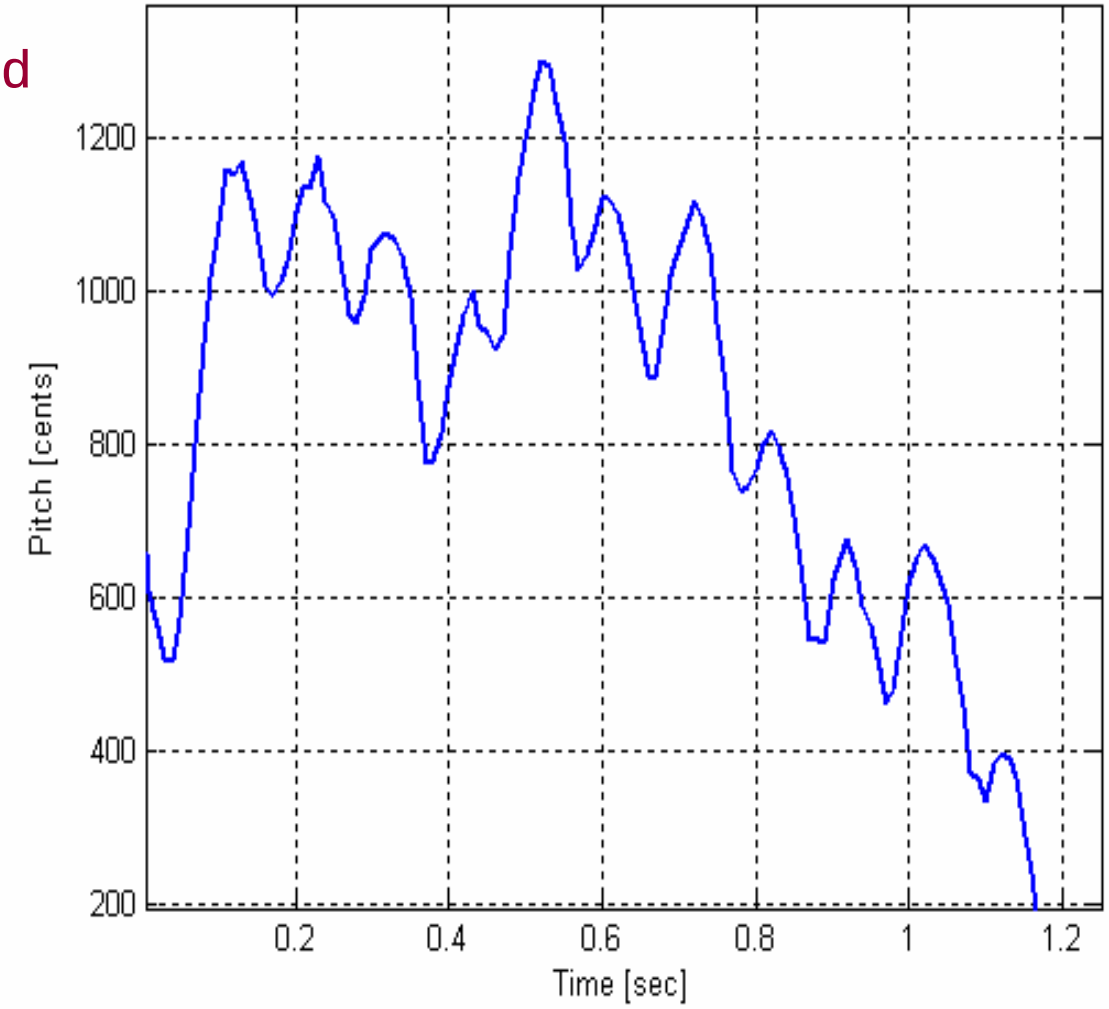
Typical north Indian classical vocal music performance

Percussion (*Tabla*)

Singer

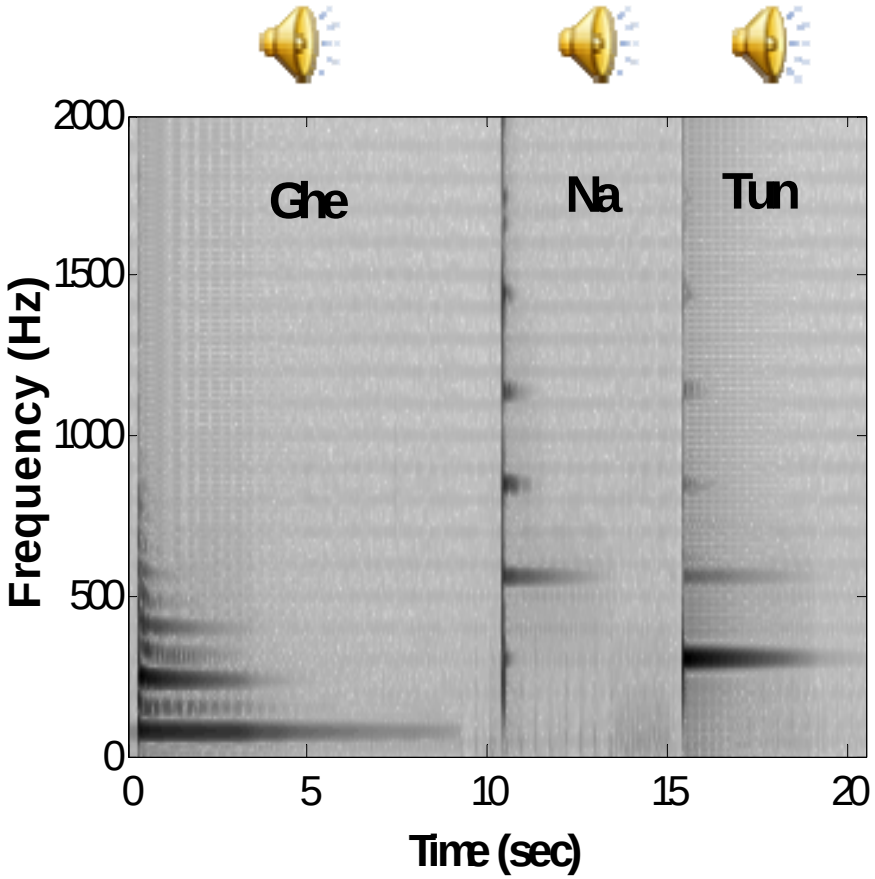
SIGNAL CHARACTERISTICS
VOICE

- Frequent, large, rapid pitch modulations



PERCUSSION (*TABLA*)

Category	**Tonal**	**Impulsive**
Rate of decay	slow	fast
Spectrum	near harmonic	noisy
Examples	Na, Tun, Ge	Tit, Ke

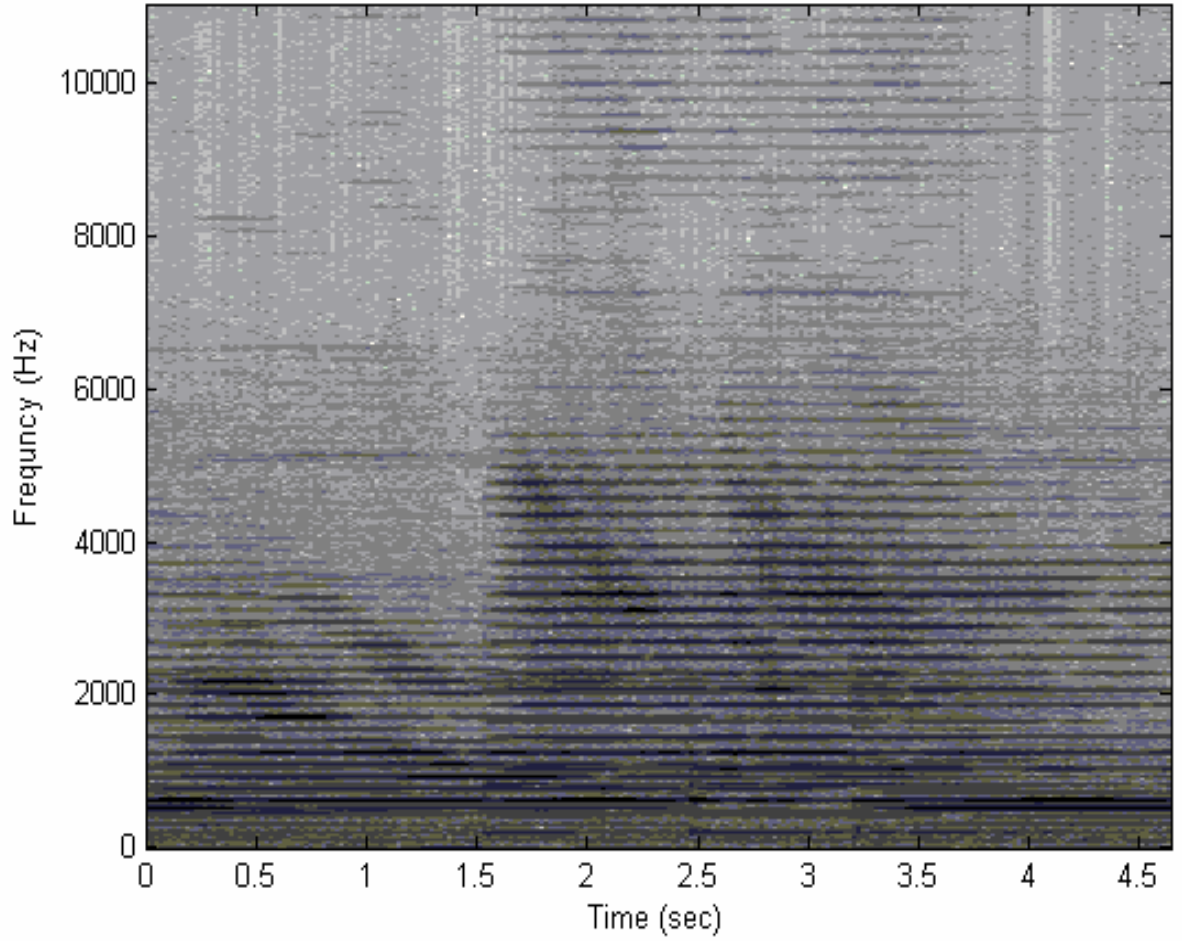


Signal to Interference Ratio (SIR) -10 dB

SIGNAL CHARACTERISTICS

DRONE (*TANPURA*)

- Constant background
- Rich in overtones
- Tuned to the tonic and fourth or fifth
- SIR around 20 – 30 dB



Harmonic Matching PDAs

- Based on explicit frequency domain matching of measured spectrum with an ideal harmonic spectrum
- Input
  - Magnitudes and frequencies of detected sinusoids (window main-lobe matching [Griffin & Lim 1988])
- Differ on the basis of spectral fitness measure or error function

Pattern Recognition (PR) [Brown 1992]

- Based on maximizing the cross correlation between an *n*-pulse template and the measured spectrum, where *n* is the number of included harmonics

- Operation

$$C(\psi) = \sum_{i=1}^{M} I(f_i) X(f_i + \psi)$$

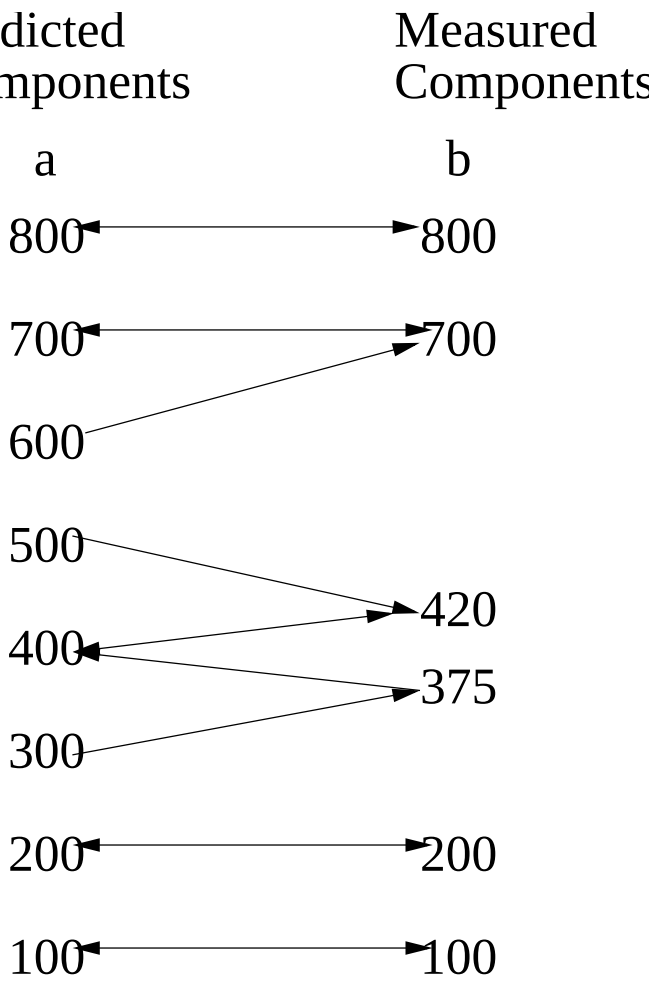
  - Cross correlation (*C*) between ideal spectrum (*I*) and measured spectrum (*X*) for different trial F0 (ψ).
  - Frequency axis is logarithmically spaced
  - Ideal spectrum consists of impulses at expected harmonic locations

Two-Way Mismatch (TWM) [Maher & Beauchamp 1994]

- Minimizes **ERROR** between measured spectral peaks and predicted harmonic spectral pattern for different trial F0
- Based on
  - (1) normalized frequency error
  - (2) normalized amplitude
- Maximum when
  - (1) is large
  - (1) is small and (2) is small
- Minimum when
  - (1) is small and (2) is large

$$Err_{total} = \frac{Err_{p\to m}}{N} + \rho\frac{Err_{m\to p}}{K}$$

$$Err_{p\to m} = \sum_{n=1}^{N} \Delta f_n \cdot (f_n)^{-p} + (\frac{a_n}{A_{\max}}) \times [q\Delta f_n \cdot (f_n)^{-p} - r]$$

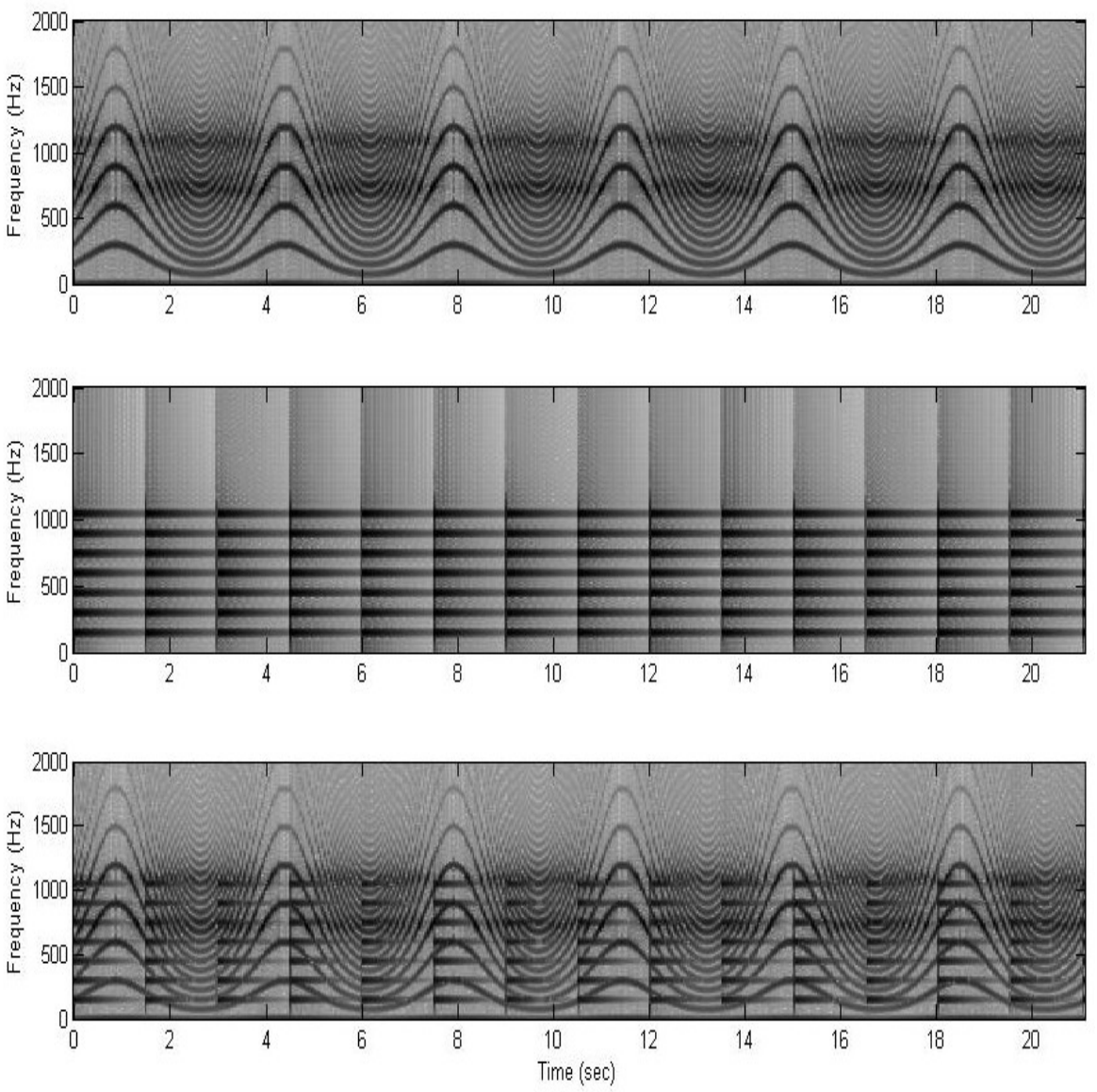


Nearest Neighbour Matching

From [Maher & Beauchamp 1994]

SIMULATION
DATA

- Target (Voice)
  - Formant synthesis
  - F0 smoothly varying
    - Max rate : 3 ST/sec
  - Base F0s
    - 150 Hz
    - 330 Hz
  - Range ±1 octave
- Interference (*Tabla*)
  - Complex tones
  - Same F0 as voice base F0
  - Amplitude envelope decays over 2 sec
- SIR
  - -10 dB



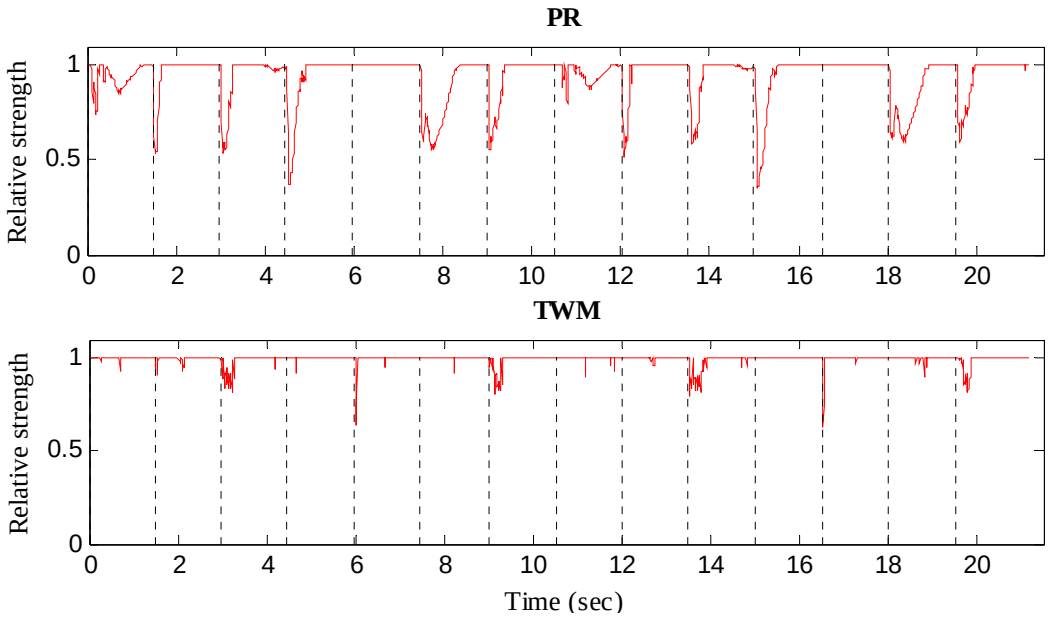
SIMULATION RESULTS

Interference	Base F0 = 150 Hz				Base F0 = 330 Hz			
	PR	PR-DP	TWM	TWM-DP	PR	PR-DP	TWM	TWM-DP
None	100.0	100.0	99.6	100.0	100.0	100.0	98.5	100.0
1 harmonic	70.8	68.1	92.5	100.0	69.9	74.2	92.3	100.0
3 harmonics	64.7	63.1	88.7	94.3	66.3	69.1	90.8	97.4
5 harmonics	62.8	61.4	86.9	93.4	65.1	70.2	86.6	93.7

- Pitch accuracy [Poliner et.al. 2007]
  - Tolerance 50 cents
- Relative Strength

$$RS = 1 - \frac{MC_{tr} - MC_{mf}}{MC_{tr}}$$

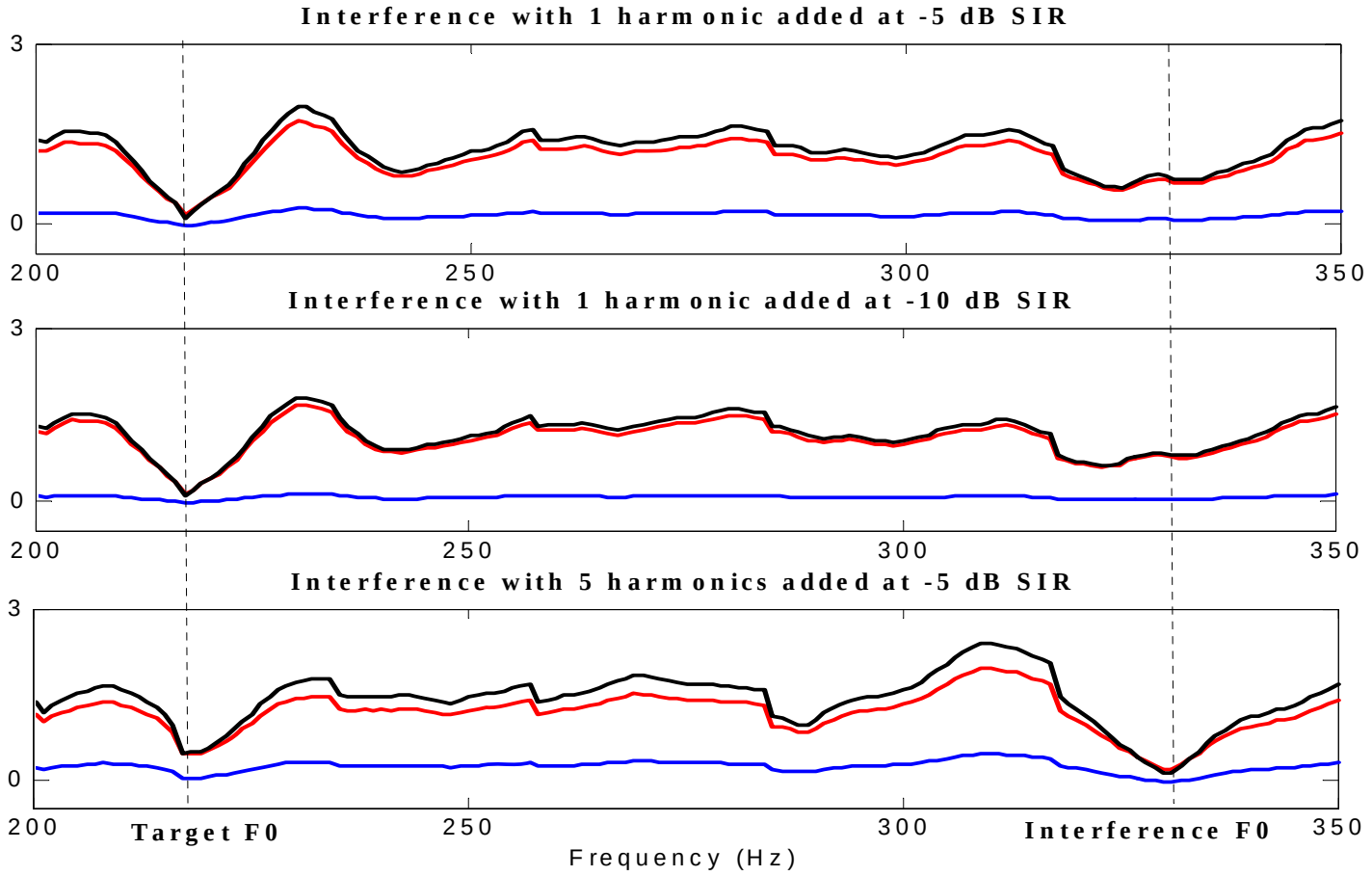
  - MC_{tr} – Measurement cost of Top Ranked candidate
  - MC_{mf} – Measurement cost of Melodic F0 candidate



SIMULATION

ROBUSTNESS OF TWM ERROR FUNCTION

$$\left| Err_{p \to m} = term1 + term2 \right| \quad term1 = \sum_{n=1}^{N} \frac{\Delta f_n}{(f_n)^p} \quad \left| term2 = \sum_{n=1}^{N} \left(\frac{a_n}{A_{\max}} \right) \left(q \frac{\Delta f_n}{(f_n)^p} - r \right) \right|$$



REAL SIGNALS
DATA & RESULTS

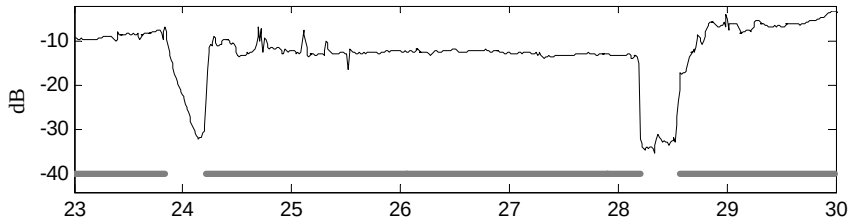
- Multi-track time-synch. data of voice, *tabla*, *tanpura*
  - One min. excerpts (low and high tempo regions) of 2 artists
  - Acoustic isolation by distancing artists
- Ground truth from vocal tracks
  - Majority vote between YIN [deCheveigne & Kawahara 2002], SHS [Hermes 1988] and TWM, with DP
  - Concurrence threshold : 50 cents (~3%)
- Mixtures
  - V – Voice only
  - VT – Voice + *tabla* (5 dB SIR)
  - VTT – Voice + *tabla* (5 dB SIR) + *tanpura* (20 dB SIR)

Content	PR		TWM	
	Raw	DP	Raw	DP
V	90.81	98.24	98.34	99.66
VT	78.01	80.45	97.41	99.51
VTT	76.74	79.71	92.90	98.20

VOCAL DETECTION
METHOD & POST-PROCESSING

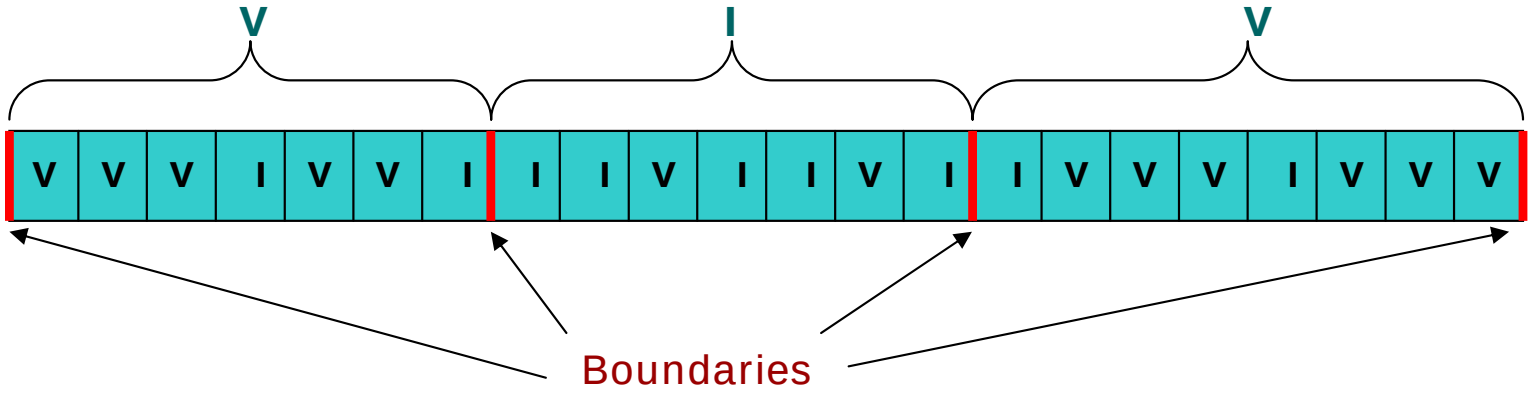
- Frame-level decisions
  - Input feature : Normalized Harmonic Energy (NHE)

$$NHE = \sum_{i=1}^{N} \left| X\left[k_i\right] \right|^2$$



    - $|X()|$ - magnitude spectrum
    - k_i – bin number of local maxima closest to i^{th} expected harmonic for given F0
  - GMM classifier

- Post-processing
  - Grouping of frame-level labels over automatically segments [Foote 2000] by majority vote [Li & Wang07]

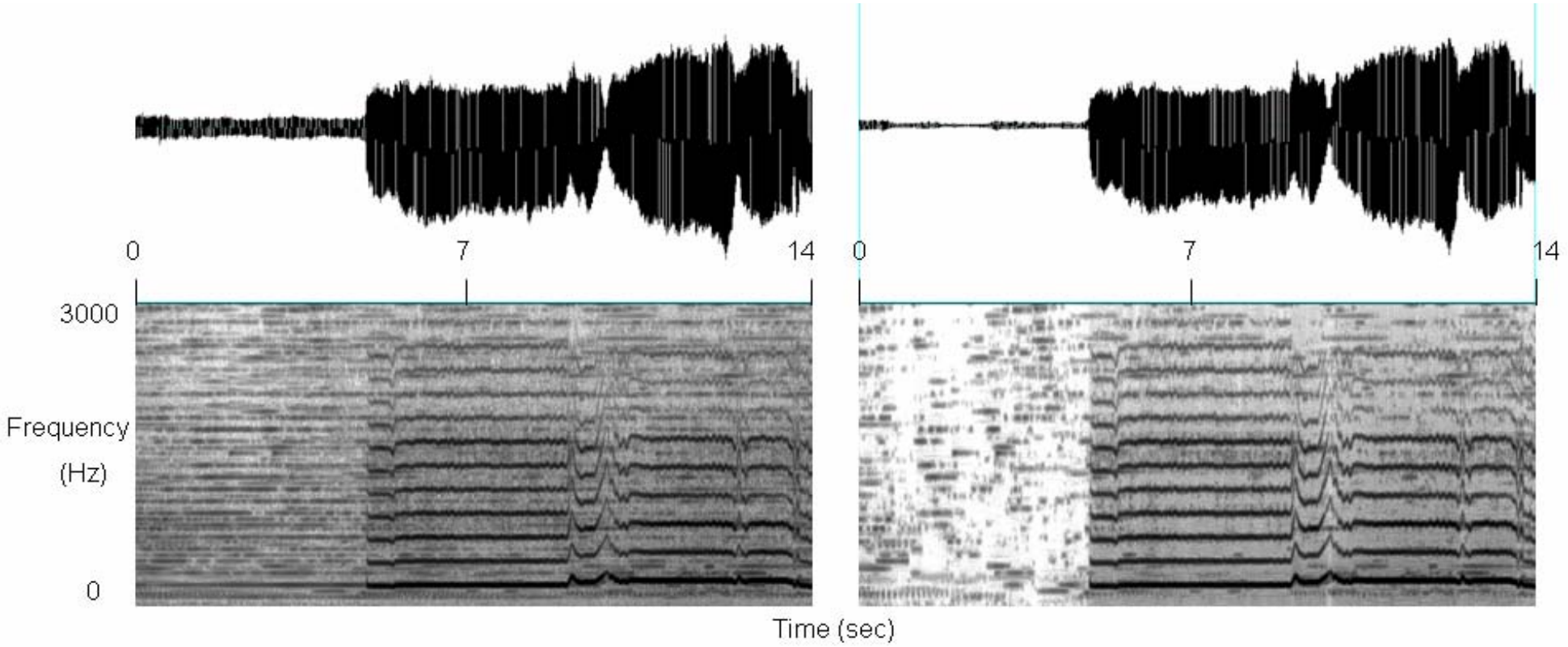


PRE-PROCESSING

Tanpura suppression by spectral subtraction (SS) [Boll 1979]

Before

After

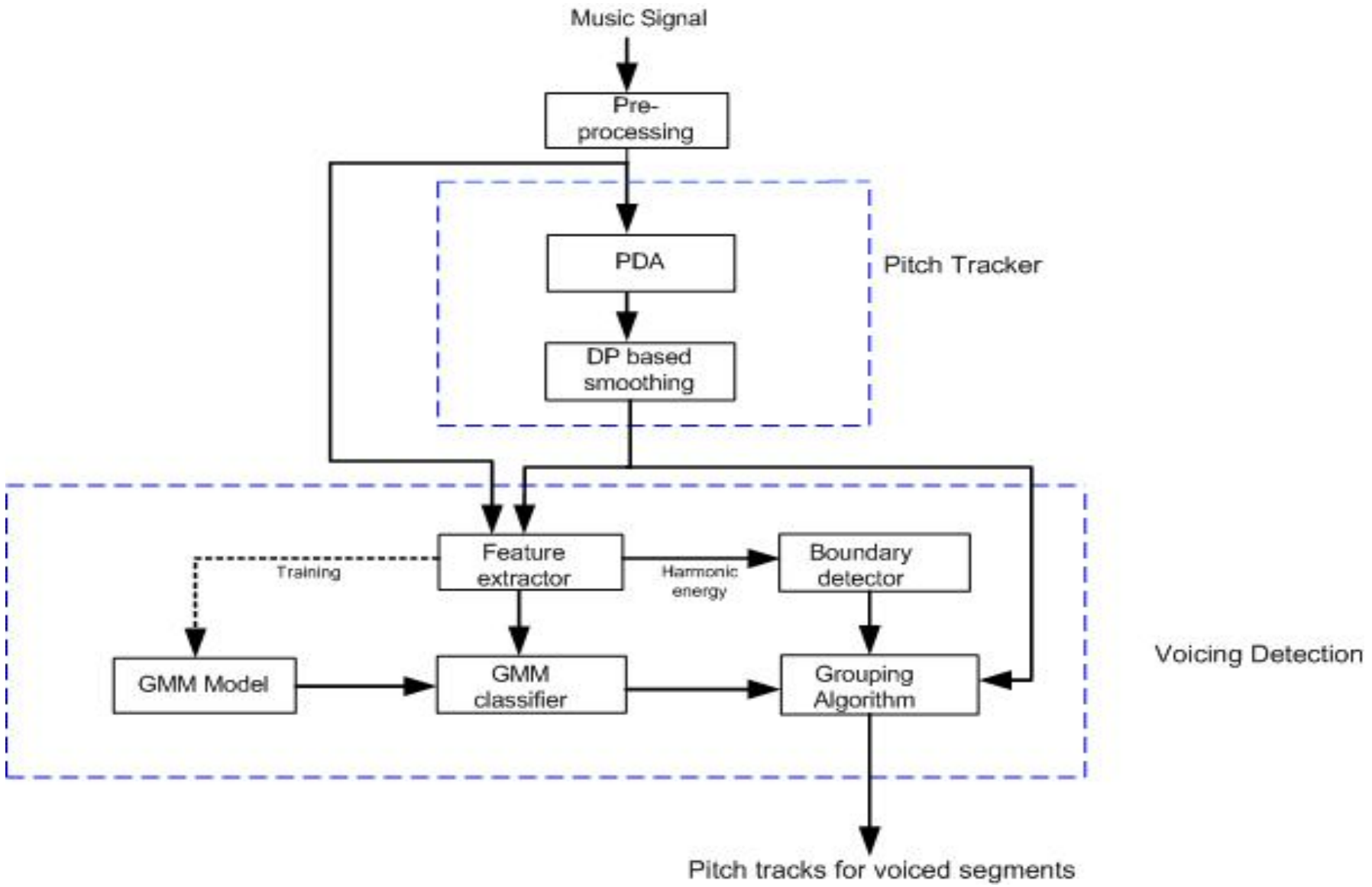


VOCAL DETECTION
RESULTS

- V – Vocal accuracy
  - % of actually vocal frames detected as vocal
- I – Instrumental accuracy
  - % of actually Instrumental frames detected as Instrumental

		Before Grouping			After Grouping		
		V	I	Overall	V	I	Overall
Before SS	V	86.9	13.1	87.8	92.3	7.7	92.4
	I	5.6	94.5		6.2	93.8	
After SS	V	91.6	8.4	91.9	96.6	3.4	96.2
	I	6.1	93.9		6.6	93.4	

FINAL SYSTEM BLOCK DIAGRAM



CONCLUSIONS & FUTURE WORK

- Conclusion
  - TWM is more robust to sparse, harmonic interference even at low SIRs
    - This is attributed to the specific form of its error function
  - NHE serves as a reliable indicator of voicing
- Future work
  - Secondary melodic instrument problem
    - Investigating methods of instrument suppression based on sinusoidal modeling

MELODY EXTRACTION

COMPARISON OF TWM-DP TO MIREX'06 SUBMISSIONS

- RPA – Raw pitch accuracy (Tolerance : 50 cents) [Poliner 2007]
- RCA – Raw chroma accuracy (All pitches folded down to one octave)

Dataset	Algorithm	Vocal		Instrumental		Overall	
		RPA	RCA	RPA	RCA	RPA	RCA
ISMIR 04	Dressler	77.1	78.0	88.7	90.1	82.9	84.0
	Ryynanen	78.3	79.3	82.8	85.3	80.6	82.3
	Poliner	65.4	69.0	81.0	83.9	73.2	76.4
	Sutton	67.5	68.0	57.7	62.9	62.6	65.4
	Brossier	56.3	63.5	58.5	73.8	57.4	68.7
	TWM-DP	83.1	88.9	69.9	80.9	78.0	85.8
MIREX 05	TWM-DP	80.2	82.3	74.5	79.7	78.5	81.5

VOCAL DETECTION

COMPARISON OF NHE TO MIREX'06 SUBMISSIONS

- Recall = % of actually voiced frames labeled as voice
- False alarms = % of actually instrumental frames labeled as voice
- NHE threshold = -15 dB

Algorithm	Vocal only		Non-Vocal only		All data	
	Recall (%)	False Alm (%)	Recall (%)	False Alm (%)	Recall (%)	False Alm (%)
ISMIR 2004 Testing dataset						
dressler	89.8	10.9	92.0	9.5	90.9	10.5
ryynanen	85.9	11.5	82.9	15.2	84.4	12.6
poliner	88.4	34.5	91.4	40.4	89.9	36.3
sutton	90.8	32.0	54.6	8.1	73.2	24.9
brossier	99.8	93.9	99.7	82.9	99.7	88.4
NHE	79.9	20.2	85.9	18.6	82.2	19.7
MIREX 2005 Training dataset						
NHE	82.3	17.4	93.1	41.1	85.5	26.1

REFERENCES

[Bol79] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. on Audio, Speech and Signal Processing*, vol. 27, no. 2, pp. 113-120, 1979.

[Bro92] J. Brown “Musical fundamental frequency tracking using a pattern recognition method,” *J. Acoust. Soc. Am.,* vol. 92, no. 3, pp. 1394-1402, 1992.

[Chev02] A. de Cheveigné and H. Kawahara, “YIN, a fundamental frequency estimator for speech and music,” *J. Acoust. Soc. Am.,* vol. 111, no. 4, pp. 1917-1930, 2002.

[Foot00] J. Foote, “Automatic audio segmentation using a measure of audio novelty,” in *Proc. of IEEE International Conference on Multimedia and Expo (ICME)*, New York City, 2000.

[Herm88] D. J. Hermes, “Measurement of pitch by Sub-Harmonic Summation,” *J. Acoust. Soc. Am.,* vol. 83, no.1, pp. 257-264, 1988.

[Grif88] D. Griffin and J. Lim, “Multiband Excitation Vocoder,” *IEEE Trans. On Acoustics, Speech and Signal processing*, vol. 36, no. 8, pp. 1223-1235, 1988.

[LiWang07] Y. Li and D. Wang, “Separation of singing voice from music accompaniment for monoaural recordings,” *IEEE Trans .on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1475-1487, 2007.

[Mah94] R. Maher and J. Beauchamp, “Fundamental frequency estimation of musical signals using a Two-Way Mismatch procedure,” *J. Acoust. Soc. Am.,* vol. 95, no. 4, pp. 2254-2263, 1992.

[Ney83] H. Ney, “Dynamic Programming Algorithm for Optimal Estimation of Speech Parameter Contours,” *IEEE Trans. on Systems, Man and Cybernetics,* vol. SMC-13, no. 3, pp. 208-214, 1983.

[Pol07] G. Poliner et. al., “Melody transcription from music audio: Approaches and evaluation,” *IEEE Trans. on Audio, Speech and Language Processing,* vol. 15, no. 4, pp. 1247-1256, 2007.