



# 2-D Processing of Speech with Application to Pitch and Formant Estimation\*

Thomas F. Quatieri and Tianyu Tom Wang  
MIT Lincoln Laboratory

Harvard Workshop on  
Next-Generation Statistical Models for  
Speech and Audio Signal Processing

November 9-10 2007

\*This work was supported by the Department of Defense under Air Force contract FA8721 05 C 0002. The opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government..

1

MIT Lincoln Laboratory



## Motivation From Image Processing

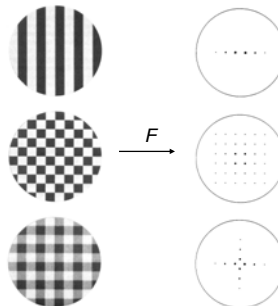
- Certain 2-D geometric patterns transform to dots in a 2-D spatial frequency plane\*

- Time-frequency distributions contain "geometric patterns" due to harmonic content

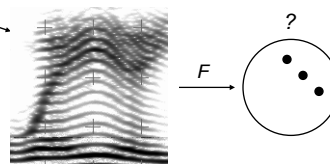
- Possible use
  - Pitch estimation
  - Noise reduction
  - Multi-speaker separation

\*From R.L. DeValois and K.K. DeValois, *Spatial Vision*, Oxford University Press, 1988.

### 2-D Grating Patterns



### Narrowband Spectrogram



2

MIT Lincoln Laboratory

# Report Documentation Page

Form Approved  
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

|  |                                    |                                     |                            |   |                                 |
|--|------------------------------------|-------------------------------------|----------------------------|---|---------------------------------|
| 1. REPORT DATE<br><b>NOV 2007</b>  |                                    | 2. REPORT TYPE                      |                            | 3. DATES COVERED<br><b>00-00-2007 to 00-00-2007</b> |                                 |
| 4. TITLE AND SUBTITLE<br><b>2-D Processing of Speech with Application to Pitch and Formant Estimation</b>  |                                    |                                     |                            | 5a. CONTRACT NUMBER                                 |                                 |
|  |                                    |                                     |                            | 5b. GRANT NUMBER                                    |                                 |
|  |                                    |                                     |                            | 5c. PROGRAM ELEMENT NUMBER                          |                                 |
| 6. AUTHOR(S)   |                                    |                                     |                            | 5d. PROJECT NUMBER                                  |                                 |
|  |                                    |                                     |                            | 5e. TASK NUMBER                                     |                                 |
|  |                                    |                                     |                            | 5f. WORK UNIT NUMBER                                |                                 |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)<br><b>Massachusetts Institute of Technology, Lincoln Laboratory, 244 Wood Street, Lexington, MA, 02420-9108</b> |                                    |                                     |                            | 8. PERFORMING ORGANIZATION REPORT NUMBER            |                                 |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)  |                                    |                                     |                            | 10. SPONSOR/MONITOR'S ACRONYM(S)                    |                                 |
|  |                                    |                                     |                            | 11. SPONSOR/MONITOR'S REPORT NUMBER(S)              |                                 |
| 12. DISTRIBUTION/AVAILABILITY STATEMENT<br><b>Approved for public release; distribution unlimited</b>  |                                    |                                     |                            |   |                                 |
| 13. SUPPLEMENTARY NOTES  |                                    |                                     |                            |   |                                 |
| 14. ABSTRACT   |                                    |                                     |                            |   |                                 |
| 15. SUBJECT TERMS  |                                    |                                     |                            |   |                                 |
| 16. SECURITY CLASSIFICATION OF:  |                                    |                                     | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES                                 | 19a. NAME OF RESPONSIBLE PERSON |
| a. REPORT<br><b>unclassified</b>   | b. ABSTRACT<br><b>unclassified</b> | c. THIS PAGE<br><b>unclassified</b> |                            |   |                                 |



## OUTLINE

- **2-D Spectrogram Model and Mapping**
- Application to Pitch Estimation
- Application to Formant Estimation
- Extension to Alternate Time-Frequency Distributions
- Conclusions and Directions

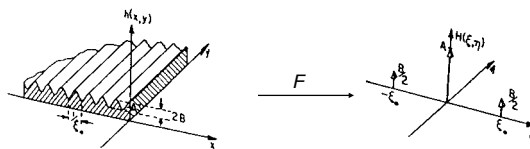
3

MIT Lincoln Laboratory



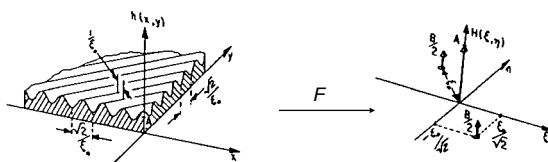
## 2-D Spectrogram Model Inspiration from Image Processing

- **2-D sine on a pedestal\*: Zero degree rotation**



Distance of 2-D impulses from origin varies inversely with sine frequency

- **2-D sine on a pedestal\*: 45 degree rotation**



Angle of 2-D impulses w/r axes proportional to extent of sine rotation

From J.D. Gaskill, *Linear Systems, Fourier Transforms, and Optics*, John Wiley and Sons, New York, NY, 1978.

4

MIT Lincoln Laboratory



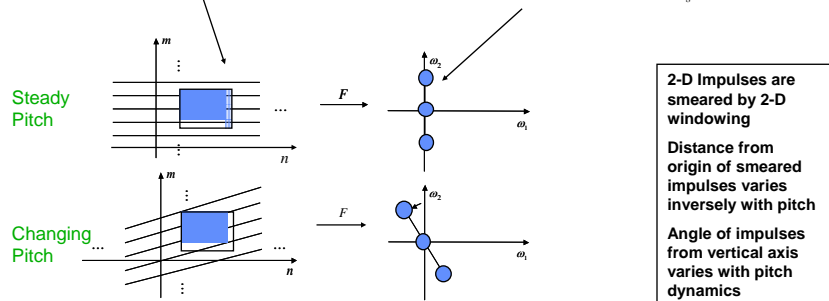
## 2-D Spectrogram Model

### Short-Space 2-D Sine

- Harmonic line structure of the narrowband spectrogram is modeled over a small region by a 2-D sine function sitting on a flat pedestal of unity
- 2-D window is applied to extract a short-time segment and 2-D Fourier transform is then computed

$$x[n, m] = w[n, m](1 + \cos(\omega_g m))$$

$$X(\omega_1, \omega_2) = 2W(\omega_1, \omega_2) + W(\omega_1, \omega_2 - \omega_g) + W(\omega_1, \omega_2 + \omega_g)$$



5

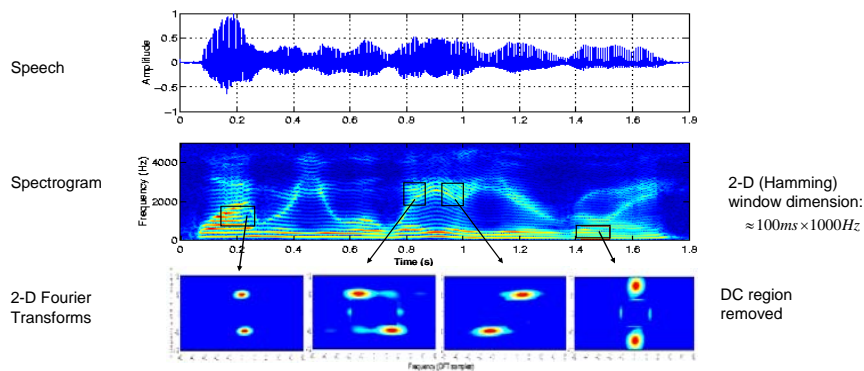
MIT Lincoln Laboratory



## 2-D Processing

### Example

- 2-D analysis of narrowband spectrogram of all-voiced female speech



Henceforth, refer to 2-D mapping as the “Grating Compression Transform” (GCT) to highlight mapping “gratings” to concentrated “dots”

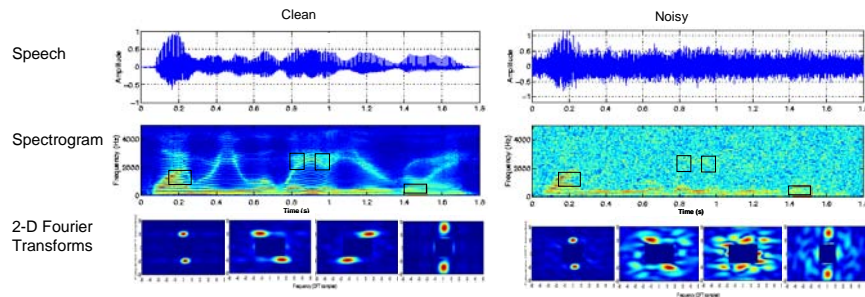
6

MIT Lincoln Laboratory



## 2-D Processing Example with Noise

- 2-D analysis of all-voiced female speech in noise
  - GCT without and with additive white Gaussian noise at average SNR of  $\sim 3$  dB



- Energy concentration of GCT is typically preserved at roughly the same location as for the clean case
  - However, when noise dominates so that little harmonic structure remains within the 2-D window, energy concentration deteriorates, as in the vicinity of 0.95 s and 2000 Hz

MIT Lincoln Laboratory

7



## OUTLINE

- 2-D Spectrogram Model and Mapping
- Application to Pitch Estimation
- Application to Formant Estimation
- Extension to Alternate Time-Frequency Distributions
- Conclusions and Directions

MIT Lincoln Laboratory

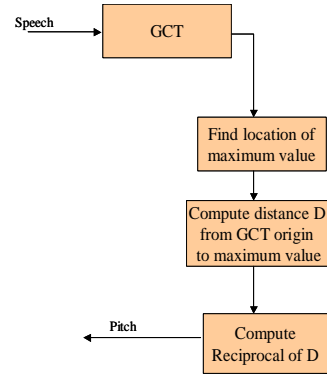
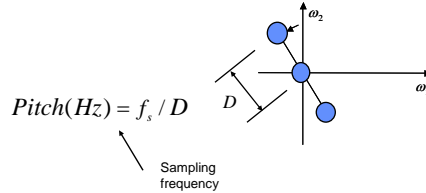
8



## Pitch Estimation GCT-Based Approach

- GCT of speech examples motivate a simple pitch estimator

- Pitch estimate is reciprocal to the distance from the origin to the maximum value in the GCT



- Pitch rate of change is proportional to angle of GCT peak from vertical axis

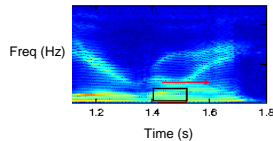
9

MIT Lincoln Laboratory



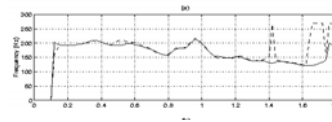
## GCT-Based Pitch Estimation Example

- GCT-based estimator over time
  - 2-D analysis window slid along the speech spectrogram at a 10-ms frame interval at low-frequency location

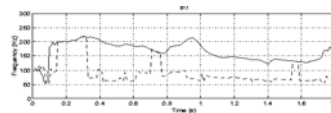


- Relatively robust in noise, outperforming a sinewave-based pitch estimator

GCT-Based Estimator



Sinewave-Based Estimator



Solid: Clean Speech  
Dashed: 3 dB SNR

10

MIT Lincoln Laboratory



## GCT-Based Pitch Estimation Performance

- GCT-based estimator
  - 2-D analysis window slid along the speech spectrogram at a 10-ms frame interval at a low-frequency location given by the 2-D window in previous slide
  - Average magnitude difference measured between pitch-contour estimates with and without white Gaussian noise for both the GCT- and sinewave-based estimators

### Performance Measurements

|      | FEMALES |      | MALES |      |
|------|---------|------|-------|------|
|      | 9dB     | 3dB  | 9dB   | 3dB  |
| GCT  | 0.5     | 6.7  | 0.9   | 6.7  |
| SINE | 5.8     | 40.5 | 2.6   | 12.8 |

Average magnitude error (in dB) in GCT- and sine-wave-based pitch contour estimates for clean and noisy all-voiced passages. The two passages "Why were you away a year Roy?" and "Nanny may know my meaning." from two male and two female speakers were used under noise conditions 9 dB and 3 dB average signal-to-noise ratio.

11

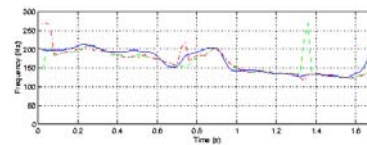
MIT Lincoln Laboratory



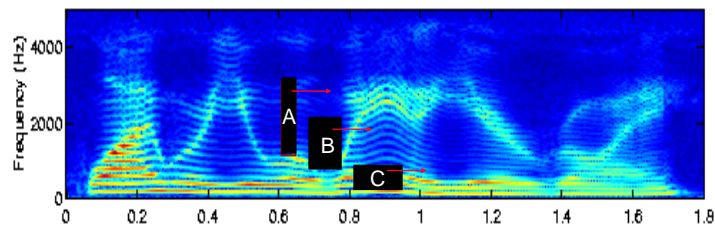
## Pitch Estimation Multi-resolution properties

- Pitch in the 2-D plane
  - Pitch can be obtained anywhere in the 2-D plane
  - "Wavelet-like tiling" of 2-D window found to give the most consistent estimate  
Reflects increase pitch FM with increasing frequency

### Three pitch contours



Window A: Dashed Green  
Window B: Dashed-Dot Red  
Window C: Solid Blue



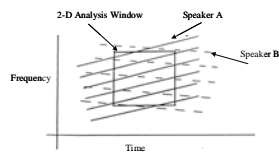
12

MIT Lincoln Laboratory

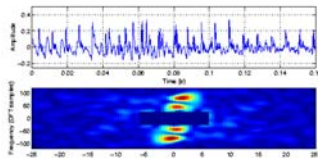


## Two-Speaker Pitch Estimation

- Sum of two speakers has spectrogram with two harmonic sets



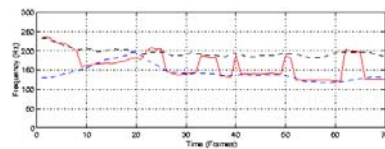
- GCT gives two pairs of dots, one pair for each speaker
  - All-voiced example (male + female)



- Blind use of one-speaker pitch estimator on two-speaker signal

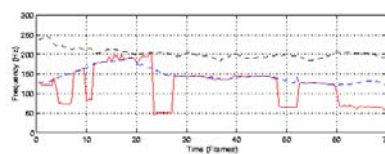


### GCT-Based Estimator



Maximum of GCT latches on to speaker with (locally) largest energy

### Sine-Wave-Based Estimator



13

MIT Lincoln Laboratory



## OUTLINE

- 2-D Spectrogram Model and Mapping
- Application to Pitch Estimation
- **Application to Formant Estimation**
- Extension to Alternate Time-Frequency Distributions
- Conclusions and Directions

14

MIT Lincoln Laboratory

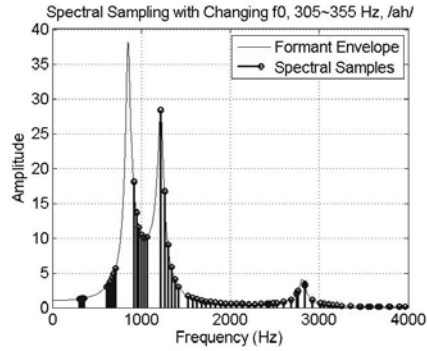
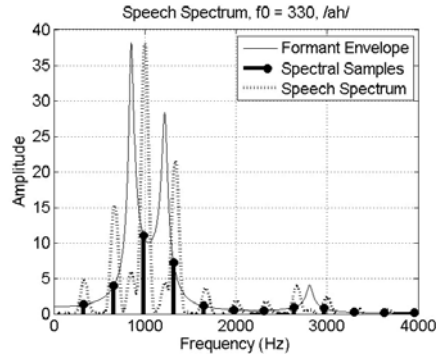




## Formant Estimation The High-Pitched Problem

Synthesized vowel /ah/ with 330-Hz pitch. Speech spectrum generated from short-time Fourier analysis with a 20-ms Hamming window.

Collection of harmonic samples from pitch sweep ranging from 305–355 Hz. Contrast to  $f_0 = 330$  Hz shown in Figure 1.



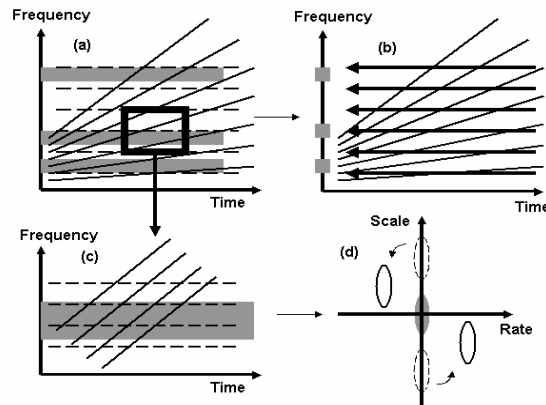
15

MIT Lincoln Laboratory



## 2-D Framework Exploiting Changing Pitch in Formant Estimation

a) Schematic of changing and fixed  $f_0$  across a steady vowel in a STFT; (b) Averaging of harmonic samples to a single 1-D frequency axis; (c) Localized spectrotemporal region from (a); (d) Mapping of source-filter speech components in the GCT from (c).



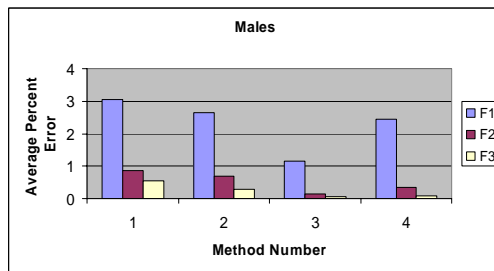
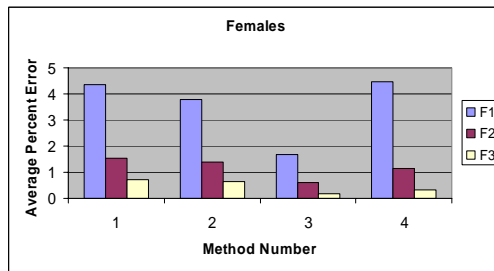
16

MIT Lincoln Laboratory



## Spectral Estimation: Results

### Average Percent Formant Error



- **Methods**
  1. Single STFT slice
  2. Cepstral liftering
  3. Proposed method
  4. Spectral slice averaging
- **Relative gains (method 1) for [F1, F2, F3] via proposed:**
  - Females: [61%, 61%, 73%]
  - Males: [62%, 82%, 87%]
- **Gains for F3 greatest (wider harmonic sampling)**
- **Individual formant scores across vowels also consistent**

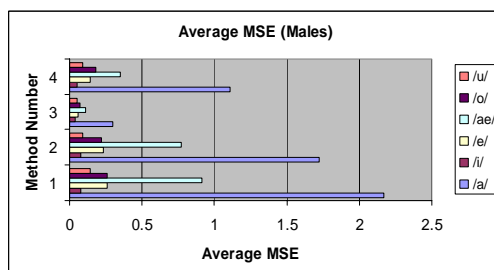
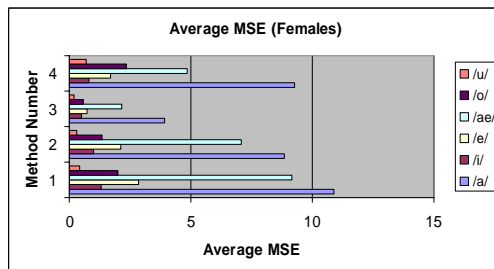
17

MIT Lincoln Laboratory



## Spectral Estimation: Results

### Average MSE



- **Methods**
  1. Single STFT slice
  2. Cepstral liftering
  3. Proposed method
  4. Spectral slice averaging
- **Results are consistent with formant frequency estimation**
  - Method 3 outperforms others for all vowels
- **Data not shown:**
  - Consistent results with children's formants

18

MIT Lincoln Laboratory



## Speaker Recognition: Methods

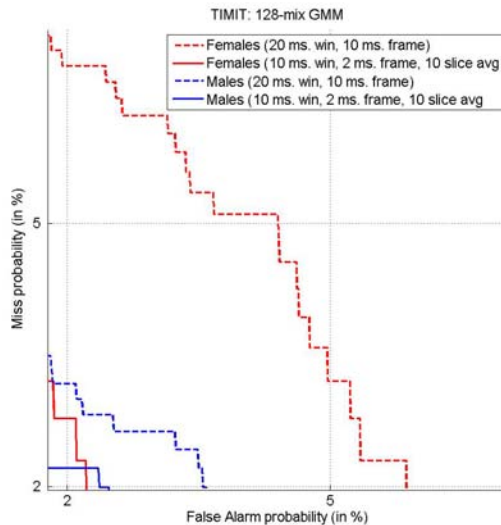
- Data set - male and female subsets of TIMIT corpus
- Baseline system
  - Mel-cepstrum feature extraction with 20 ms. window and 10 ms frame interval + delta features
  - Adaptive Gaussian Mixture Modeling
    - 128 mixture components
    - Universal background model
- System modifications in *feature extraction*
  - Short-time analysis using 10 ms frame window and 2 ms frame interval
  - Compute *average* of spectral slices spanning ~30 ms
  - Derived spectra are used for computing standard mel-cepstrum + deltas

19

MIT Lincoln Laboratory



## Speaker Recognition: Results



- Equal error rate (EER)

|         | Baseline (EER)        | Proposed (EER)        |
|---------|-----------------------|-----------------------|
| Males   | 1.86% < 2.45% < 3.39% | 1.53% < 2.15% < 2.80% |
| Females | 3.12% < 4.41% < 5.64% | 1.55% < 2.15% < 3.30% |

- Absolute EER reduction in females 2.26% (not yet significant)

20

MIT Lincoln Laboratory



## OUTLINE

- 2-D Spectrogram Model and Mapping
- Application to Pitch Estimation
- Application to Formant Estimation
- **Extension to Alternate Time-Frequency Distributions**
- Conclusions and Directions

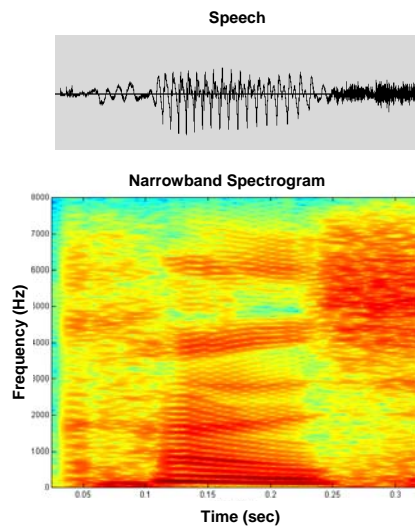
21

MIT Lincoln Laboratory



## Limitation of the Spectrogram Observations

- **Two curious effects are seen:**
  - Frequency tracks moving in the wrong direction, e.g., up rather than down and
  - Crossing tracks, i.e., tracks moving up and down simultaneously.
- **The problem is that the basis functions of the Fourier transform, stationary sinusoids, cannot resolve the speech harmonics which have rapid frequency modulation and are closely spaced in frequency.**
  - This lack of resolution leads to the complex line phenomena seen in Figure 2.



22

MIT Lincoln Laboratory

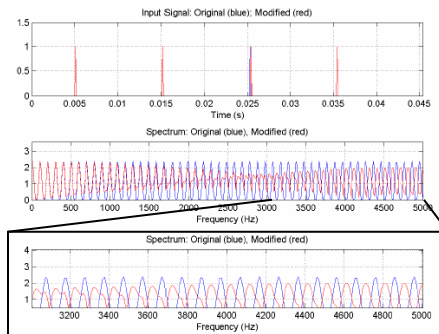


## Spectral Sensitivity\*

### Example

- Harmonic speech spectra can be quite sensitive to aberrations in periodicity of the glottal source
- Even small perturbations can lead to short-time spectral changes that mislead the viewer in terms of signal composition

Example: One-sample shift (0.1 ms)



\*We have developed formal spectral for these sorts of effects: To be published in January 2008 IEEE TSLP, "Spectral representations of nonmodal phonation," Malyska and Quatieri.

We see that the shift in time domain seems to move the harmonics in the higher frequencies

23

MIT Lincoln Laboratory



## An Alternate Transform

### The Fan-Chirp Transform

- The Fan-Chirp Transform (FChT)
  - “Adaptive Chirp-Based Time–Frequency Analysis of Speech Signals”  
Marian Kepesia and Luis Weruaga, *Speech Communication*, vol. 48, no. 5, pp. 474-492, May 2006.
  - “The Fan-Chirp Transform for Non-Stationary Harmonic Signals”  
Luis Weruaga and Marian Kepesia, (submitted to Elsevier)
- FChT is a generalization of the Fourier transform
  - Fourier transform basis functions are stationary sine waves
  - FChT basis functions are sine waves with linear frequency modulation
    - the set of basis functions has a fan geometry
    - 1<sup>st</sup> order match to harmonic frequency modulation in speech

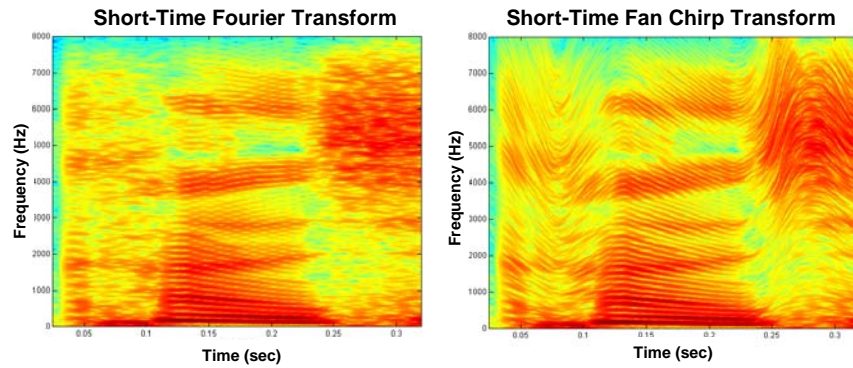
24

MIT Lincoln Laboratory



## Spectrogram Comparison STFT versus Short-Time Fan-Chirp

- Observations
  - FChT resolves high frequency harmonics even when frequency modulation is large
  - Frequency tracks appear as predicted for FChT



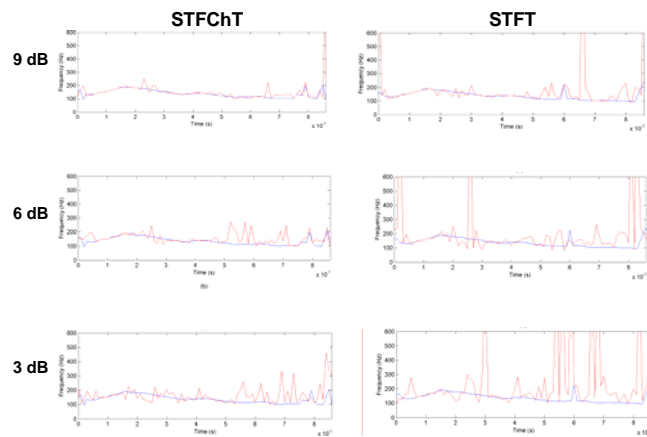
25

MIT Lincoln Laboratory



## Fan-Chirp for Grating Compression Transform Pitch Estimation

- Based on preliminary results, for pitch estimation in noise, the short-time Fan-chirp transform appears to outperform the STFT



26

MIT Lincoln Laboratory



## OUTLINE

- 2-D Spectrogram Model and Mapping
- Application to Pitch Estimation
- Application to Formant Estimation
- Extension to Alternate Time-Frequency Distributions
- **Conclusions and Directions**

27

MIT Lincoln Laboratory



## Conclusions and Directions

- The grating compression transform (GCT) maps harmonically-related signal components to a concentrated entity in a spatial 2-D frequency plane
- The GCT forms the basis of a pitch estimator that uses the radial distance to the largest peak of the GCT
  - The resulting pitch estimator appears robust under noise conditions and amenable to extension to two-speaker pitch estimation
- The GCT forms the basis of a formant estimator that exploits separability of speech source and vocal tract information via changing pitch
- Although the spectrogram provides a useful starting point for the GCT, alternate transforms can provide improved performance
  - Fan-chirp transform is one possibility
- Possible GCT directions
  - Alternate time-frequency distributions
  - Pitch estimation
    - Extended evaluation to a larger corpus and use of voiced/unvoiced speech
    - Two-speaker pitch estimation
  - Formant estimation in noise
  - GCT as model of auditory cortical processing (Sthamma, Ezzat, and Poggio)

28

MIT Lincoln Laboratory