SSC: The Science of Talking

How do we express emotions with speech?

Yi Xu, PhD

Emotional expressions in speech

- Human speech conveys not only linguistic messages, but also emotional information
- The emotional content of speech can be perceived even when the message of the utterance is emotionally ambiguous.

- and even when the listener does not know the language (Chuenwattanapranithi et al, 2006; Frick, 1985; van Bezooijen et al., 1983; Scherer, 1999; Scherer et al. 2001)

Why study emotional expressions in speech?

- Scherer (1982:138): During evolution, language and speech were superimposed on a primitive, analog vocal signaling system. Because speech uses the same voice production mechanism and many of the same acoustic features as the more primitive nonverbal system, we find an intriguing intermeshing of verbal and nonverbal aspects in human sound production
- If we can separate the emotional components from the non-emotional ones, it will be a big step forward in understanding speech coding in general
- We may even learn things that can help us understand emotions in general

Emotional cues in speech have been difficult to identify

Current practice: Examine as many acoustic parameters as possible and measure their correlations with multiple emotions

Scherer (2003:233): Synthetic compilation of the review of empirical data on acoustic patterning of basic emotions

	Stress	Anger/rage	Fear/panic	Sadness	Joy/elation	Boredom
Intensity	↗	↗	↗	↘	↗	
F0 floor/mean	↗	↗	↗	↘	↗	
F0 variability		↗		↘	↗	↘
F0 range		↗	↗(↘)	↘	↗	↘
Sentence contours		↘		↘		
High frequency energy		↗	↗	↘	(↗)	
Speech and articulation rate		↗	↗	↘	(↗)	↘

Why have motions in speech been so difficult to study?

1. Hard to disentangle emotional coding from speech coding in the complex acoustic signal
2. Lack of clear, theory-based hypotheses

Are emotions uniquely human?

- It is often believed that only humans have emotions, and emotional intelligence makes humans stand out from other animals
- Whether this belief is right depends on knowing what exactly emotions are
- In general, it is believed that emotions are, first and foremost, internal feelings we experience
- Therefore, emotional expressions are for displaying our internal feelings

Current theories of emotion

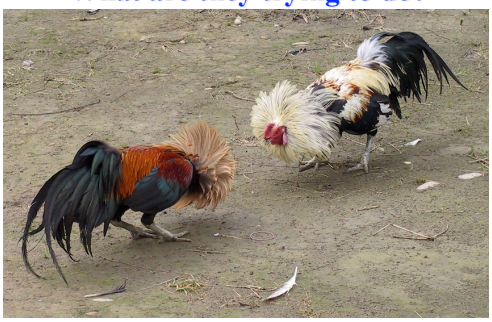
- Discrete emotion theories — There exist a set of basic or fundamental emotions such as anger, fear, joy, sadness, disgust and surprise
- Dimensional theories — All emotions can be placed into a space defined by a set of dimensions, e.g., valence (positive/negative), activation (active/rest), and power/control
- Following these theories, emotional expressions are displays of either discrete emotions, or internal feelings described by the dimensions

A new theoretical perspective

1. Internal feelings are an evolution-engineered mechanism to quickly mobilize all the reactions needed to cope with interactions with other individuals (either conspecific or cross-species), including the act of generating emotional expressions
2. Vocal emotional expressions are evolutionarily designed to elicit behaviours that may benefit the vocalizer

Morton (1977): Many animals express dominance by trying to appear as large as possible

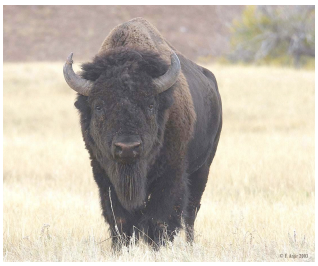
What are they trying to do?



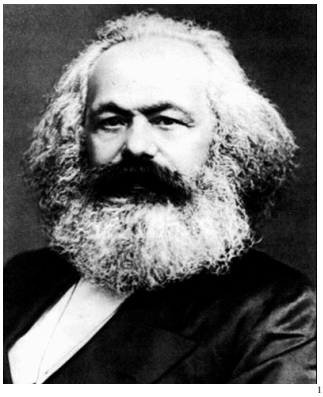
Cock fight

Permanent size markers based on the principle of body size projection

The visual strategy

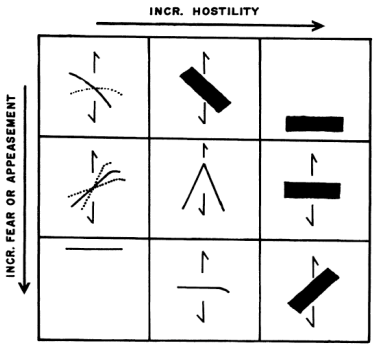


Permanent size markers based on the principle of body size projection



Morton (1977): *The body size projection principle* (motivational-structural rules)

Animal sounds express hostility & fear/appeasement through pitch and voice quality — based on body-size projection



Height = frequency

Thick = harsh

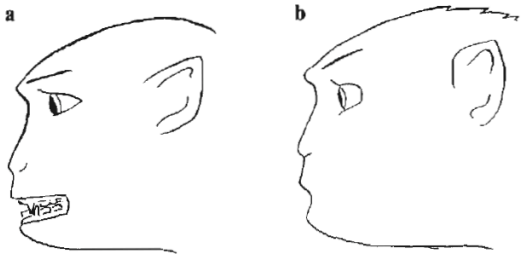
Thin = tonal

Arrow = changeable

Ohala (1984): Human smile -- similar to monkey submission face -- is to understate body size

Monkey facial expressions (van Hooff, 1962, cited by Ohala, 1984)

Submission **Aggression**

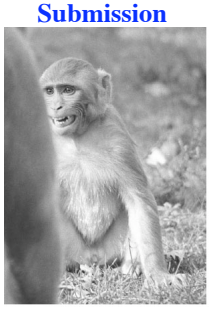


Lip corner retraction **Lip protrusion (O-face)**

Ohala (1984): Human smile -- similar to monkey submission face -- is to understate body size

Monkey facial expressions (van Hooff, 1962, cited by Ohala, 1984)

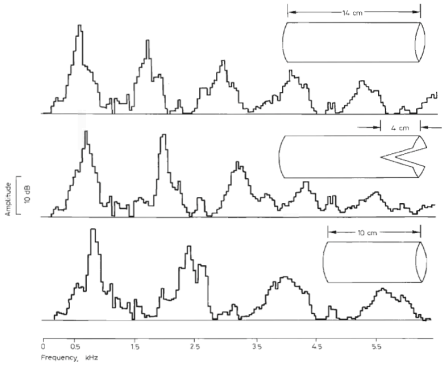
Submission **Aggression**



Lip corner retraction **Lip protrusion (O-face)**

Ohala (1984): Extending body-size projection to formant frequencies

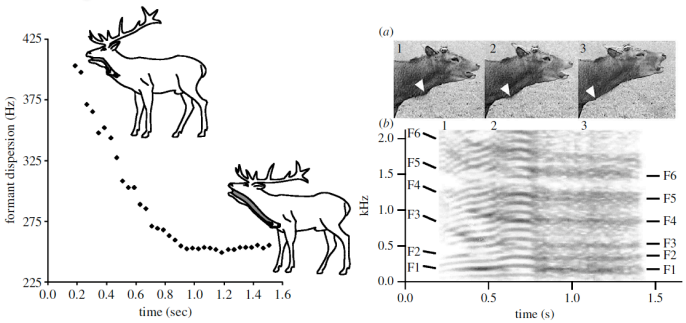
- **Retracting lip corners increases formant frequencies**



Fitch & Reby (2001): Red deer larynx is mobilely descended to exaggerate body size

Male Red deer roars during mating competition

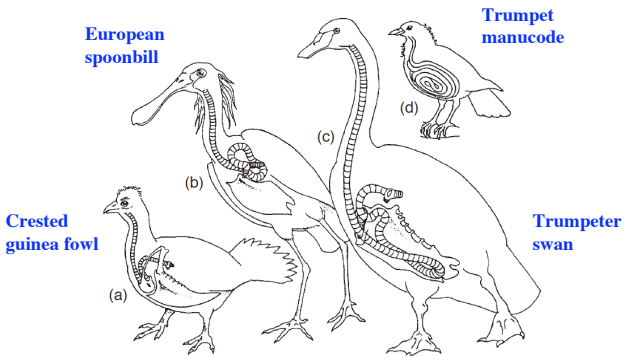
Fitch & Reby (2001): Red deer larynx is mobilely descended to exaggerate body size



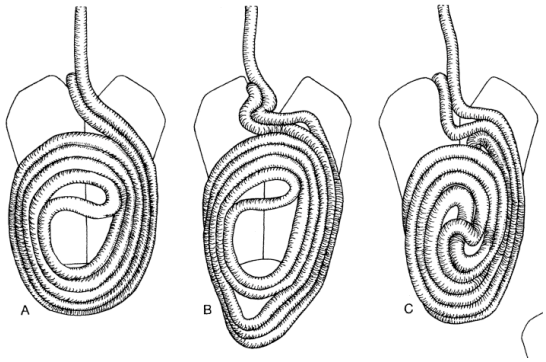
Male Red deer roars during mating competition

Tracheal elongation in birds

(Fitch, 1999)



Elongated and coiled tracheae of birds-of-paradise (Clench, 1978)



Permanently descended larynx in human

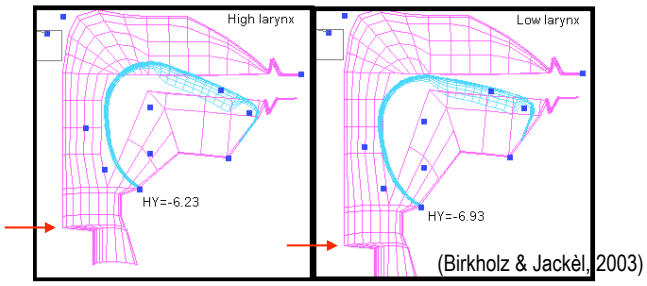
- **Human (as well as Chimpanzee) males have lower larynx than females; their vocal folds are also longer than females' (Fitch, 1994)**
- **The dimorphism occurs at puberty (Negus, 1949; Goldstein, 1980), i.e., just at a time when males have the need to attract females** (Feinberg et al., 2005)
- **Female subjects found male speech with lower pitch and denser spectrum more attractive** (Feinberg et al., 2005)

The size code hypothesis of emotional speech

- **Vocal expression of *anger* is a display of agressiveness and vocal expression of *happiness* a display of sociability**
- **Anger and happiness are vocally conveyed by acoustically exaggerating or understating the body size of the speaker, just as nonhuman animals exaggerate or understate their body size to communicate threat or appeasement.**

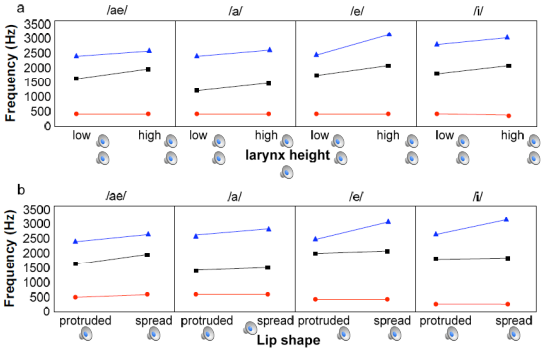
A perceptual test using a 3D articulatory synthesizer

(Chuenwattanapranithi et al. 2008)



Difference in larynx height = 7 mm

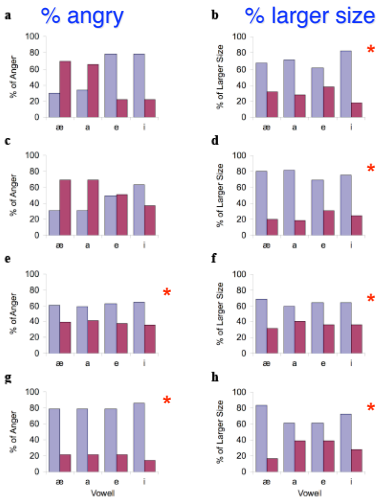
Manipulation of larynx height and lip protrusion resulted in difference in formant frequencies



Experiment 1 — Effect of spectral density and F_0 on perception of emotion and body size

Subjects and procedure

- **393 native Thai speakers (314 males and 79 females; age 19-22 undergraduate students King Mongkut's University of Technology Thailand)**
- **Listened to 8 vowels synthesized with different vocal tract lengths and F_0**
- **196 listeners judged whether the speaker is *larger or smaller in body size***
  197 listeners judged whether the speaker *angry or happy*



Results of emotion and body size perception

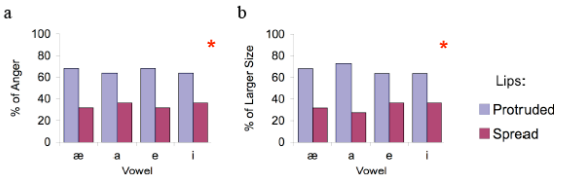
Condition 1 (a, b) — Static low/high larynx; fixed F_0

Condition 2 (c, d) — Static low/high larynx; static high/low F_0

Condition 3 (e, f) — Dynamically lowered /raised larynx; fixed F_0

Condition 4 (g, h) — Dynamically lowered /raised larynx; dynamically lowered /raised F_0

Experiment 2 — Effect of lip protrusion and F_0 on perception of emotion and body size



- Dynamically protruded / spread lips (by 7 mm); dynamically lowered /raised F_0 (by 5 Hz)
- 92 undergraduate Thai students as subjects (72 males and 22 females)

Interpretation

- Listeners are highly sensitive to formant and F_0 variability that may signal body size of the speaker
- They can also use the size information to determine whether the speaker is happy or angry, as long as the acoustic cues are dynamic
- The finding is consistent with the size code hypothesis:

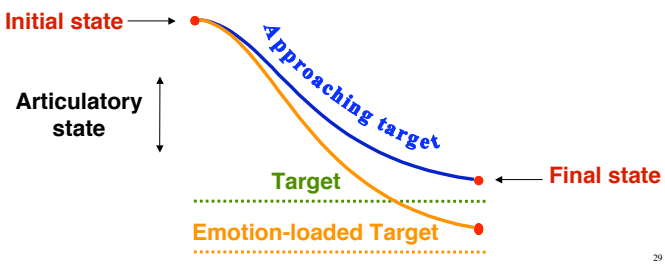
  Anger and happiness are conveyed in speech by exaggerating or understating the body size of the speaker, as if to communicate threat or appeasement

Why dynamic?

- It is as if listeners can "tell" when the acoustic cues indicate "true" body size and when they are used as a code
- From the perspective of encoding, it seems that conveying anger and happiness is not to sound convincingly large or small, but to show an effort to do so
- This is exactly what speakers do when encoding lexical contrasts conveyed by tonal and segmental phonemes (Xu, 1997, 1999; Xu and Liu, 2007; Xu and Wang, 2001)

Encoding emotions by modifying linguistic targets

- The target approximation process is intrinsically dynamic, and it is intrinsic to speech
- Emotions are probably encoded through modification of the existing linguistic targets



A new study — Direct modification of real speech based on the size code

(Kelly, Xu & Huckvale, 2008)

Conditions:

1) Original speech: English numerals: 1, 2, 3, … 10

2) Spectral density: original +10% -10%

3) F_0: original +10 Hz -10 Hz

Tasks:

1. Body size: Larger / Smaller
2. Emotion: Angry / happy

Static stimuli “nine”

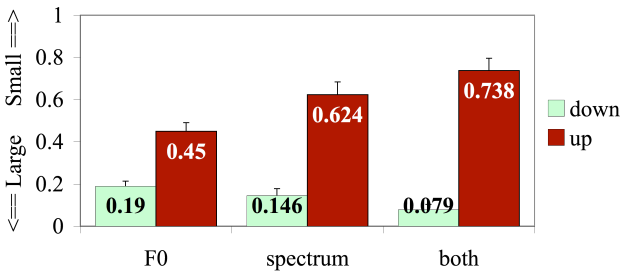
- Original
- Expanded / Condensed spectrum
- Raised / Lowered F_0
- Expanded spectrum + Raised F_0
- Condensed spectrum + Lowered F_0

Dynamic stimuli “nine”

- Original
- Expanded / Condensed spectrum
- Raised / Lowered F_0
- Expanded spectrum + Raised F_0
- Condensed spectrum + Lowered F_0

Results: size judgment

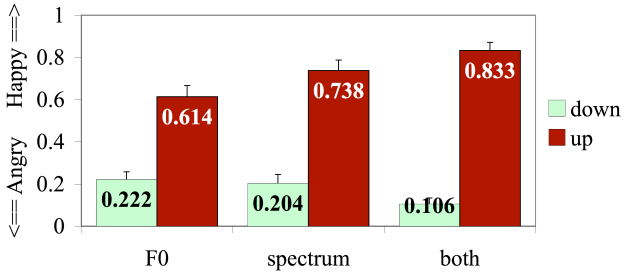
Interaction of direction and parameter of manipulation



Down = Lowered F_0 and/or spectral dispersion

Results: emotion judgment

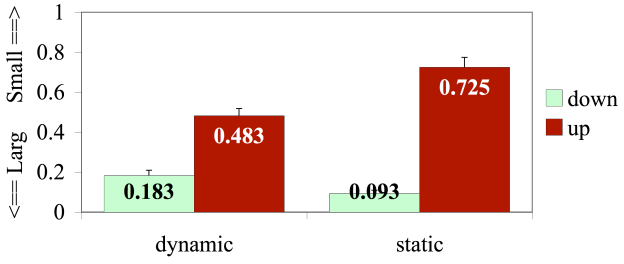
Interaction of direction and parameter of manipulation



Down = Lowered F_0 and/or spectral dispersion

Results 2: size judgment

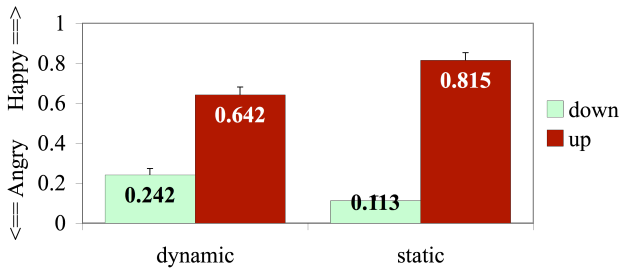
Interaction of manner and parameter of manipulation



Down = Lowered F_0 and/or Spectral dispersion

Results 2: emotion judgment

Interaction of manner and parameter of manipulation



Down = Lowered F_0 and/or Spectral dispersion

Bio-informational dimensions theory — An extension of the size code hypothesis

(Xu, Kelly & Smillie, in press)

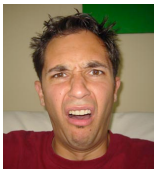
- Vocal emotional expressions are evolutionarily designed to elicit behaviours that may benefit the vocalizer
- They influence the behaviour of the receivers by manipulating the vocal signal along a set of bio-informational dimensions:
  - *size projection*
  - *dynamicity*
  - *audibility*
  - *association*

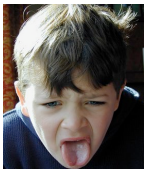
The bio-informational dimensions

- ***size projection*** = size code
- ***dynamicity:*** controls how vigorous the vocalization sounds, depending on whether it is beneficial for the vocalizer to appear strong or weak.
- ***audibility:*** controls how far a vocalization can be transmitted from the vocalizer, depending on whether and how much it is beneficial for the vocalizer to be heard over long distance.
- ***association:*** controls associative use of sounds typically accompanying a non-emotional biological function in circumstances beyond the original ones. For example, the disgust vocalization seems to mirror the sounds made when a person orally rejects unpleasant food (Darwin, 1872).

The association dimension

Darwin (1872:262 The Expression of the Emotions in Man and Animals) We have now seen that scorn, disdain, contempt, and disgust are expressed in many different ways, by movements of the features, and by various gestures; and that these are the same throughout the world. They all consist of actions representing the rejection or exclusion of some real object which we dislike or abhor, but which does not excite in us certain other strong emotions, such as rage or terror; and through the force of habit and association similar actions are performed, whenever any analogous sensation arises in our minds.





An initial test of bio-informational dimensions

- Original speech: English sentence "I owe you a yoyo" with focus on "owe"
- Parameter manipulations:

Formant shift ratio	Pitch median (Hz)	Pitch range factor	Duration factor
1.2	400	4	1.1
1.078	200	1.17	0.9
0.956	100	0.341	
0.833	50	0.1	

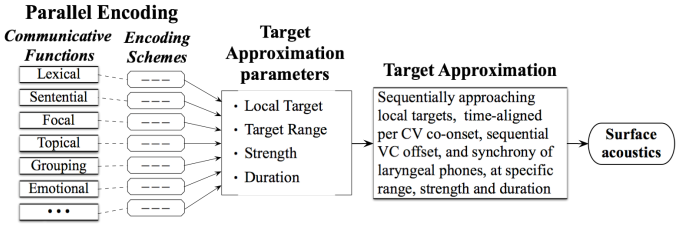
Results

Emotion	Best score (%)	Formant shift	Pitch median	Pitch range	Duration ratio
Happy	73.3	1.2***	200	4.0***	0.9***
Depressed	71.1	0.96	100***	0.1***	1.1***
Grief-stricken	60	0.83***	400**	0.34***	1.1**
Scared	48.9	0.83**	400***	1.17	0.9
Angry	48.9	0.83	50***	0.1	1.1

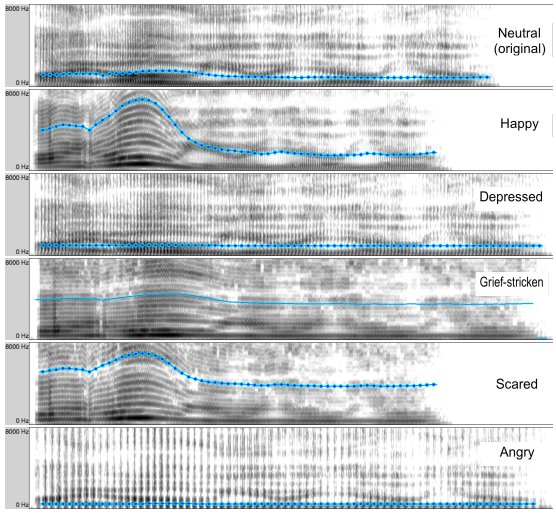
Key findings

- Happiness is located toward the small end of the *size-projection* dimension and high end of *dynamicity* dimension
- Two types of sadness: *depressed*, corresponding to most commonly reported sadness, *grief-stricken*, with lengthened vocal tract, suggesting that it is demanding (not begging) for sympathy (straightMorph.swf)
- Fear has lengthened vocal tract and relatively large pitch range. Combined with high median pitch, it sends a mixed signal: I may be small (high pitch), but I am willing to fight (long vocal tract). This separates fear from submission, counter Morton (1977).
- While submission probably indeed signals total surrender, a fear expression signals a demand for the aggressor to back off. This makes evolutionary sense, because a total surrender to a predator can only mean one thing: to be eaten.

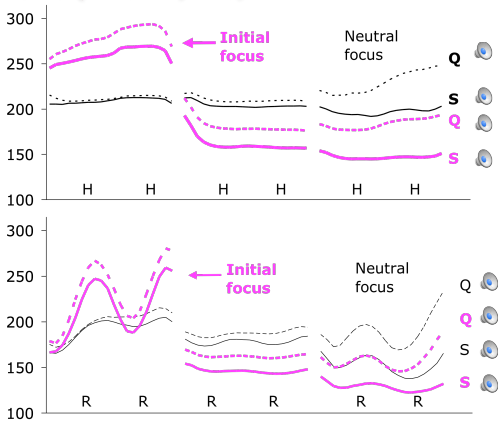
The Parallel Encoding and Target Approximation model of speech prosody (PENTA) (Xu, 2005)



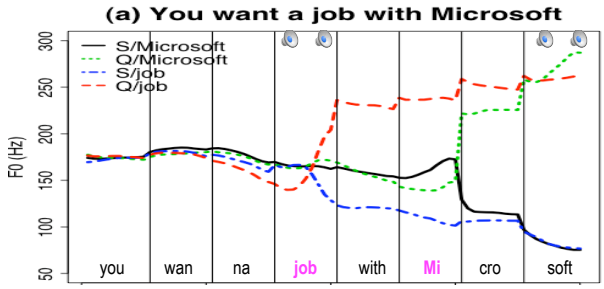
- Speakers convey to the listeners not only words, but also many layers of additional information.
- They do so by controlling Target Approximation parameters specified by encoding schemes of various communicative functions, including the emotional functions.



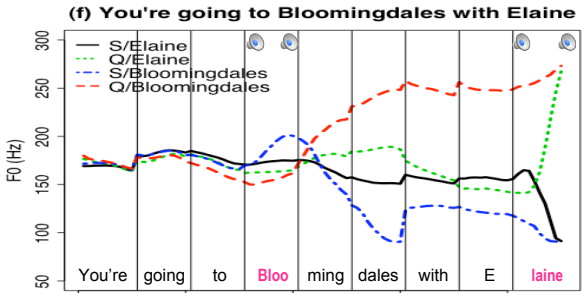
Mandarin: Statement vs. question + focus + tone (Liu & Xu, 2005)



English: Statement vs. question + focus + word stress (Liu & Xu, 2007)

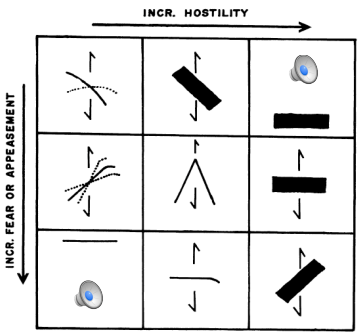


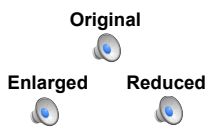
English: Statement vs. question + focus + word stress (Liu & Xu, 2007)



What about voice quality?

Morton (1977):





Without a harsh voice quality, speech does not sound obviously angry

Gobl & Ni Chasaide (2003): harsh/tense voice ⇒ anger

?

Summary

We are now starting to crack the emotion code:

1. **Vocal expression of emotion is not for the sake of revealing one's internal feelings**

   Rather, it is likely to be an evolutionarily engineered way of eliciting behaviors from the listener that may benefit the vocalizer

2. **This is done mainly by projecting a large body size to scare away the listener, or a small body size to attract the listener**

3. **Size projection is also accompanied by other vocal manipulations to enhance the beneficial effects**

4. **All of this is done in parallel with the transmission of linguistic information**

References

Birkholz, P.; Jackèl, D.: A three-dimensional model of the vocal tract for speech synthesis. Proc. The 15th International Congress of Phonetic Sciences, pp. 2597-2600 (Barcelona, Spain 2003).

Chuenwattanapranithi, S., Xu, Y., Thipakorn, B. and Maneewongvatana, S. (2008). Encoding emotions in speech with the size code — A perceptual investigation. Phonetica 65: 210-230.

Clench, M. H. (1978). Tracheal elongation in birds-of-paradise. *Condor* **80**: 423-430.

Feinberg, D. R.; Jones, B. C.; Little, A. C.; Burt, D. M.; Perrett, D. I.: Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices. Animal Behavior 69: 561-568 (2005).

Fitch, W. T.: Vocal tract length perception and the evolution of language. Ph. D. Dissertation (Brown University, 1994).

Fitch, W. T. (1999). Acoustic exaggeration of size in birds by tracheal elongation: Comparative and theoretical analyses. *Journal of Zoology (London)* **248**: 31-49.

Fitch, W. T. and Reby, D. (2001). The Descended Larynx Is Not Uniquely Human. *Proceedings of the Royal Society, Biological Sciences* **268**(1477): 1669-1675.

Frick, R. W. (1985). Communicating emotion: The role of prosodic features. *Psychological Bulletin* **97**: 412-429.

Goldstein, U. (1980). *An articulatory model for the vocal tracts of growing children.* Ph.D. dissertation, Massachusetts Institute of Technology.

Morton, E. W. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *American Naturalist* **111**: 855-869.

Negus, V. E.: The comparative anatomy and physiology of the larynx (Hafner, New York 1949).

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. Phonetica 41: 1-16.

References (cont.)

Scherer, K. R. (1982). Methods of research on vocal communication: paradigms and parameters. In *Handbook of Methods in Nonverbal Behavior Research.* K. R. Scherer and P. Ekman. Cambridge, UK: Cambridge University Press pp. 136-198.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication* **40**: 227-256.

Scherer, K. R. (1999). Universality of emotional expression. In *Encyclopedia of Human Emotions.* D. Levinson, J. Ponzetti and P. Jorgenson. New York: Macmillan pp. 669-674.

Scherer, K. R., Banse, R. and Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology* **32**: 76-92.

van Bezooijen, R., Otto, S. and Heenan, T. A. (1983). Recognition of vocal expressions of emotions: A three-nation study to identify universal characteristics. *Journal of Cross-Cultural Psychology* **14**: 387-406.

Xu, Y.: Effects of tone and focus on the formation and alignment of F0 contours. Journal of Phonetics 27: 55-105 (1999).

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics* 25: 61-83.

Xu, Y. (2005). Speech Melody as Articulatorily Implemented Communicative Functions. *Speech Communication* 46: 220-251.

Xu, Y. (2007). Speech as articulatory encoding of communicative functions. In *Proceedings of The 16th International Congress of Phonetic Sciences,* Saarbrucken**:** 25-30.

Xu, Y., Kelly, A. and Smillie, C. (forthcoming). Emotional expressions as communicative signals. To appear in S. Hancil and D. Hirst (eds.) *Prosody and Iconicity.*

Xu, Y.; Liu, F.: Determining the temporal interval of segments with the help of F0 contours. Journal of Phonetics 35: 398-420 (2007).

Xu, Y. and Wang, Q. E. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication* 33: 319-337.