Introduction of Audio Features

Wei-Ta Chu

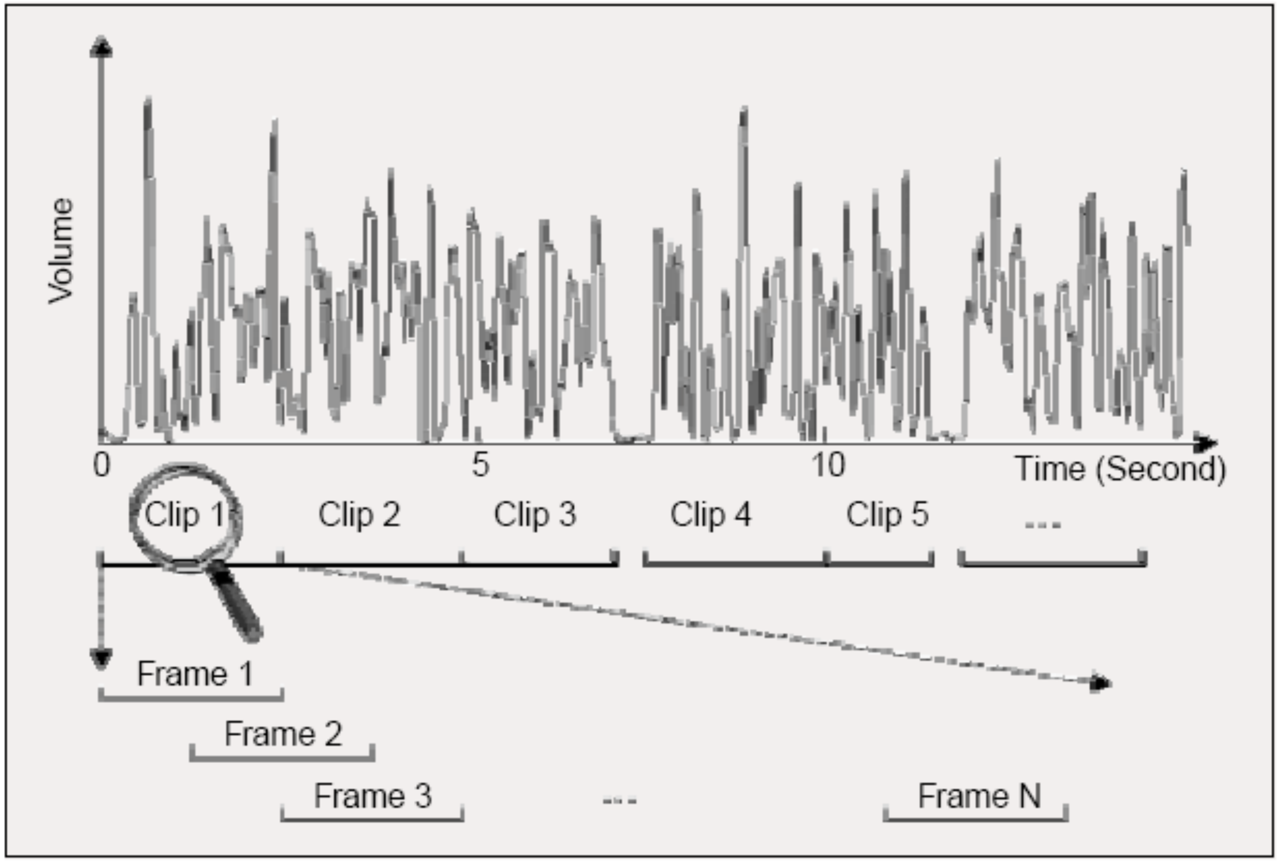
Introduction of Audio Features

- Short-term ***frame*** level vs. long-term ***clip*** level
  - A frame is defined as a group of neighboring samples which last about 10 to 40 ms
    - For audio clips with sampling frequency 16kHz, how many samples are in a 20ms audio frame?
  - Within an audio frame we can assume that the audio signal is stationary.
- A clip consists of a sequence of frames, and clip-level features usually characterize how frame-level features change over a clip.

Y. Wang, Z. Liu, and J.-C. Huang, "Multimedia content analysis – using both audio and visual clues," IEEE Signal Processing Magazine, Nov., 2000, pp. 12-36.

Frames and Clips

- □ Fixed length clips (1 to 2 seconds) or vary-length clips
- □ Both frames and clips may overlap with their previous ones



Frame-Level Features

- □ Most of the frame-level features are inherited from speech signal processing.
- □ Time-domain features
- □ Frequency-domain features
- □ We use N to denote the frame length, and $s_n(i)$ to denote the ith sample in the nth audio frame.

Volume (Loudness, Energy)

- □ Volume is a reliable indicator for silence detection, which may help to segment an audio sequence and to determine clip boundaries.
- □ It is approximated by the root mean square of the signal magnitude within each frame

$$v(n) = \sqrt{\frac{1}{N} \sum_{i=0}^{N-1} s_n^2(i)}$$

- □ Volume of an audio signal depends on the gain value of the recording and digitizing devices. We may normalize the volume for a frame by the maximum volume of some previous frames.

Zero Crossing Rate

- □ Count the number of times that the audio waveform crosses the zero axis.

$$Z(n) = \frac{1}{2}\left(\sum_{i=1}^{N-1} |sign(s_n(i)) - sign(s_n(i-1))|\right)\frac{f_s}{N}$$

ZCR = the number of zero crossings per second

- □ ZCR is one of the most indicative and robust measures to discern unvoiced speech. Typically, unvoiced speech has a low volume but a high ZCR.
- □ Using ZCR and volume together, one can prevent low energy unvoiced speech frames from being classified as silent.

Pitch

- Pitch is the fundamental frequency (基頻) of an audio waveform.
- Normally only voiced speech and harmonic (泛音) music have well-defined pitch.
- Temporal estimation methods rely on computation of the short time autocorrelation function $R_n(l)$ or AMDF $A_n(l)$

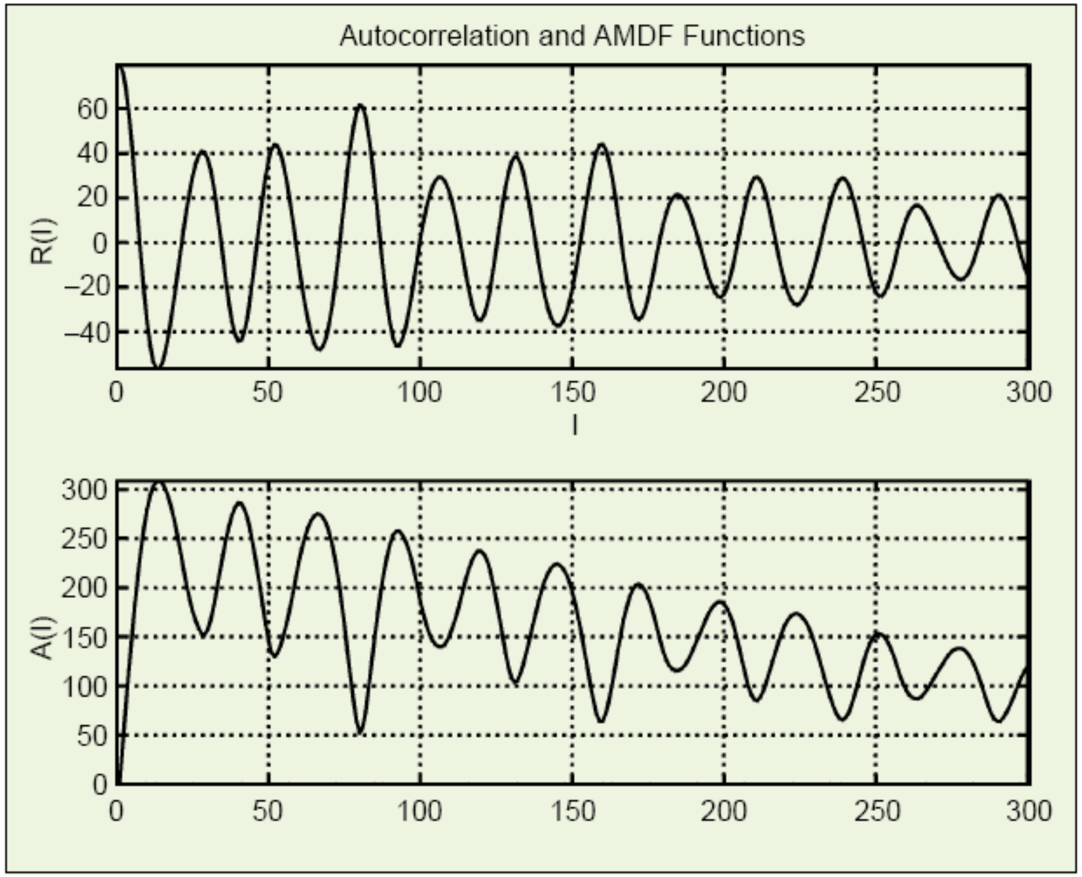
$$R_n(l) = \sum_{i=0}^{N-l-1} s_n(i) s_n(i+l)$$

$$A_n(l) = \sum_{i=0}^{N-l-1} |s_n(i) - s_n(i+l)|$$

AMDF: average magnitude difference function

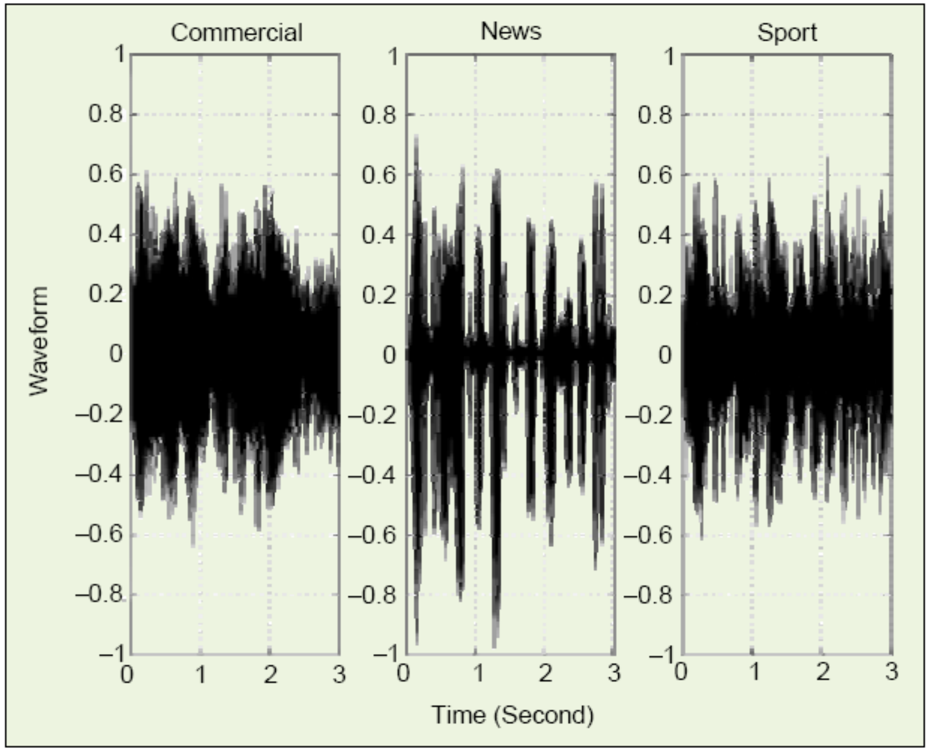
Pitch

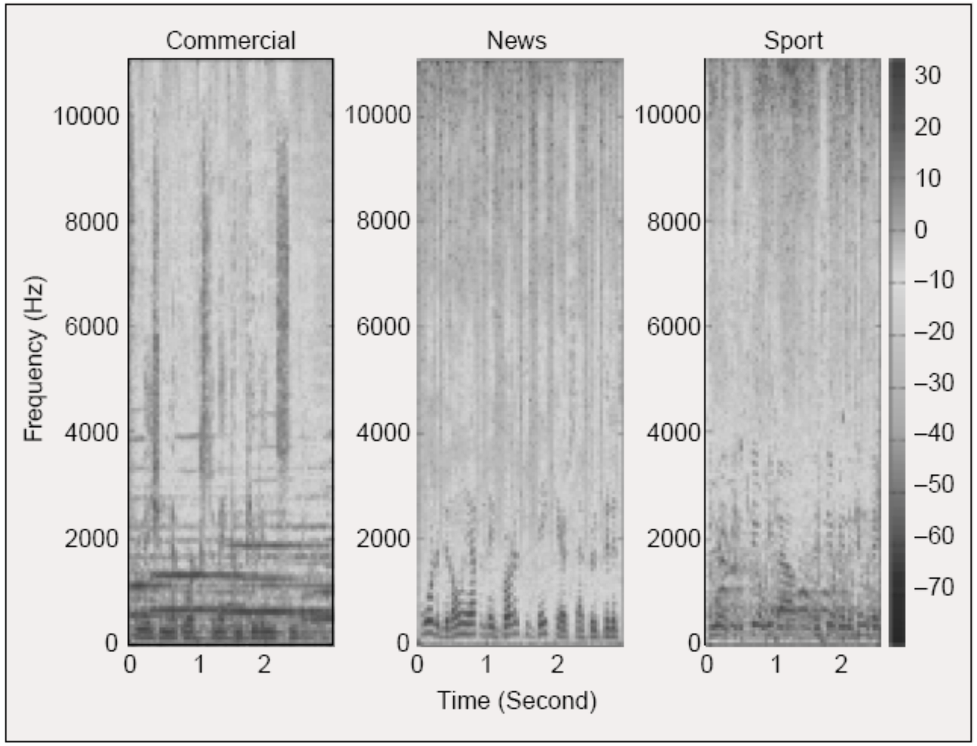
- □ Valleys exist in voiced and music frames and vanish in noise and unvoiced frames.



Spectral Features

- □ Spectrum: the Fourier transform of the samples in this frames
- □ The difference among these three clips is more noticeable in the frequency domain than in the waveform domain





Spectrogram

Spectral Features

- □ Let $S_n(\omega)$ denote the power spectrum (i.e. magnitude square of the spectrum) of frame *n*.
- □ If we think of ω as a random variable and $S_n(\omega)$ normalized by the total power as the probability density function of ω, we can define mean and standard deviation of ω.

$$FC(n) = \frac{\int_0^\infty \omega S_n(\omega)d\omega}{\int_0^\infty S_n(\omega)d\omega}$$

Frequency centroid, brightness

$$BW^2(n) = \frac{\int_0^\infty (\omega - FC(n))^2 S_n(\omega)d\omega}{\int_0^\infty S_n(\omega)d\omega}$$

Bandwidth

Subband Energy Ratio

- The ratio of the energy in a frequency subband to the total energy

$$BE_i = \int_{\omega_L}^{\omega_U} S(\omega)^2 d\omega$$

$$BER_i = \frac{BE_i}{\sum_i BE_i} \quad 1 \leq i \leq 4$$

- When the sampling rate is 22050 Hz, the frequency ranges for the four subbands are 0-630 Hz, 630-1720 Hz, 1720-4400 Hz, and 4400-11025 Hz.

Z. Liu, Y. Wang, and T. Chen, "Audio feature extraction and analysis for scene segmentation and classification," Journal of VLSI Signal Processing, vol. 20, 1998, pp. 61-79.

Spectral Flux

- Spectrum flux (SF) is defined as the average variation value of spectrum between the adjacent two frames.

$$SF = \frac{1}{(N-1)(K-1)} \sum_{n=1}^{N-1} \sum_{k=1}^{K-1} [\log(A(n,k)+\delta) - \log(A(n-1,k)+\delta)]^2$$

$A(n,k)$ is the discrete Fourier transform of the nth frame of input signal

- The SF values of speech are higher than those of music.
- The environment sound is among the highest and changes more dramatically than the other two types of signal.

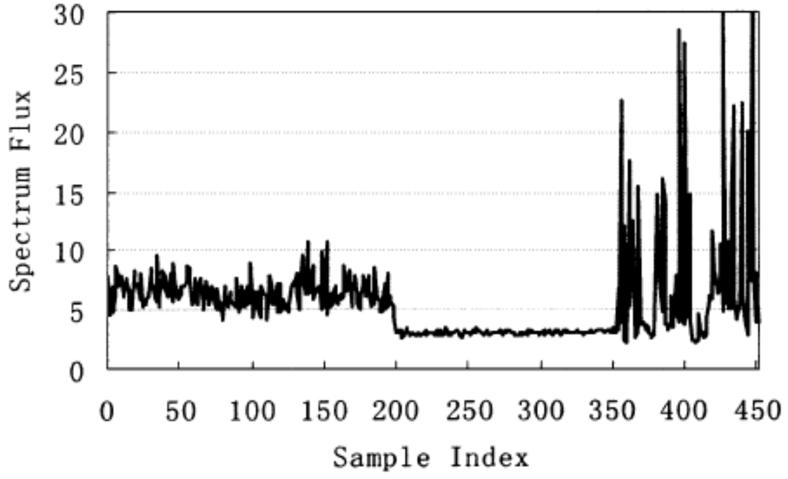


Fig. 4. Spectrum flux curve (0–200 s is speech, 201–350 s is music, and 351–450 s is environment sound).

L. Lu, H.-J. Zhang, H. Jiang, "Content analysis for audio classification and segmentation," IEEE Trans. on Speech and Audio Processing, vol. 10, no. 7, 2002, pp. 504-516.

Spectral Rolloff

- The 95th percentile of the power spectral distribution.
- This measure distinguishes voiced from unvoiced speech. The value is higher for right-skewed distributions.
  - Unvoiced speech has a high proportion of energy contained in the high-frequency range of the spectrum
  - This is a measure of the “skewness” of the spectral shape

E. Scheirer and M. Slaney, “Construction and evaluation of a robust multifeatures speech/music discriminator,” Proc. of ICASSP, vol. 2, 1997, pp. 3741-3744.

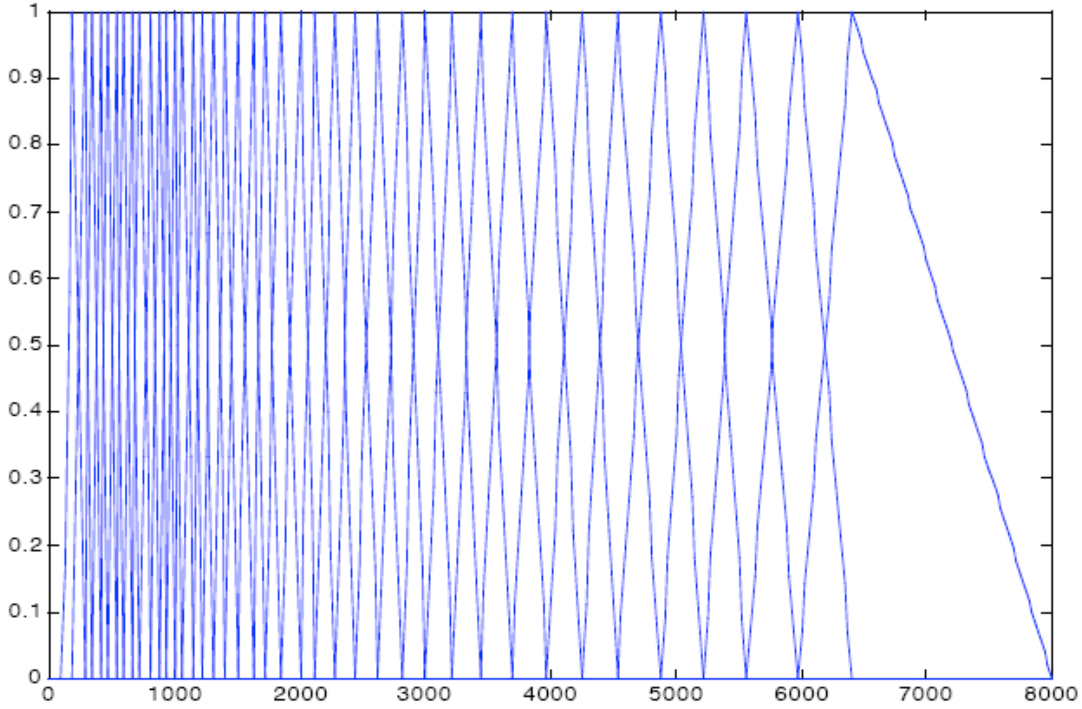
MFCC (Mel-Frequency Cepstral Coefficients)

- The most popular features in speech/audio/music processing.
- Segment incoming waveform into frames
- Compute frequency response for each frame using DFTs
- Group magnitude of frequency response into 25-40 channels using triangular weighting functions
- Compute log of weighted magnitudes for each channel
- Take inverse DCT/DFT of weighted magnitudes for each channel, producing ~14 cepstral coefficients for each frame

Par of slides are from Prof. Hsu, NTU
http://www.csie.ntu.edu.tw/~winston/courses/mm.ana.idx/index.html

The Mel Weighting Functions

- Human pitch perception is most accurate between 100Hz and 1000Hz.
  - Linear in this range
  - Logarithmic above 1000Hz
- A mel is a unit of pitch defined so that pairs of sounds which are perceptually equidistant in pitch are separated by an equal number of mels



Clip-Level Features

- To extract the semantic content, we need to observe the temporal variation of frame features on a longer time scale.
- **Volume-based features:**
  - VSTD (volume standard deviation)
  - VDR (volume dynamic range)

    $$(\max(v) - \min(v)) / \max(v)$$
  - Percentage of low-energy frames: proportion of frames with rms volume less than 50% of the mean volume within one clip
  - NSR (nonsilence ratio): the ratio of the number of nonsilent frames
  - …

Clip-Level Features

- **ZCR-based features**:
  - With a speech signal, low and high ZCR periods interlaced.
  - ZSTD (standard deviation of ZCR)
  - Standard deviation of first order difference
  - Third central moment about the mean
  - Total number of zero crossing exceeding a threshold
  - Difference between the number of zero crossings above and below the mean values

J. Saunders, “Real-time discrimination of broadcast speech/music,” Proc. of ICASSP, vol. 2, 1996, pp. 993-996.

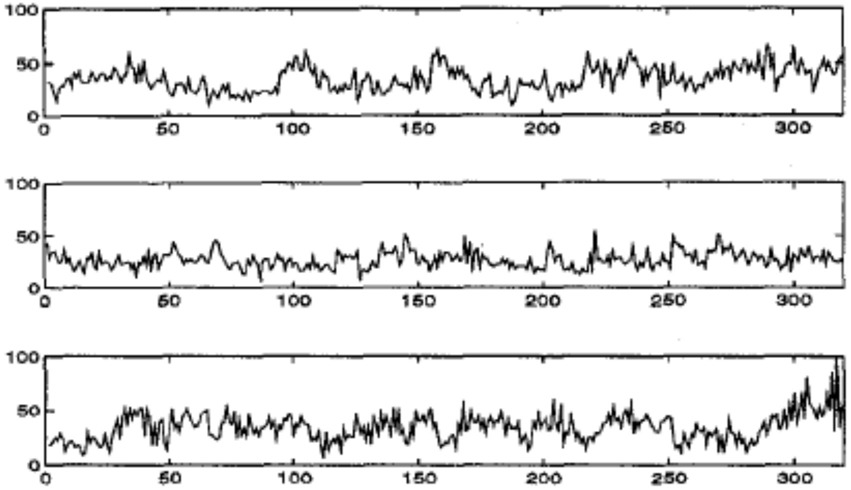


Figure 1: Zero crossing rate contour for music

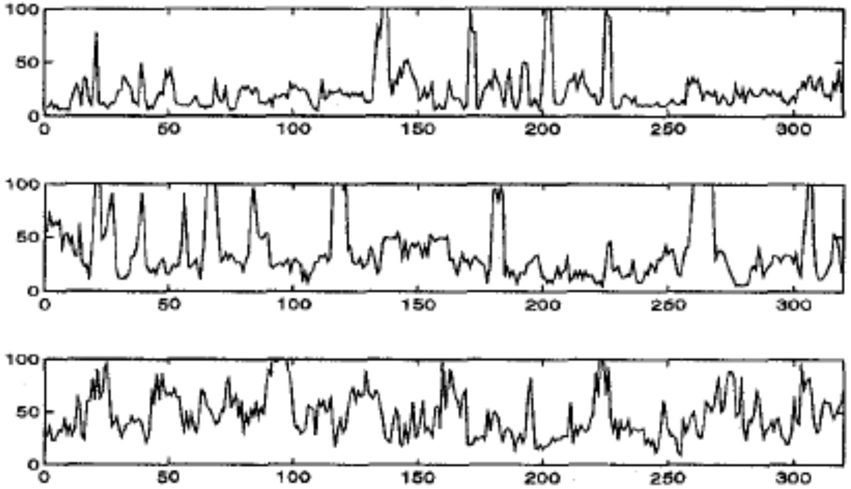
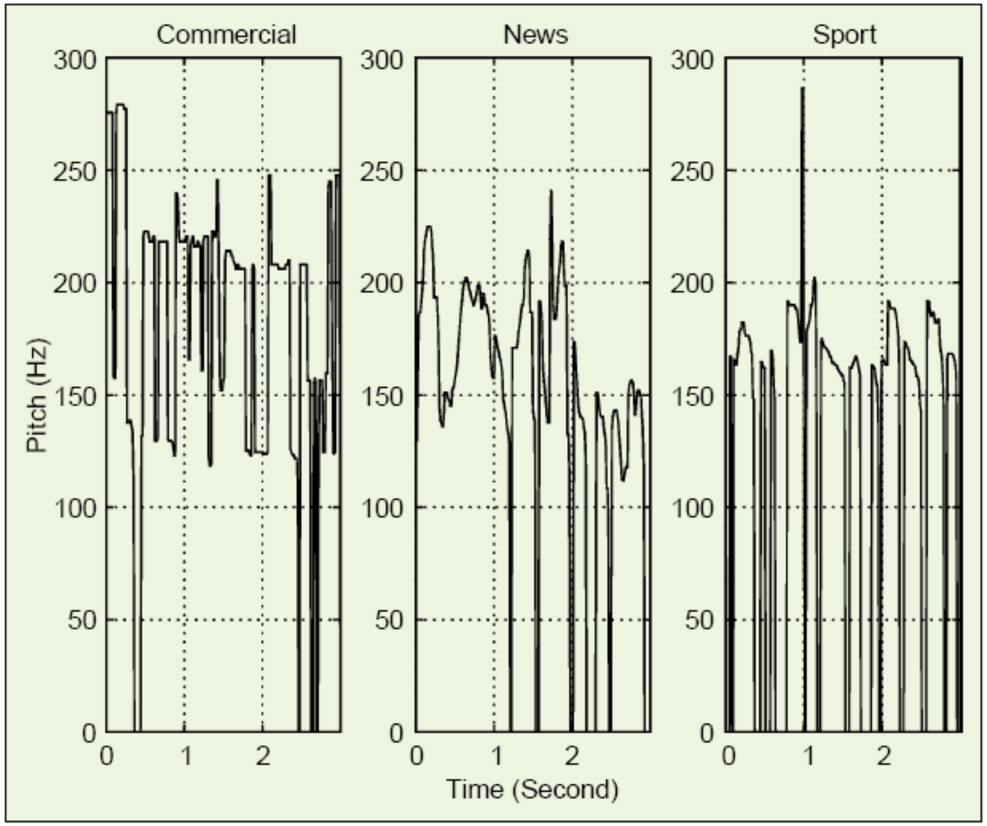


Figure 2: Zero crossing rate contour for speech

Clip-Level Features

- **Pitch-based features:**
  - PSTD (standard deviation of pitch)
  - SPR (smooth pitch ratio): the percentage of frames in a clip that have similar pitch as the previous frames
    - Measure the percentage of voiced or music frames within a clip
  - NPR (nonpitch ratio): percentage of frames without pitch.
    - Measure how many frames are unvoiced speech or noise within a clip


Commercial
News
Sport
Pitch (Hz)
Time (Second)