Role of pitch in speech

- **Pitch** is the dimension of auditory perception that makes it possible to rank sounds on a scale from low to high.

Pitch perception

- Sounds that evoke a clear sensation of pitch are **periodic**; models of pitch perception describe how the auditory system estimates the **periodicity** (repetition rate, $T$) or its inverse, the fundamental frequency, $F_0$.

![Waveforms](image)

Pitch Perception - reviews


Models of pitch perception

- Spectral models
  - Fourier analysis of (most) pitch-evoking sounds generates a harmonic spectrum. The frequency location of the lowest harmonic ($F_0$) generally corresponds to the pitch: a sine wave synthesized at the same frequency will have the same pitch as a complex sound with many harmonics.
Models of pitch perception

- Spectral models
  - Problem #1: missing fundamental
  - Problem #2: shifted harmonics
- Harmonic spacing
- Common multiples
- Pattern matching

Models of pitch perception

- Temporal models
  - Estimate the period or repetition rate from the waveform
  - find landmarks (e.g., peaks) and measure the interval between them
  - Problem #1: what if there are multiple peaks?
  - Solution: compute the autocorrelation function

Pitch in speech is determined by the rate of vocal fold oscillation

Pitch variation in speech

- \( F_0 \) – repetition rate of vocal fold vibration
- Pitch – perceived aspect

- log scale:
  - 1 octave = 12 semitones (1 semitone = 6%)
  - The semitone difference, \( D \), between two frequencies \( f_1 \) and \( f_2 \) is calculated as:

\[
D = 12 \times \log_2 \frac{f_2}{f_1}
\]

Log scale

ERB-rate scale

- Moore and Glasberg (1989) auditory model
- ERB: Equivalent Rectangular Bandwidth
- ERB units provide approximately equal distances along the basilar membrane

\[
E = 16.7 \log_{10} (1 + f/165.4)
\]

and

\[
f = 165.4 (10^{0.06E} - 1)
\]
Intrinsic pitch (Intrinsic $f_0$)
- On average, high vowels like /i/ and /u/ are produced on a higher intrinsic pitch than low vowels like /a/.
- Involuntary side-effect or deliberate strategy?
  - Infants show similar pattern (Whalen)
  - Katz & Assmann (2001) neutralized intrinsic pitch differences in vowels but found no change in identification accuracy.

Role of Pitch Information
- Pitch contour is the primary cue for tone recognition
  - Tonal languages rely on pitch level and differences to convey lexical meanings within syllables
- Pitch helps to segregate auditory components from different sound sources

Tone languages
- In tone languages such as Mandarin, a difference in the level and/or movement of pitch associated with a given phoneme can serve to differentiate pairs of words.
- Tone languages may have several tones (e.g., Mandarin: high level, high rising, low falling, high falling)

F0 range in speech
- 80-200 Hz for adult males
- 180-400 Hz for adult females
- 200-600 Hz for young children
- Even-tempered Scale for the Octave Above Middle C

F0 measurement
- $F_0$ estimates
- 48 sentences
- 1 adult male (blue)
- 1 adult female (red)

Tone contrasts in Mandarin

<table>
<thead>
<tr>
<th>Chinese Character</th>
<th>Tone symbol</th>
<th>Tone description</th>
<th>English gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>媽</td>
<td>丅</td>
<td>high level</td>
<td>mother</td>
</tr>
<tr>
<td>麻</td>
<td>丌</td>
<td>high rising</td>
<td>hemp</td>
</tr>
<tr>
<td>馬</td>
<td>丶</td>
<td>low falling rising</td>
<td>horse</td>
</tr>
<tr>
<td>罵</td>
<td>丰</td>
<td>high falling</td>
<td>cold</td>
</tr>
</tbody>
</table>

Source: The tones of Mandarin Chinese (U-IOWA)

Lexical tones in Mandarin

Pitch and vowel identification

• There is a systematic relationship between F0 and formant frequencies across voices (low-pitched voices tend to have lower formants than high-pitched voices and *vice versa*).

Pitch and vowel identification

• Frequency-shifted speech is more intelligible and perceived as more natural when the normal co-variation of F0 and formant frequency is preserved, even when frequency shifts approach or exceed the range found in human speech.
Comparison of vowel identification accuracy (red circles) and sentence recognition (blue circles).

Pitch and consonant voicing

- Voice pitch is higher following a voiceless consonant compared to a voiced consonant.
- Listeners perceive these small changes; voicing judgments are influenced by F0 of the following vowel.

Scatterplot of vowel identification accuracy and HINT sentence recognition.
Perceptual units in speech perception
- words
- syllables
- phonemes, segments
- context-sensitive allophones
- diphones

Suprasegmentals
- Segmental properties
  - phoneme (phone or segment) units
- Suprasegmental properties
  - properties of speech that cross segment boundaries
  - Includes individual voice characteristics (loud, soft, breathy, hoarse, shrill)
  - Includes linguistic properties (stress, intonation, rhythmic and melodic structure)

Prosody
- Properties of speech not derived from the segmental sequence of phonemes.
  - modulations of voice pitch
  - compression and expansion of syllable duration
  - fluctuations in overall loudness
  - perceived patterns of syllable prominence
  - perceived melodic and rhythmic aspects of speech

Melody of speech
- $F_0$ (pitch) variation over the course of a sentence
- Excludes tone contrasts (segmental use of pitch): lexical forms distinguished by different pitch levels and/or pitch movements

Pitch variation in connected speech
- The $F_0$ contour tends to vary smoothly and continuously in connected speech but may terminate abruptly during voiceless segments or silent gaps (e.g. stop closures).

Intonation patterns
- Declination: pitch tends to fall over the course of a sentence or utterance; “declination reset”
Intonation patterns
• topline, midline, baseline

Stress, accent, focus

(Who wrote that novel?)
(a) The dean of our FACULTY wrote that silly book

(Who wrote that novel?)
(b) The DEAN of our faculty wrote that silly book

Stress, accent, focus

1) WHY are you doing this?
2) Why ARE you doing this?
3) Why are YOU doing this?
4) Why are you DOING this?
5) Why are you doing THIS?

Cocktail Party Effect

- Colin Cherry (1953) coined the term cocktail party effect to describe the ability of listeners to attend to a single talker in a mixture of conversations and other background noises.
  - Cues from lip reading
  - Spatial localization
  - Differences in voice pitch

Intonation patterns
• Pitch accents and prominence
• Rules of intonation
• Pitch rises and falls
  - direction (up or down)
  - size (number of semitones)
  - rate of change (semitones/sec)
  - timing (in ms, relative to syllable onset or offset)
F₀ and voice separation

- When two people speak at the same time, it is often easier to understand what they say if the pitches of their voices differ, for example if one voice is male and the other is female.

F₀ and voice separation

- Hypothesis 1: voice separation becomes easier and intelligibility improves as the pitch (F₀) difference between the voices is increased.

F₀ and voice separation

- Voice pitch is rarely constant in natural speech, but changes over time (melody of speech, or prosody). Time variation in voice pitch may help listeners to “track” a target voice in a mixture of voices.

F₀ and voice separation

- Hypothesis 2: voice separation is easier when the natural variation in pitch (F₀) is present, and becomes more difficult when the pitch is held constant (monotone).

F₀ and voice separation

- A high quality speech vocoder was used to construct pairs of sentences on different F₀s.
- The F₀ difference between the two voices was manipulated (0, 1, 2, 4, or 8 semitones).
- F₀ was either constant or variable (natural pitch contour)
Results

- Significant improvement with $\Delta F_0$.
- No effect of $F_0$ modulation.
- No interaction of $\Delta F_0 \times F_0$ modulation.

Results

- Marginally higher scores for intoned sentences at 0 and 1 semitones may stem from momentary differences in $F_0$ between the sentences.

Size variation in natural speech

![Graph showing fundamental frequency and formant frequencies for adults' and children's voices.]

STRAIGHT Vocoder

- Kawahara (1997, 1999)
- High-resolution analysis of time-varying spectrum envelope
- Wavelet-based instantaneous frequency $F_0$ extraction
  - Spectrum envelope ($FF$) scaling
  - Fundamental frequency ($F_0$) scaling

Assmann & Nearey JASA-EL 2007

- Stimuli: vowel triplets, /i/-/a/-/u/ extracted from hVd words spoken by 2 men and 2 women
(1) FF continuum: FF scale factor fixed at 1.0, F0 varied over ± 2 octaves
(2) F0 continuum: F0 scale factor fixed at 1.0, and FF scale factors spanning ± 0.66 octaves.

• Same, but for gender-swapped voices (N=14)

 FF = 0.28F0 + 5.85

r = 0.78; N=208

Developmental changes in F0 and FF

• Is there a systematic relationship between F0 and formant frequencies in children’s voices?
• Age/sex differences?
• Can we manipulate these properties to implement age transformations?

Original (unscaled) voices

Age-transformed (gender preserved)
Age-transformed (gender swapped)

\[ FF = 0.29F_0 + 5.85 \]

\[ r = 0.78, N = 208 \]

Medians of 75 vowels per talker

Boys

Girls

Geometric mean formant frequency (Hz)

Geometric mean F0 (Hz)