Universität Bielefeld

Multimodal Prosody

Petra Wagner

Phonetics and Phonology Workgroup, Faculty of Linguistics and Literary Studies
Center of Excellence Cognitive Interaction Technology (CITEC)
Bielefeld University

2nd Workshop on Uralic Prosody 2017 — Budapest

Speaking is Intrinsically Multimodal Activity

Speaking is Intrinsically Multimodal Activity

Modalities are inseparable, but...

- What is their functional relationship? (complementary, parallel, additive)
- What is their relationship in form? (temporal, shape?)

Why is prosody interesting to study this multimodal relationship?

Both speech prosody and gesture are similar in structure and function. Unlike verbalizations, they are continuous rather than discrete and are expressed on a different time scale than verbalizations. The temporal coordination between prosody and gesture seems stronger than that between gesture and verbalizations in general, despite their being expressed in different physiological systems.

(Wagner, Malisz and Kopp, 2014)

Central topics covered

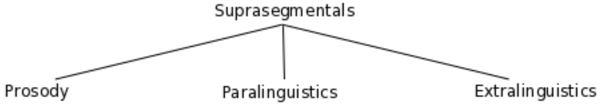
1. Which prosodic functions are expressed in a multimodal fashion, and how?
2. What is the cross-modal relationship in prosody perception and production?

Central topics covered

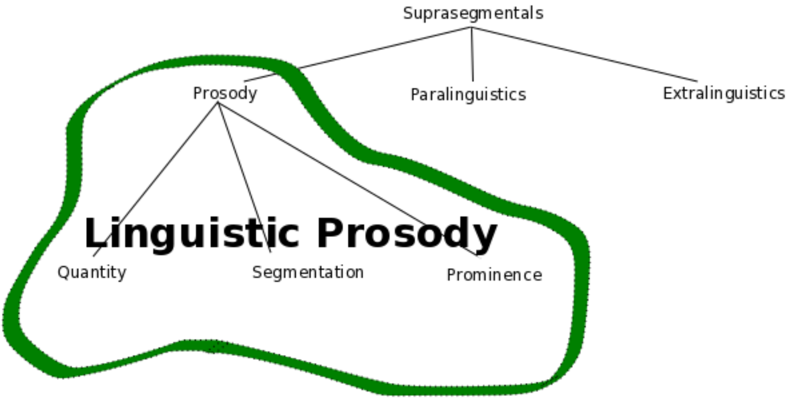
Topic 1

Which prosodic functions are expressed in a multimodal fashion, and how?

Yet another taxonomy of how we talk about “prosody”



Linguistic Prosody in a Narrow Definition

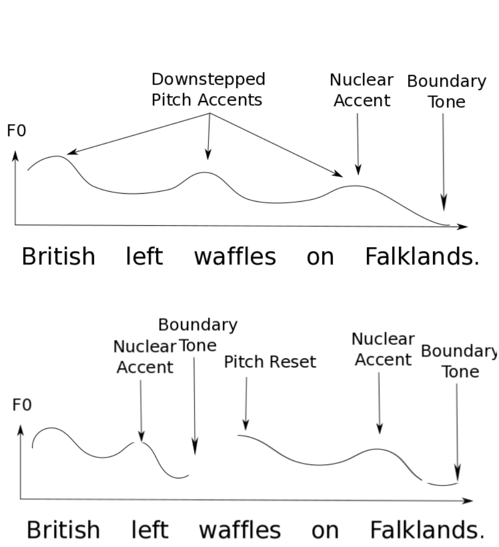


Linguistic Prosody — Phrasing

??

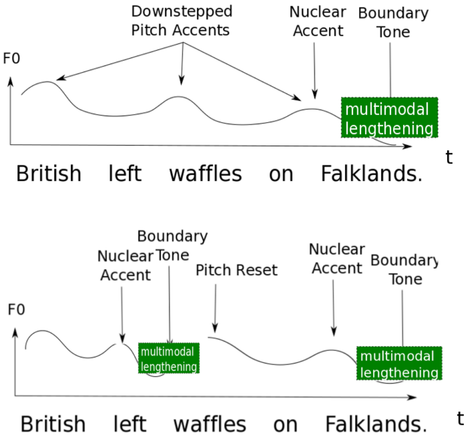
“British left waffles on Falklands”

Phrasing



Downstepped
Pitch Accents
Nuclear
Accent
Boundary
Tone
F0
British left waffles on Falklands.
Boundary
Tone
Nuclear
Accent
Pitch Reset
Nuclear
Accent
Boundary
Tone
F0
British left waffles on Falklands.

Multimodal Phrasing (Krivokapic et al., 2016)



Linguistic Prosody: Discourse Segmentation/Floor Management

Function
Function: Who's next??

1. A: British left waffles on Falklands.
2. B: Oh, okay.
3. A OR B: I really would have preferred donuts.

Prosody helps indicating:

- Feedback Function
- Floor management
- Inter-speaker relationship (e.g. familiarity, attention)

Figure 2: Who's going to talk about donuts?

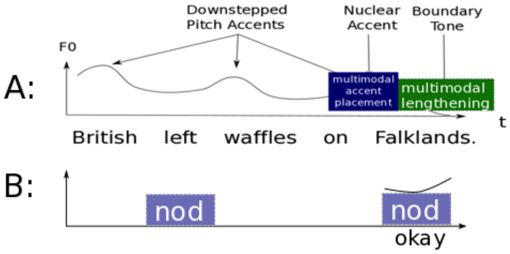
Linguistic Prosody: Discourse Segmentation/Floor Management

Multimodal Form

Expressing Feedback Function

- Attention: High frequency, loud, multimodal, high in pitch variation (Malisz et al., 2016 ; Ishi et al., 2012)
- Familiarity: less (multimodal) feedback (Ishi et al., 2012)
- Turn ending confirmation: Multimodal rather than visual only (Ishi et al., 2014); Nonverbal feedback distributed uniformly across dialogue.(Inden et al., 2013)
- Backchanneling: Short verbal feedback with rising intonation (e.g. Benus et al., 2007; Edlund & Heldner, 2011) ; nods rather than jerks/tilts Allwood & Cerrato, 2003; Wlodarczak et al., 2012

Linguistic Prosody: Discourse Segmentation/Floor Management


A:
Downstepped Pitch Accents
Nuclear Accent
Boundary Tone
F0
multimodal accent placement
multimodal lengthening
t
British left waffles on Falklands.
B:
nod
nod
okay

A: I really would have preferred donuts.

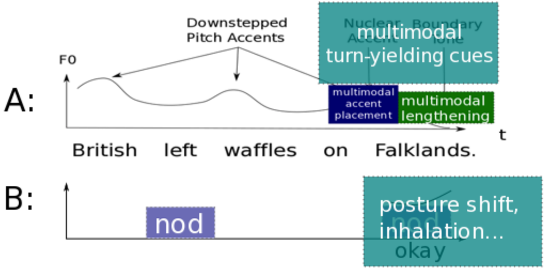
Linguistic Prosody: Discourse Segmentation/Floor Management

Multimodal Form

Floor Management

- Acoustic turn yielding cues: strong boundary signals, phonetic reduction, creaky voice, no audible inhalation (e.g. Selting, 1996; Kelly et al., 1986; Niebuhr et al., 2013; Gravano & Hirschberg, 2011; Local & Walker, 2012; Ogden, 2001; Zellers, 2016)
- Nonverbal turn yielding cues: gaze towards interlocutor, discontinued manual gesture, return to pre-turn body posture. (e.g. Argyle & Cook, 1976; Beattie et al., 1982; Cassell et al., 2001; Edlund & Beskow, 2009; Barkhuysen et al., 2008; Zellers et al., 2016)
- Multimodal cues superior to acoustic cues only (Barkhuysen et al., 2008)

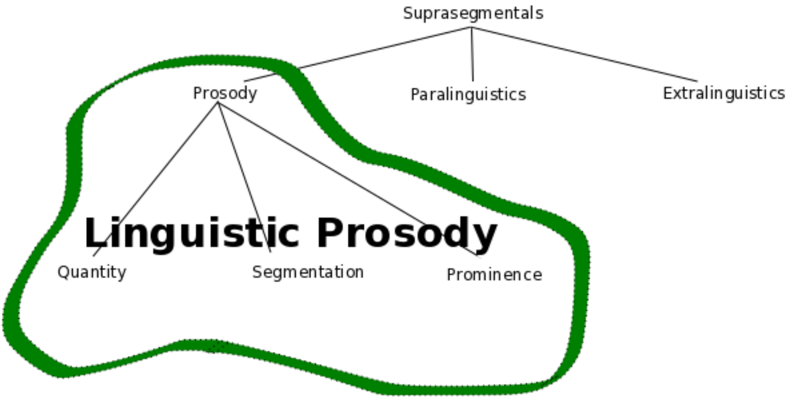
Linguistic Prosody: Discourse Segmentation/Floor Management



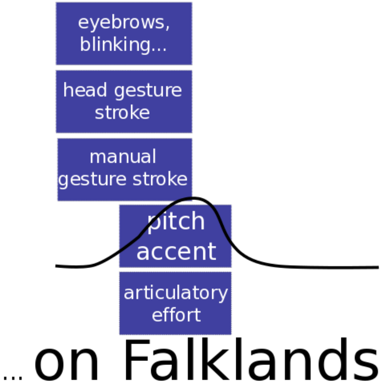
B: I really would have preferred donuts.

Linguistic Prosody: Prominence

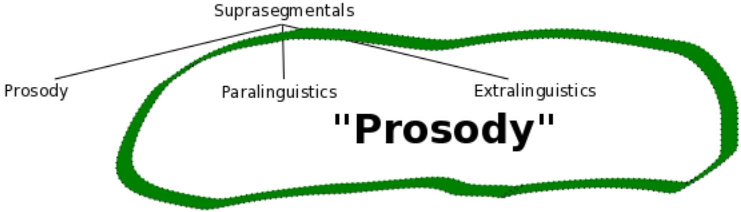
Function
Shaping information structure, monitoring attention



Linguistic Prosody: Multimodal Prominence (Leonard and Cummins, 2009; Krahmer and Swerts, 2007; Esteve-Gibert and Prieto, 2015)



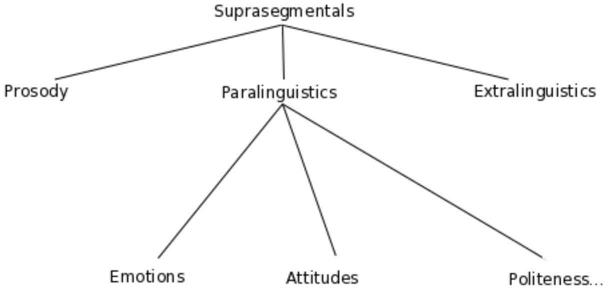
Prosody in a Less Narrow Definition



Prosody in a Less Narrow Definition: Paralinguistics

General Concept:

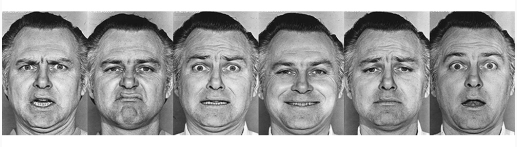
Nonlinguistic expressions that are (partly) controlled by a speaker, e.g. emotions — attitudes



Prosody in a Less Narrow Definition: Paralinguistics

Emotions:

Facial expression more robust than acoustics



Prosody in a Less Narrow Definition: Paralinguistics

Attitudes:

Facial expression and acoustics work hand-in-hand in the expression of attitudes, e.g. “amusement” (Auberge & Cathiard, 2003)

Prosody in a Less Narrow Definition: Paralinguistics

Attitudes:

Some attitudes are better heard than seen and vice versa (Hönemann & Wagner, 2017)

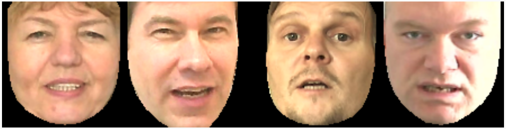


Figure 4: polite — doubt — arrogance — contempt

Conclusions I

- Speech Prosody and gestural prosody are tighly coupled in structure, function and form.
- Rich cues across modalities probably cause redundancy necessary for robust information transmission.

Central topics covered

Topic 2

What is the cross-modal relationship in prosody perception and production?

The cross-modality link in perception

How much linguistic prosody is captured in gesture?

PromDrum: Investigating the cross-modality link in perception

Gesture-based annotation method

Drumming force modulation as indicator of perceived prominence

- Task fast and easy, esp. syllable drumming
- Strong individual differences in drumming profiles
- Expert-like mean annotation profiles
- Individual types/strengths of speech-gesture linkages in perception?

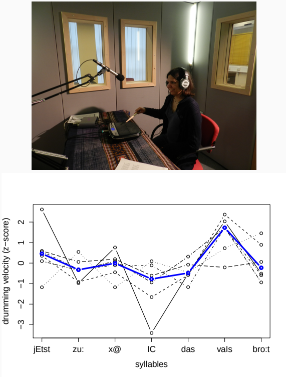


Figure 5: Annotation procedure and impact profiles (median=blue)

What do listeners model in terms of drumming force?

Random Forests trained for individual annotators' impact forces

Predictor variables: F0, syllable duration, rhythmic predictability, word class (POS)

- Individual strategies unfold

annotator	*F0*	*SyllDur*	*POS*	*Clash*	*Accent-Dist*	*% Variance Explained*
1	**59**	**48**	**62**	22	22	44
2	39	**61**	**60**	8	**55**	58
3	37	**49**	**64**	8	**39**	36
4	**14**	**14**	**12**	11	10	0
5	29	**41**	**46**	24	**54**	29
6	**34**	**12**	**15**	0	4	7
average annotator	**97**	77	**91**	**87**	80	76

Table 1: Importance (%) of the various predictor variables per annotator in the word drumming task.

Conclusions II

- Gestural prosody able to capture rich prosodic information, but shows large inter-individual variation.

Investigating the cross-modality link in production

Speech-gesture co-ordination

Congruency or economic trade-off?

- Prosody and gesture are known to adapt economically to communicative needs (Lombard, 1909; Lindblom, 1990; Hoetjes et al., 2015).
- A strong assumption of speech-gesture co-ordination expects that prosodic expression is mirrored across domains.
- What if information in either channel is redundant? Does speech economy "win"? Or speech-gesture congruency?

Figure 6: Gesture excursion increases given mutual visibility (Hoetjes et al., 2015)

Investigating the cross-modality link in production

- Data: German speakers; (10 dyads, 20 speakers, ± visibility)
- Non-visibility = redundancy of gestures
- Visibility = redundancy of verbal prosody
- First move was (quasi-randomly) preset by experimenter
- Majority of games ended as ties.

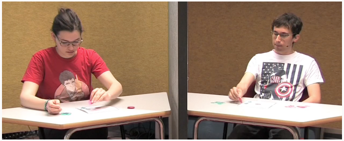


Figure 7: No mutual visibility

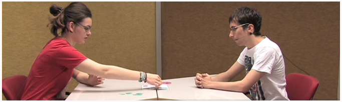


Figure 8: Mutual visibility

Investigating the cross-modality link in production

Controlling information structure

given — unpredictable — important

- TicTacToe setting differentiates "given" (first, last), unpredictable (but irrelevant), and important (but predictable) moves. (Watson et al., 2008)
- Co-speech movements (game moves) similar in excursion across visibility conditions.

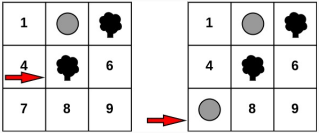


Figure 9: Unpredictable (left) and important (right) move in TicTacToe.

Investigating the cross-modality link in production: acoustic prosody

- Invisibility leads to increase in acoustic-prosodic effort (louder, slower) → pro speech-economy
- F0 not affected by visibility, only by information structure! → pro speech-gesture co-ordination

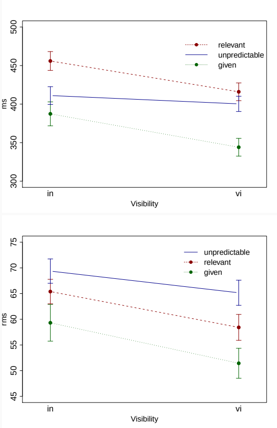


Figure 10: Duration (top) and RMS intensity (bottom) influenced by informativeness and visibility

So how is acoustic prosody influenced by gesturally transmitted information?

Acoustic prosody invests less intensity and duration if redundant. (pro speech economy models)

F0 shape and excursion is not affected by redundancy, staying aligned with the level of gestural excursion (pro speech-gesture congruency models)

Investigating the cross-modality link in production: gesture

- Does visibility and informativity influence the speech gesture synchronization?
- Dependent variable: Delay between movement apex (target) and pitch peak/prosodic boundary.

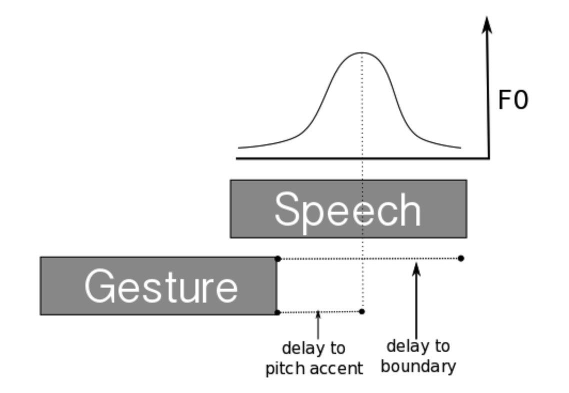


Figure 11: Measuring prosody-gesture delay.

Investigating the cross-modality link in production: gesture

- No visibility: More asynchrony between acoustics and gesture; move variability in speech-gesture co-ordination.
- Visibility makes gesture apices preceed pitch accents and boundaries (“gesture lead”).

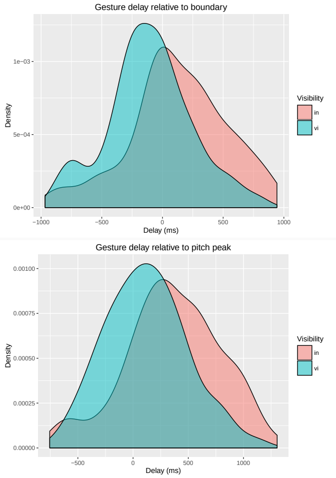


Figure 12: Delay to boundaries (top) and pitch accents (bottom) with and without visibility

Investigating the cross-modality link in production: gesture

- Uninformative (given) moves: less speech-gesture synchrony, more variability in cross-modal co-ordination
- Informative moves: Stronger speech-gesture synchronization; less variability in cross-modal co-ordination

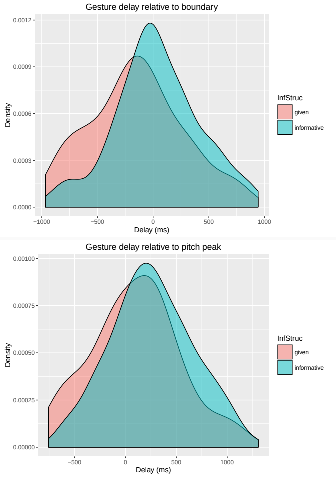


Figure 13: Delay to boundaries (top) and pitch accents (bottom) with and without informatitiy

So how are gestures affected by linguistic information structure and mutual visibility?

Both mutual visibility and information load lead to a stronger co-ordination of speech and gesture.

Mutual visibility increases gesture lead, information load increases synchronization with speech prosody.

Conclusions III

- Lack of mutual visibility and information load reduces the temporal co-ordination of speech and gesture prosody.
  → prosodic speech—gesture co-ordination has a communicative function.
  → A strong version of prosodic speech-gesture co-ordination is rejected.
- Tempo and intensity production affected by information transmitted visually
  → pro speech economy models.
- F0 excursion and shape not affected by gesturally transmitted information and serve as acoustic anchor for gesture.
  → Contra speech economy models, pro speech-gesture co-ordination and redundancy.

Caveat and “Take Home Message”

The speech-gesture link in prosody looks strong, but can apparently be modulated and vary. Before making far-reaching claims about the speech-gesture link in prosody, we should

- study more languages and cultures.
- take individual differences in speakers and listeners seriously.

Közönöm — Kiitos — Thanks — Danke

Questions — comments