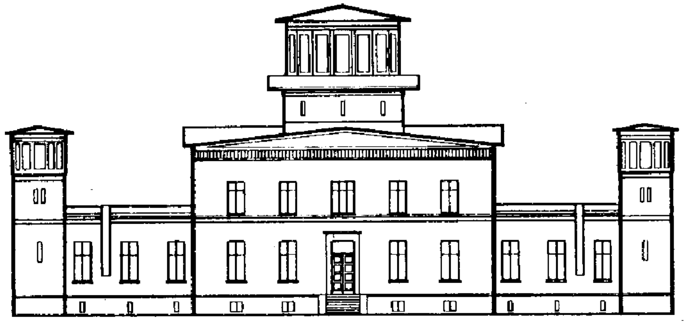
Uni Bonn

Rheinische Friedrich-Wilhelms-Universität

2004/03/26
SP2004 Nara



SP 2004 Summary
A Personal (and Impressionistic) View

Wolfgang Hess

Institut für Kommunikationsforschung und Phonetik (IKP)
Universität Bonn
Poppelsdorfer Allee 47, 53115 Bonn, Germany

Visiting Professor, University of Tokyo, Japan

wgh@ikp.uni-bonn.de
http://www.ikp.uni-bonn.de

The 2-7-1 Rule ...

... or How to Present a Ten-Minute Paper -
A Beginner's Rule [IEEE ICASSP Presentation Guide]

- First, tell them what you are going to say (2 minutes)
- Then, say it (7 minutes)
- Last, tell them what you just said (1 minute)

The Same Rule Applied to a Conference (Like SP2004)

- First, tell them what is going to be presented (Fujisaki, Lehiste, Sugito, ...)
- Then, present it (oral and poster sessions)
- Last, tell them what was just presented (Hess)

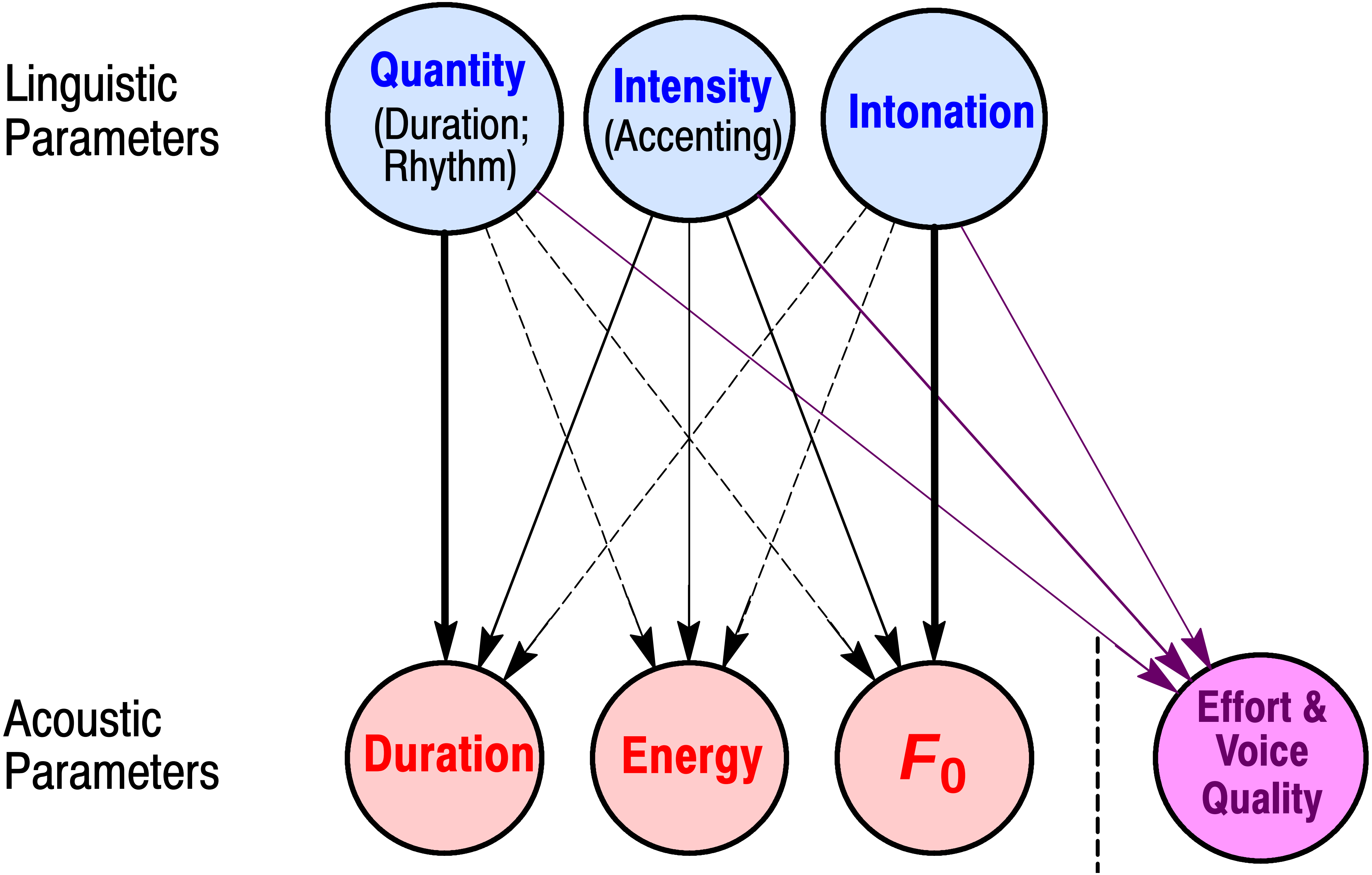
Outline of this Talk

- **General Topics and Numbers**
- A Glimpse at Individual Languages
- Modelling
- Prosody and Music
- Measurement
- Prosody and Voice Quality
- Corpora and Speech Technology
- Some Open Questions to Take Home

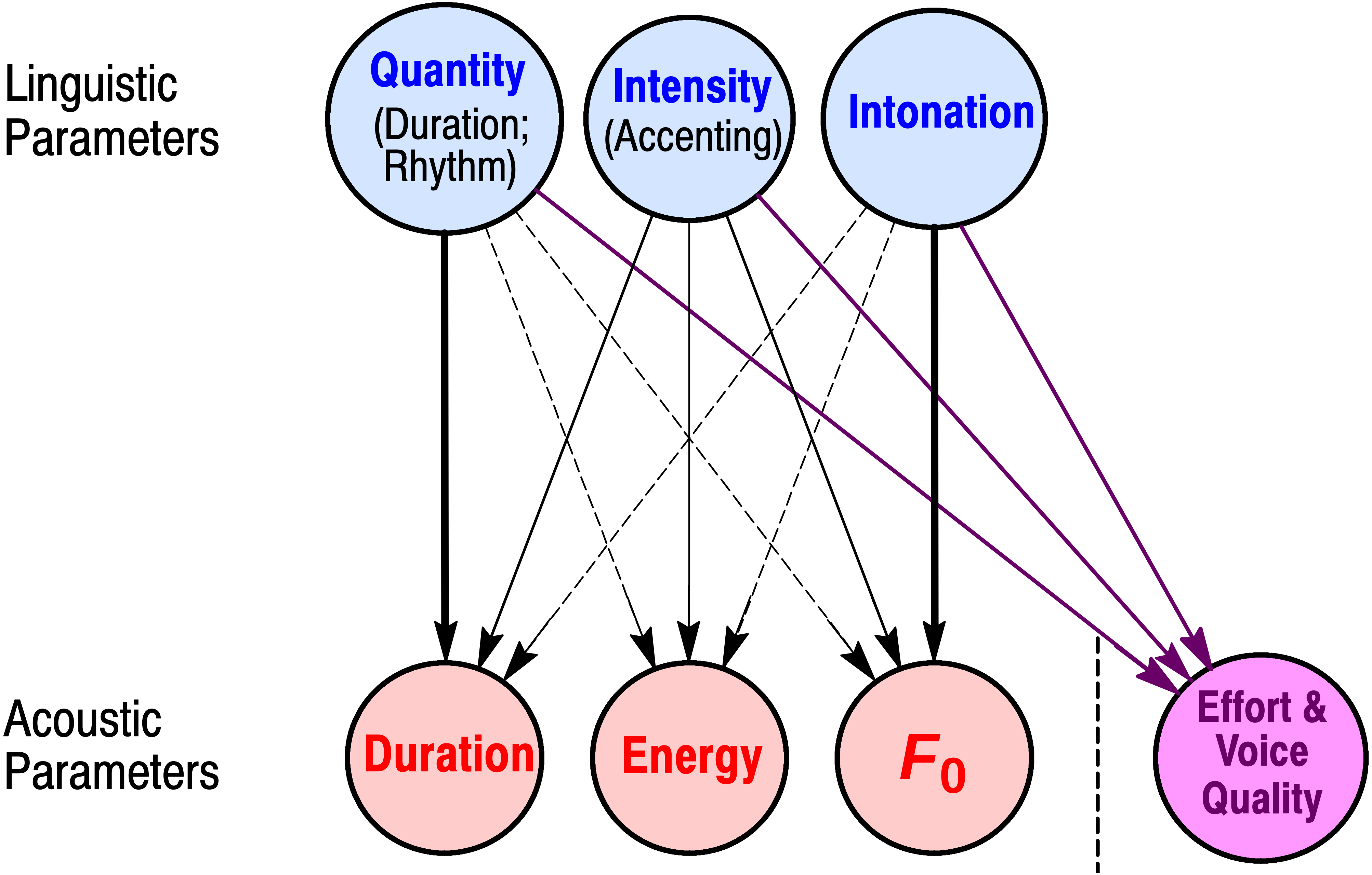
Conference Topics

- accent, stress, focus, emphasis, prominence
- tone, intonation
- syntax, semantics, pragmatics, and prosodic structure
- temporal structure, discourse, and dialogue
- phonology and phonetics of prosody
- analysis, modeling, and generation of prosody
- prosody in music
- prosody and voice quality; prosody and emotional expression
- prosody and paralinguistic/extralinguistic information
- prosody in multimodal systems
- control of prosody for high-quality and expressive speech synthesis
- prosody in speech technology: recognition, understanding, and automatic summarization
- prosodic annotation and corpora
- prosody and perception
- prosody and speech pathology

Interaction of Linguistic and Acoustic Parameters in Prosody



Interaction of Linguistic and Acoustic Parameters in Prosody



Linguistic Parameters

Quantity (Duration; Rhythm)

Intensity (Accenting)

Intonation

Acoustic Parameters

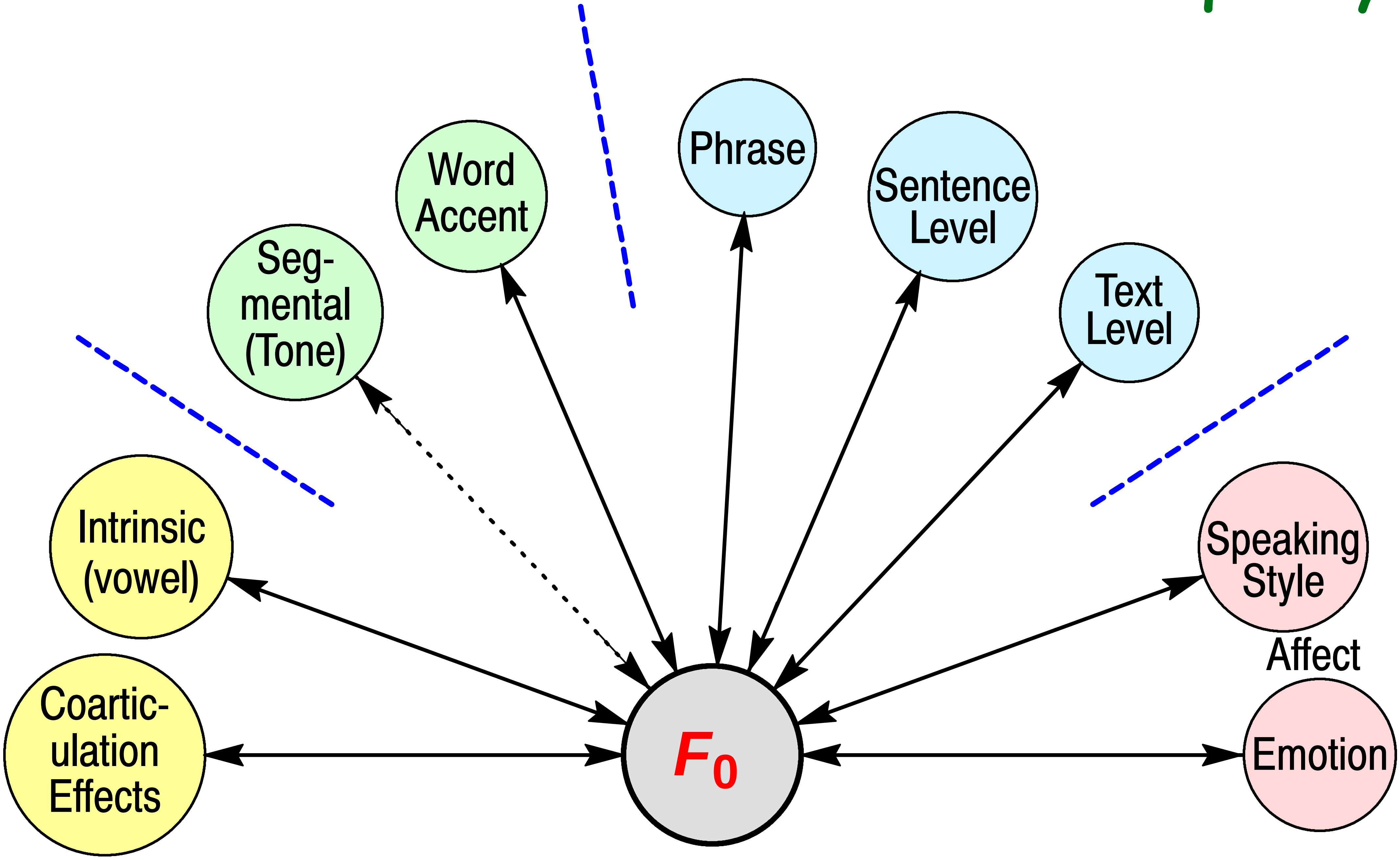
Duration

Energy

F_0

Effort & Voice Quality

The Functional Load on Fundamental Frequency



A Few Numbers

158 Contributed Papers

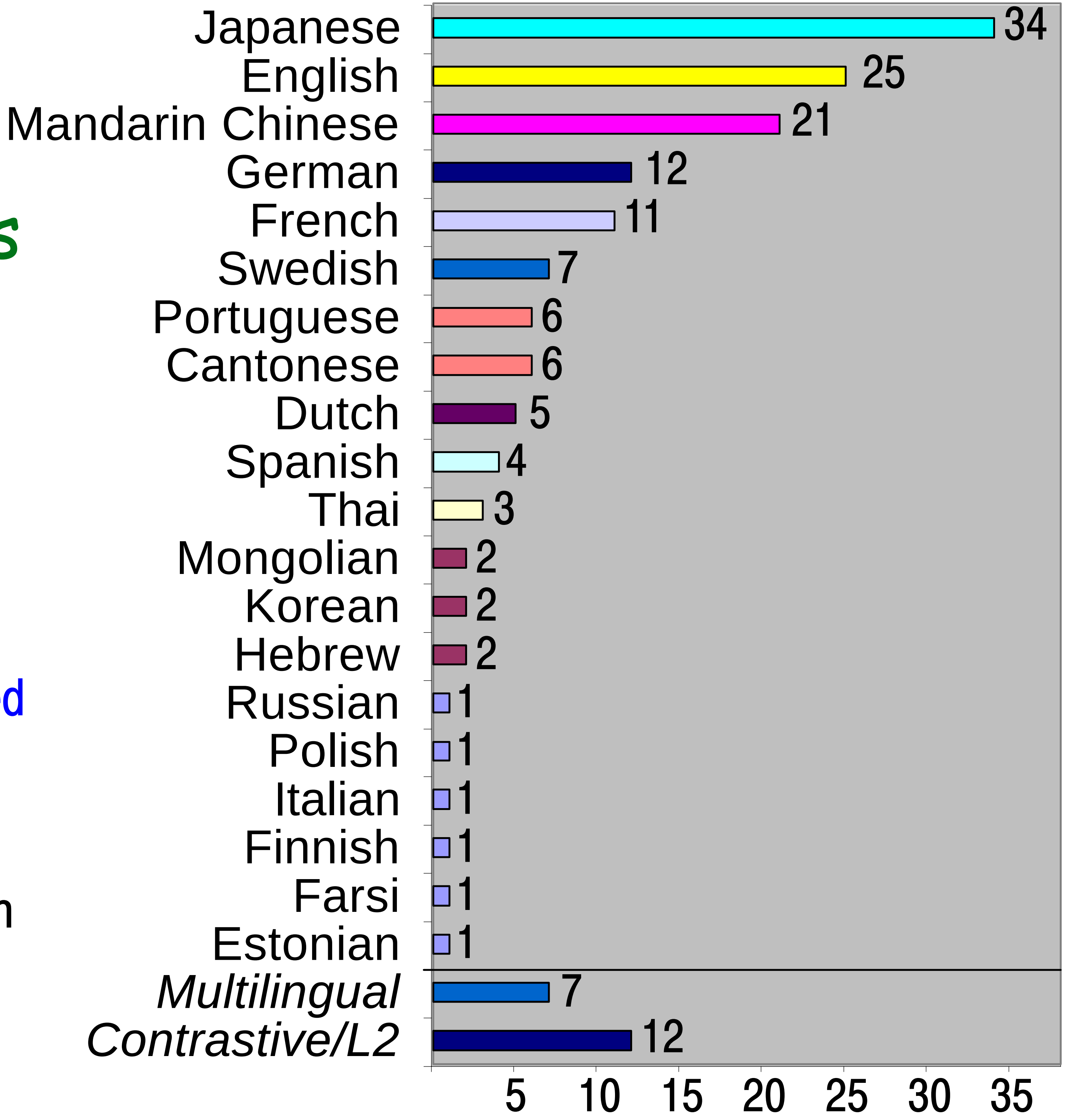
12 Invited Papers

20 Languages Investigated

270 participants

132 participants from Japan

29 countries represented



Outline of this Talk

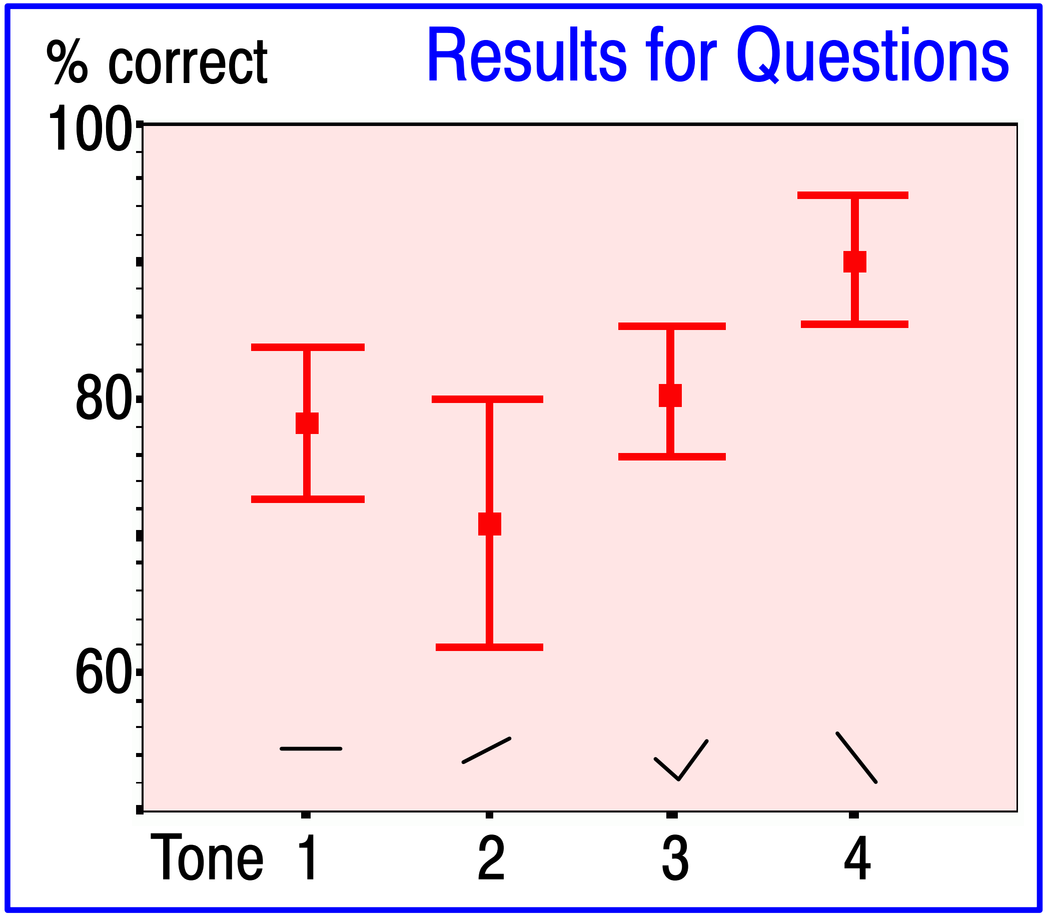
- General Topics and Numbers
- **A Glimpse at Individual Languages**
- Modelling
- Prosody and Music
- Measurement
- Prosody and Voice Quality
- Corpora and Speech Technology
- Some Open Questions to Take Home

Interaction between Tone and Intonation in Mandarin Chinese

YUAN/SHIH (131-134)

Parameters:

- Question vs. statement
- Tone of final syllable (1, 2, 3, 4)


Results for Questions

Example:

Li^{3}bai^{4}wu^3 Luo^{2}Yan^4 yao^4 mai^3 mao^1 [.?]
(Friday Lou Yan will buy cat [.?])

Experimental conditions:

- 8 speakers; 64 sentences each
- 16 listeners

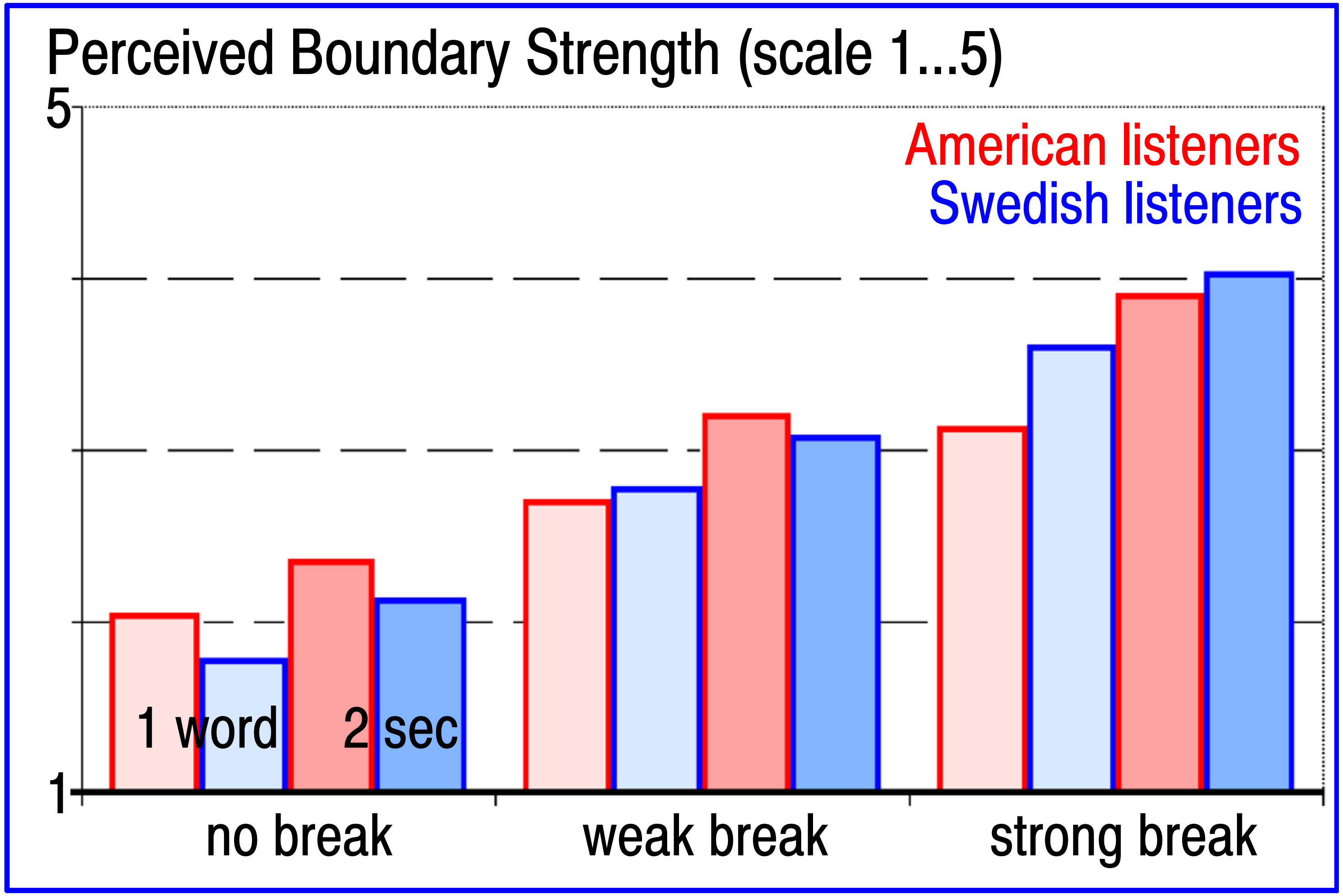
Results:

- Statements almost 100% correct
- Recognition rate for questions tone dependent

Can Boundaries be Judged from Prosodic Cues Alone? [Carlson/Hirschberg/Swerts, 329-332]

Stimuli

- From interview with a female politician; spontaneous speech
- fragments before “och”
- strong or weak or no boundary
- 2 sets differing in length: 1 word or 2 sec
- 2x60 stimuli in total
- 13 Swedish and 29 American listeners



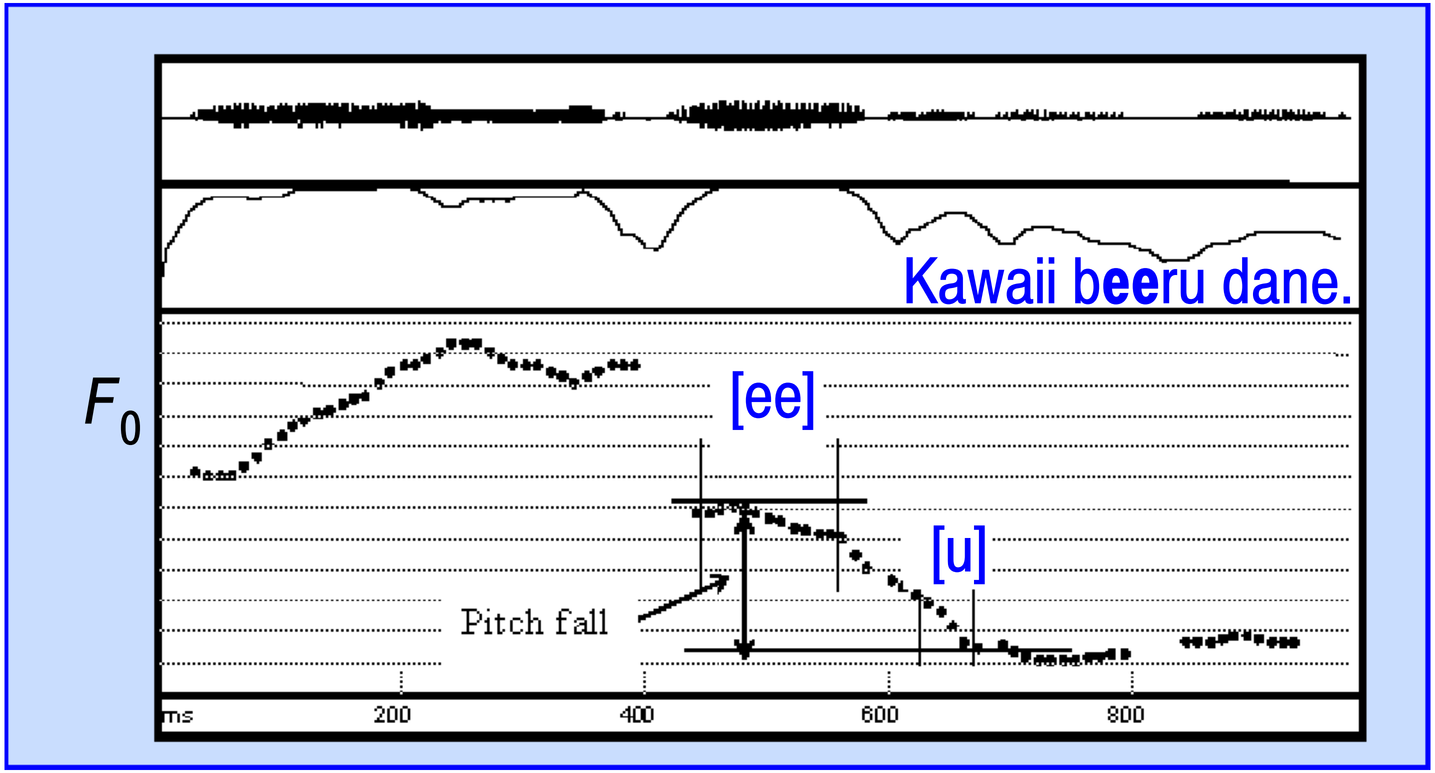
Creaky voice in the stimulus was found to be an especially salient cue in favor of boundary detection.

Pitch Accent and Quantity

Lehiste [11:14] showed that a three-mora vowel in Estonian (Q3) comes with a pitch fall most of the time.

A similar phenomenon is observed for Japanese where pitch accents in accented long vowels tend to be high on the first and low on the second half of the vowel (whereas unaccented long vowels are either high or low throughout the whole vowel) [KOZASA, 235:238].

If speech rate is varied, speakers seem to realize pitch fall on long accented vowels as a primary cue.



Some More Results

- In a tone language, the functional load of tone is as high as that of vowels [Surendran, Levow].
- Tones are much more difficult to identify in automatic speech recognition systems than segmental quality [Zhang et al.].
- Accent has a strong influences on acoustic cues to plosive voicing and place of articulation (VOT, F_0 at onset, closure duration) [Kim et al.].
- Perceived duration of high vowels is greater than for low ones, and perceived duration of a diphthong is greater than of a vowel-glide combination [Gussenhoven, Driessen].

Cross-Lingual Studies

- In audiovisual experiments (here: perception of emotion), non-native speakers tend to rely more on visual data than native speakers [Hebrew/Arabic; Amir, Almogi, Gal]
- In prosodically similar systems, boundary strength estimates can be derived from prosodic cues alone [Swedish/English; Carlson/Hirschberg/Swerts; cf. above]
- Segmental and prosodic features are about equally important in perceiving a foreign accent [Italian/Spanish, with French listeners; Boula de Mareüil et al.], cf. also [Kamiyama, on French/Japanese]

Outline of this Talk

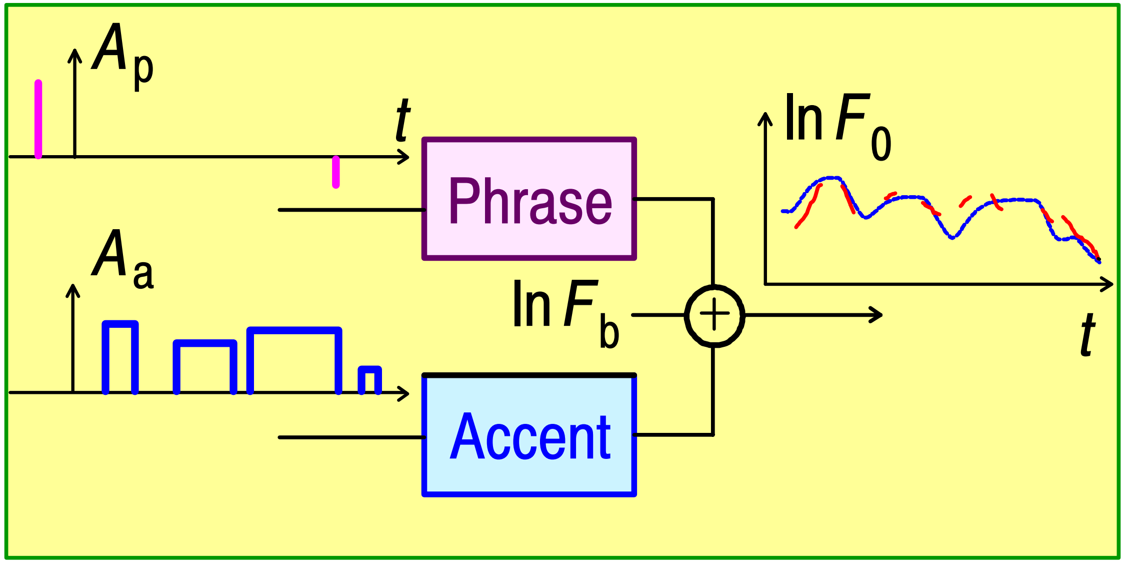
- General Topics and Numbers
- A Glimpse at Individual Languages
- **Modelling**
- Prosody and Music
- Measurement
- Prosody and Voice Quality
- Corpora and Speech Technology
- Some Open Questions to Take Home

FUJISAKI's Intonation Model

Definition (FUJISAKI, 1987:161)

$$C_\mathrm{p}(t) = \begin{cases} \alpha^2\, t\, e^{-\alpha t} & t \geq 0 \\ 0 & t < 0 \end{cases}$$

$$C_\mathrm{a}(t) = \begin{cases} 1 - (1+\beta t)\, e^{-\beta t} & t \geq 0 \\ 0 & t < 0 \end{cases}$$



$$\ln F_0(t) = \ln F_\mathrm{b} + \sum_{k=1}^{K} A_{pk}\, C_\mathrm{p}(t - T_{pk}) + \sum_{n=1}^{N} A_{an}\left[C_\mathrm{a}(t - T_{1n}) - C_\mathrm{a}(t - T_{2n})\right]$$

Base Value

Phrase Command

Accent Cmd on

Accent Cmd off

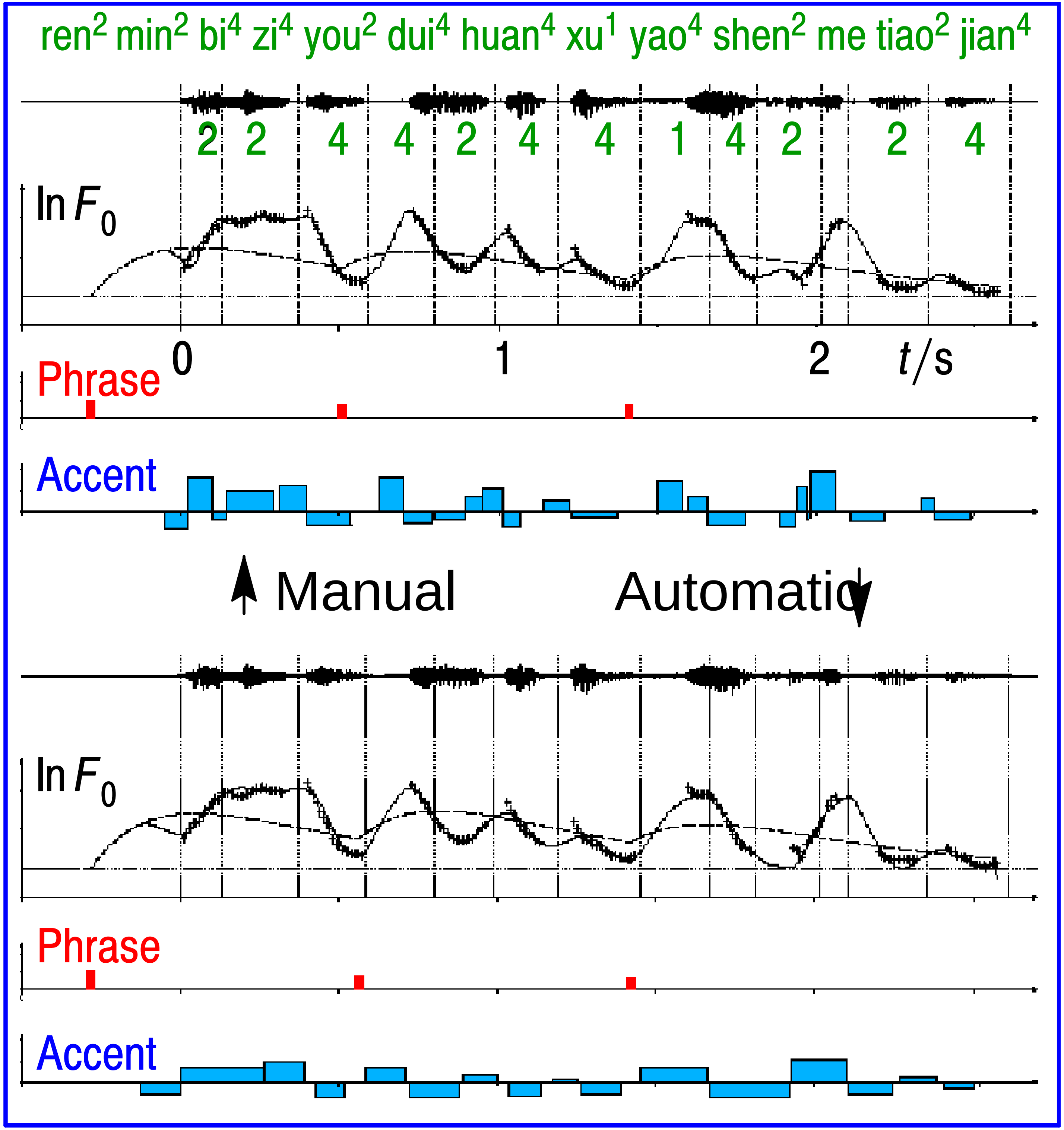
Investigations and Applications Using Fujisaki's Model

- Automatic determination of model parameters for Japanese [Bu, Yamamoto, Itahashi] / [Ogawa, Sagisaka]; evaluation of such algorithm [Narusawa, Minematsu, Hirose, Fujisaki]
- Emotional speech synthesis with corpus-based contour generation (Japanese) [Hirose, Sato, Minematsu]
- Concatenative tone model and parameter extraction (Mandarin Chinese) [G.P.Chen, Hu, Wu, Wang]
- Automatic tone command parameter extraction for Mandarin [Gu, Hirose, Fujisaki]
- Analysis and synthesis of F_0 contours for TTS in Spanish [Agüero, Wimmer, Belafonte]
- Prediction of accent commands (Portuguese) [Teixeira, Freitas, Fujisaki]
- Quantitative analysis of prosody in task-oriented dialogs (German) [Mixdorff]
- Pitch contour of Thai polysyllabic tone sequences [Seresangtakul, Takara]
- Comparing CART and Fujisaki model for synthesis of US-English names [Moberg, Pärssinen]
- Comparing two superpositional models [Raidt, Bailly, Holm, Mixdorff]
- Speech synthesis with attitude [Sagisaka, Yamashita, Kokenawa]

Tone Modelling in Mandarin Chinese Using Fujisaki's Model

[Gu/Hirose/Fujisaki, 435-438]

- Determine syllable boundaries
- Find template of accent command(s) for each syllable
- Assign amplitudes of accent commands
- Determine phrase commands
- Optimize



Outline of this Talk

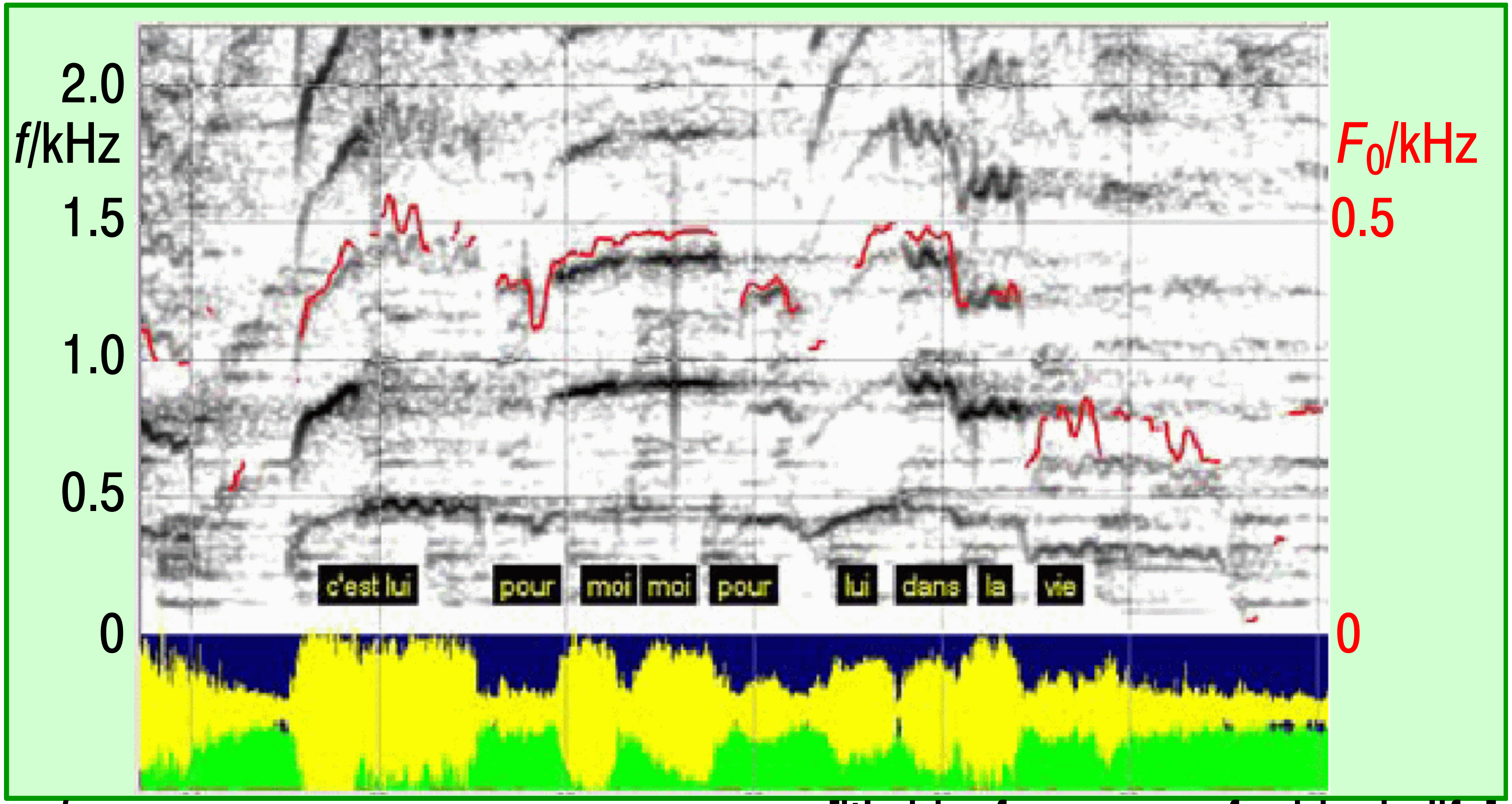
- General Topics and Numbers
- A Glimpse at Individual Languages
- Modelling
- **Prosody and Music**
- Measurement
- Prosody and Voice Quality
- Corpora and Speech Technology
- Some Open Questions to Take Home

Strategies to Export Speech Prosody into Music [Martin, 155-159]

C'est	**lui**	pour	**moi**	moi pour	**lui**	dans la	**vie**	
	2		**1**		**2**		**0**	stress level [0 – highest]
	↘		↗		↗		→	melodic contour
	[vibrato]		[vibr]		[vibr]		[vibr]	[transfer into melody]

Prosodic Structure in Speech

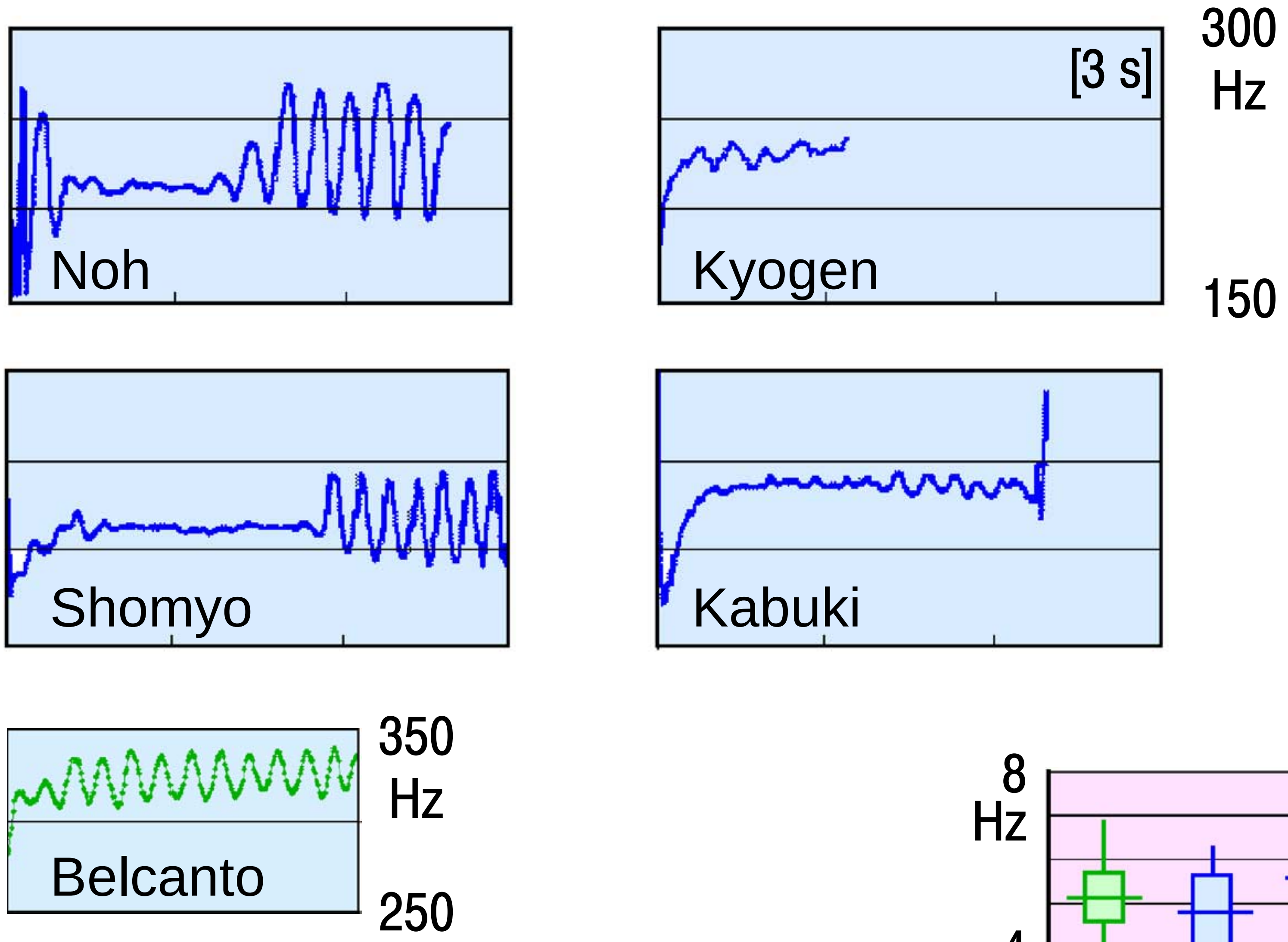
Musical Performance



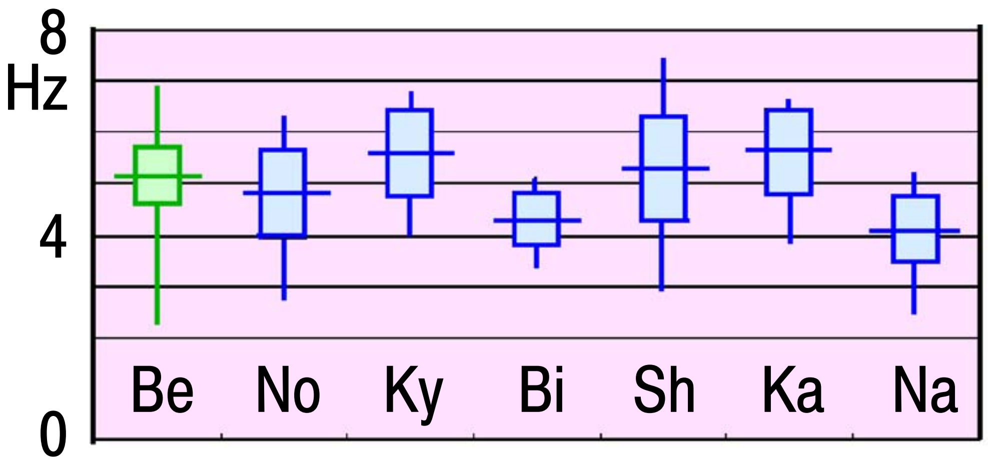
Edith Piaf, from *La vie en rose* (Text: Piaf, Music: Louigy, 1942) [it's him for me, me for him in life]

Vibrato in Different Singing Styles

Kojima/Yanagida/Nakayama (151-154)



Vibrato Patterns



Vibrato Frequency

Nagauta Synthesis

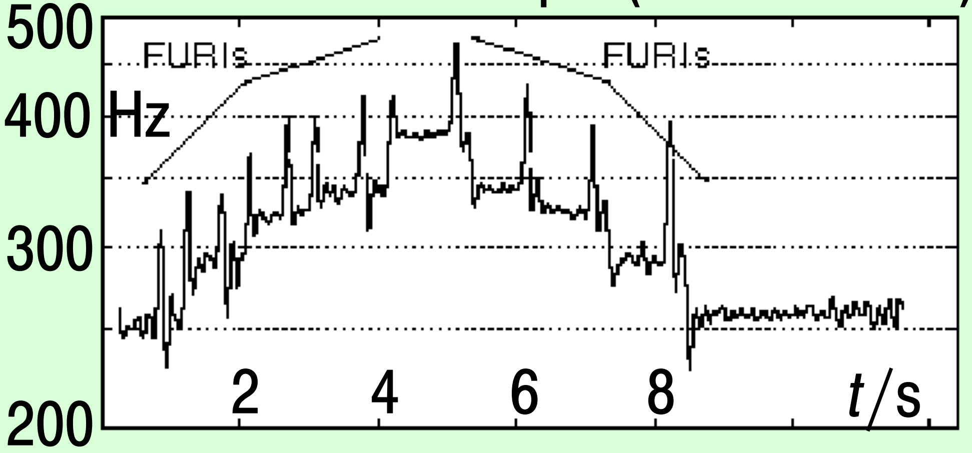
[MINEMATSU/MATSUOKA/HIROSE, 487-490]

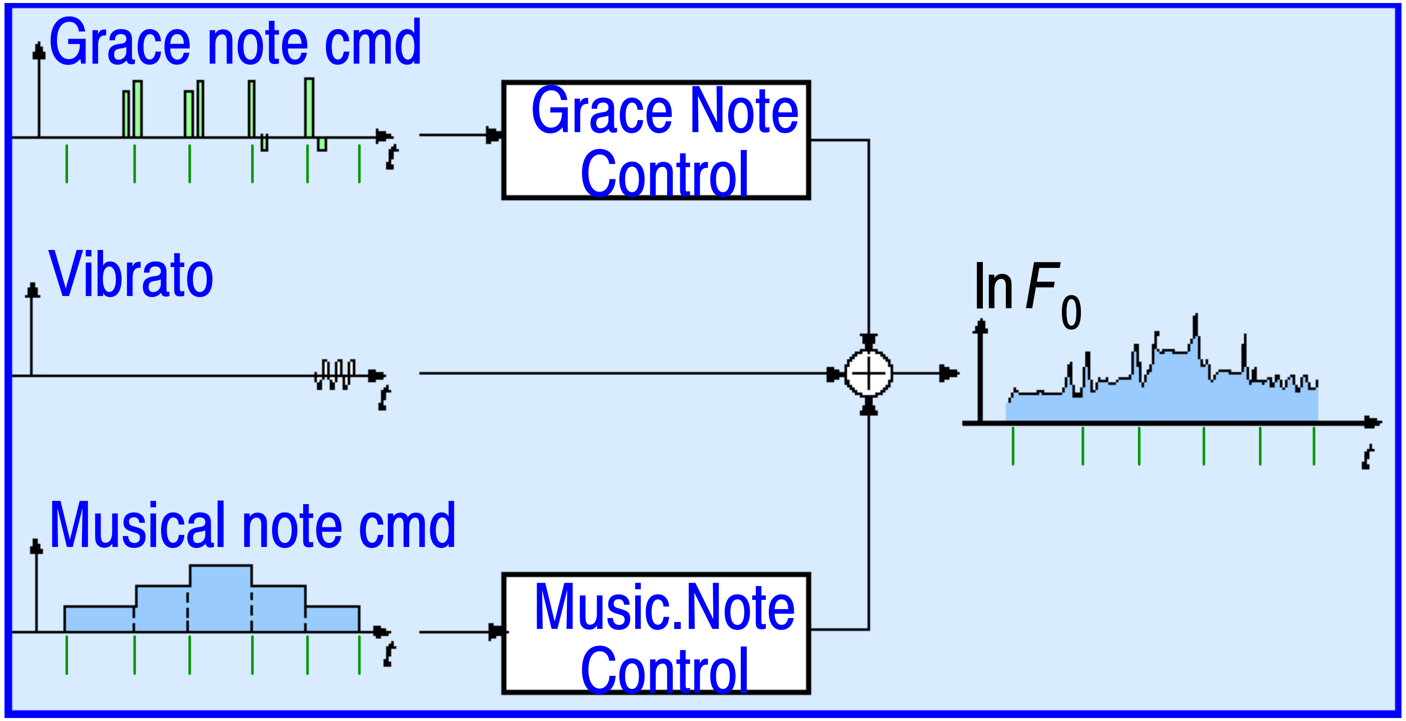
Nagauta means “long song”. Its tradition came up in the 17th century.

At transitions between notes sharp tone deviations occur (”grace notes”)

Originally pertained to a dance, but is today also performed without dance.

Example (diatonic scale)





Synthesis is done using a superposition model consisting of components for musical notes, grace notes, and vibrato.

Outline of this Talk

- General Topics and Numbers
- A Glimpse at Individual Languages
- Modelling
- Prosody and Music
- **Measurement**
- Prosody and Voice Quality
- Corpora and Speech Technology
- Some Open Questions to Take Home

Intonation Research and F_0 Measurement

- In intonation research we take for granted that F_0 measurement algorithms give reliable results.
- Nowadays they perform rather smoothly for “ordinary” tasks such as determination of F_0 contours from speech corpora.

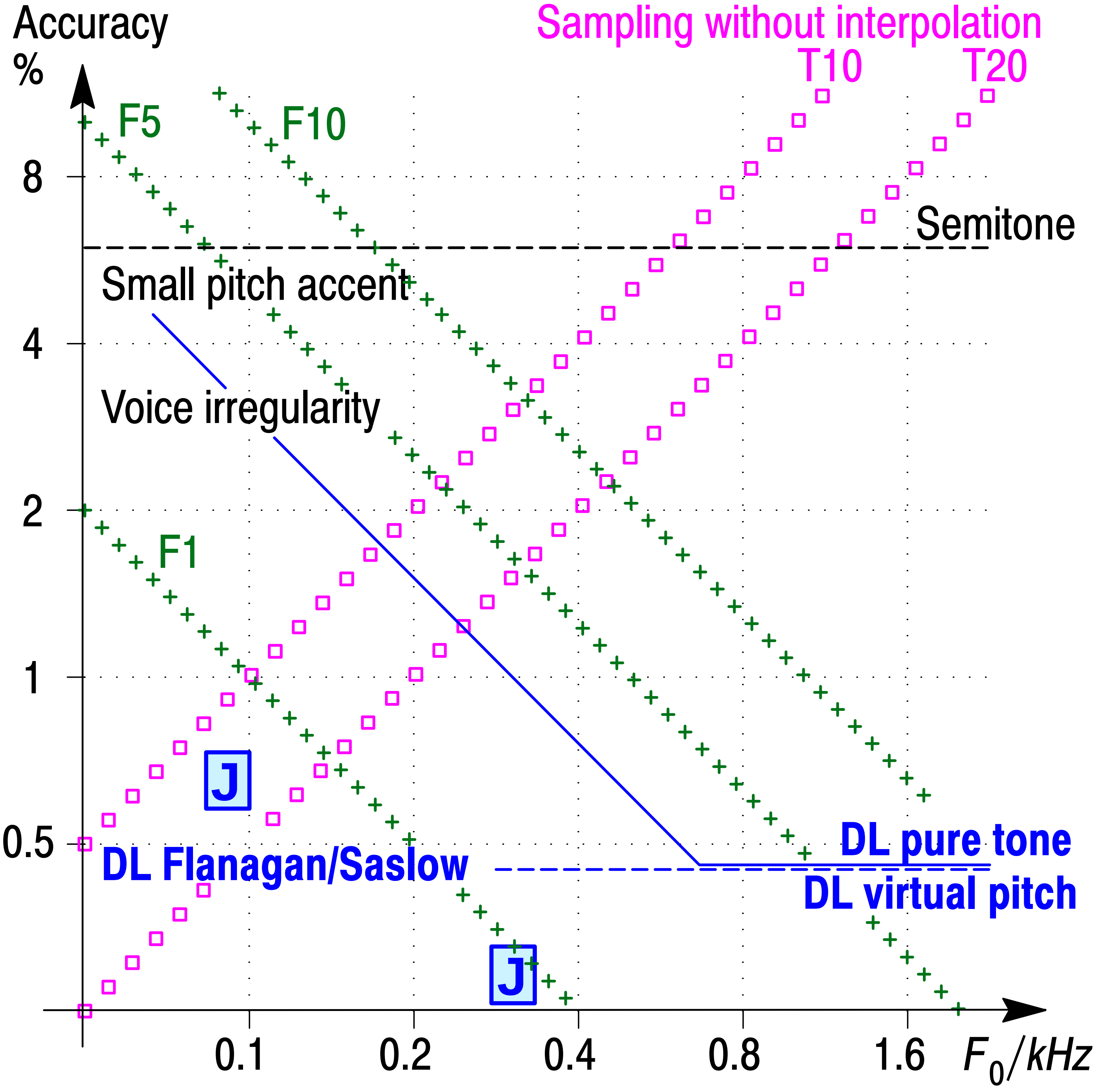
- However, there are special tasks such as analysis of vibrato, analysis of polyphonic music, or analysis of creaky voice.
- Some algorithms for such tasks were presented at this conference.

The Problem of Measurement Accuracy Regarding the Aspect of Perception

Results from the Literature (HESS, 1983)

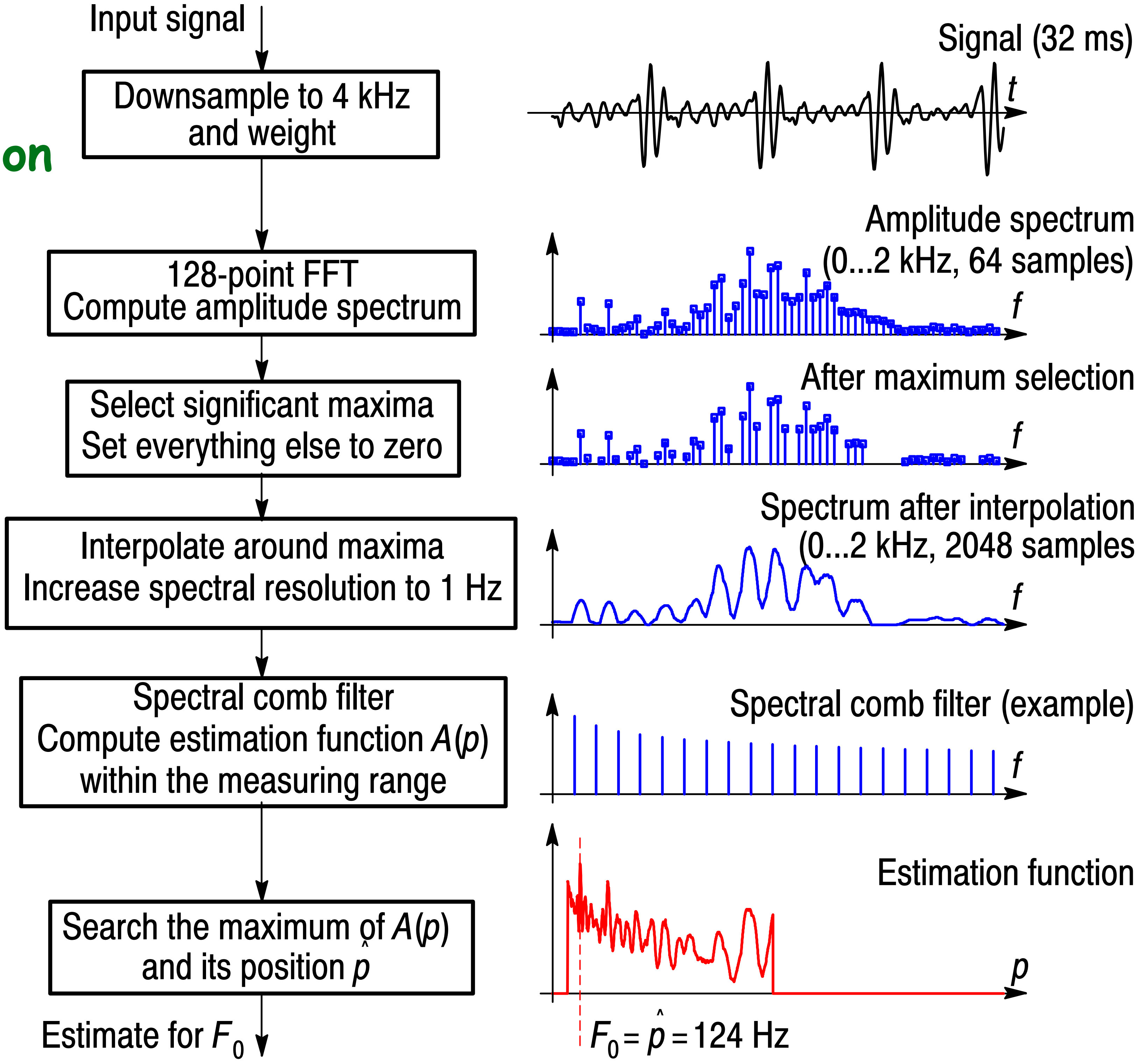
(DL) difference limen, just noticeable difference

(J) Jitter



F_0 Determination by Harmonic Compression [Martin, 1981; 545-548]

In a later version (1987), Martin used a logarithmic frequency scale, and in doing so, reached an equal resolution with respect to musical intervals over the whole frequency range. These algorithms are reported here.



Input signal

Downsample to 4 kHz and weight

128-point FFT
Compute amplitude spectrum

Select significant maxima
Set everything else to zero

Interpolate around maxima
Increase spectral resolution to 1 Hz

Spectral comb filter
Compute estimation function $A(p)$ within the measuring range

Search the maximum of $A(p)$ and its position $\hat{p}$

Estimate for F_0

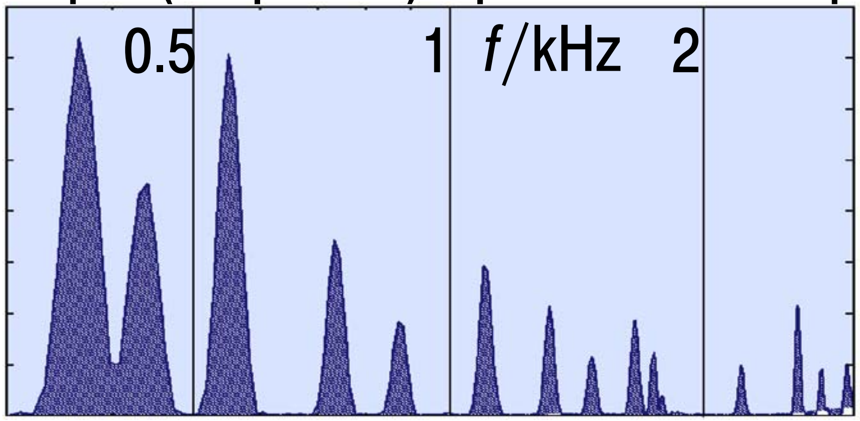
Pitch Determination for Polyphonic Music

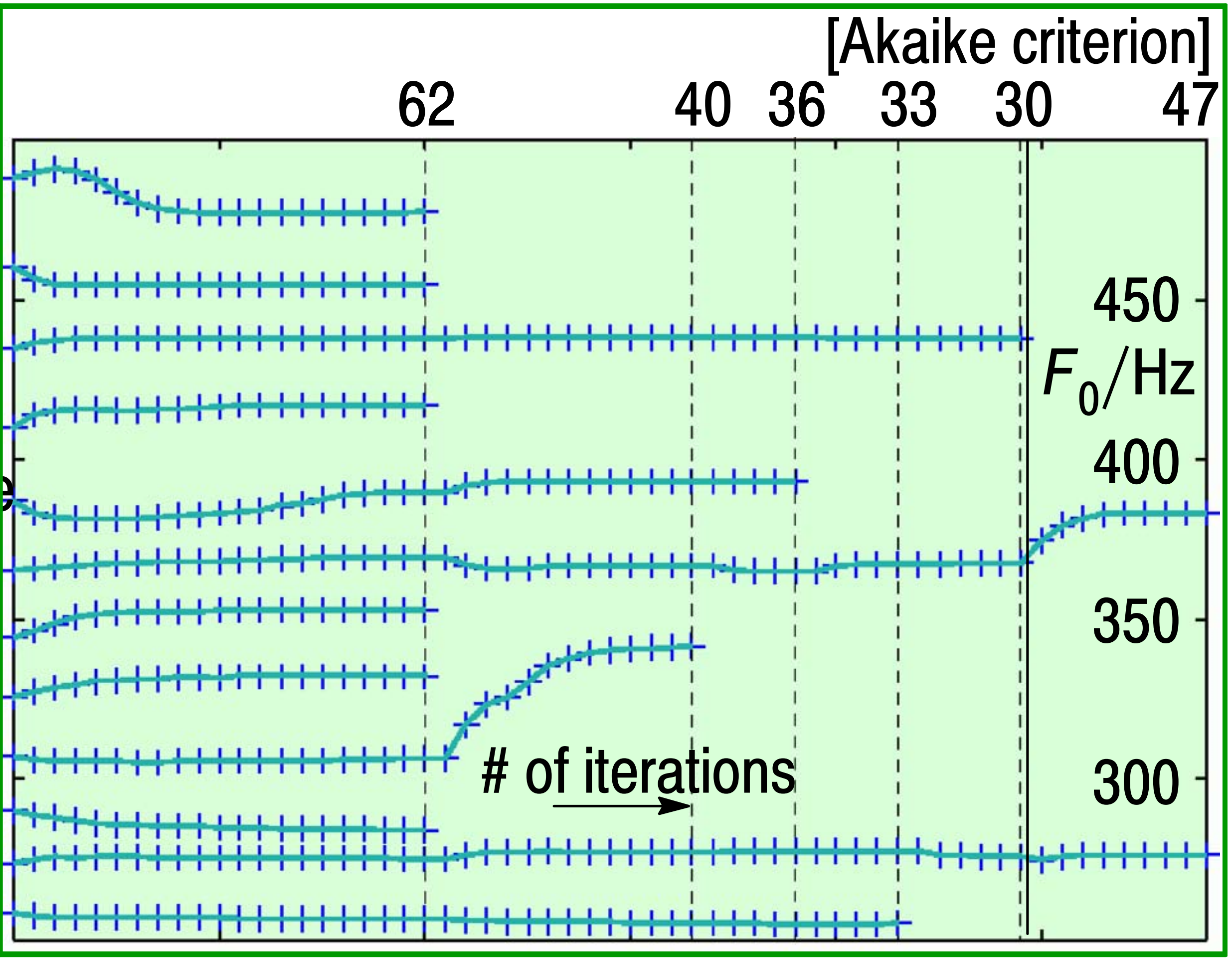
[Kameoka/Nishimoto/Sagayama, 533-536]

Iterative Expectation Maximization Algorithm

- Harmonic structure modelled as a mixture of tied Gaussian densities whose maxima are centered over prospective harmonics
- Start with the maximum number of pitches to be expressed
- Perform the iteration
- Successively remove the mixture that contributes least according to Akaike's information criterion.
- Select that number of pitches giving a minimum in this criterion.

Input (Amplitude) Spectrum: Example

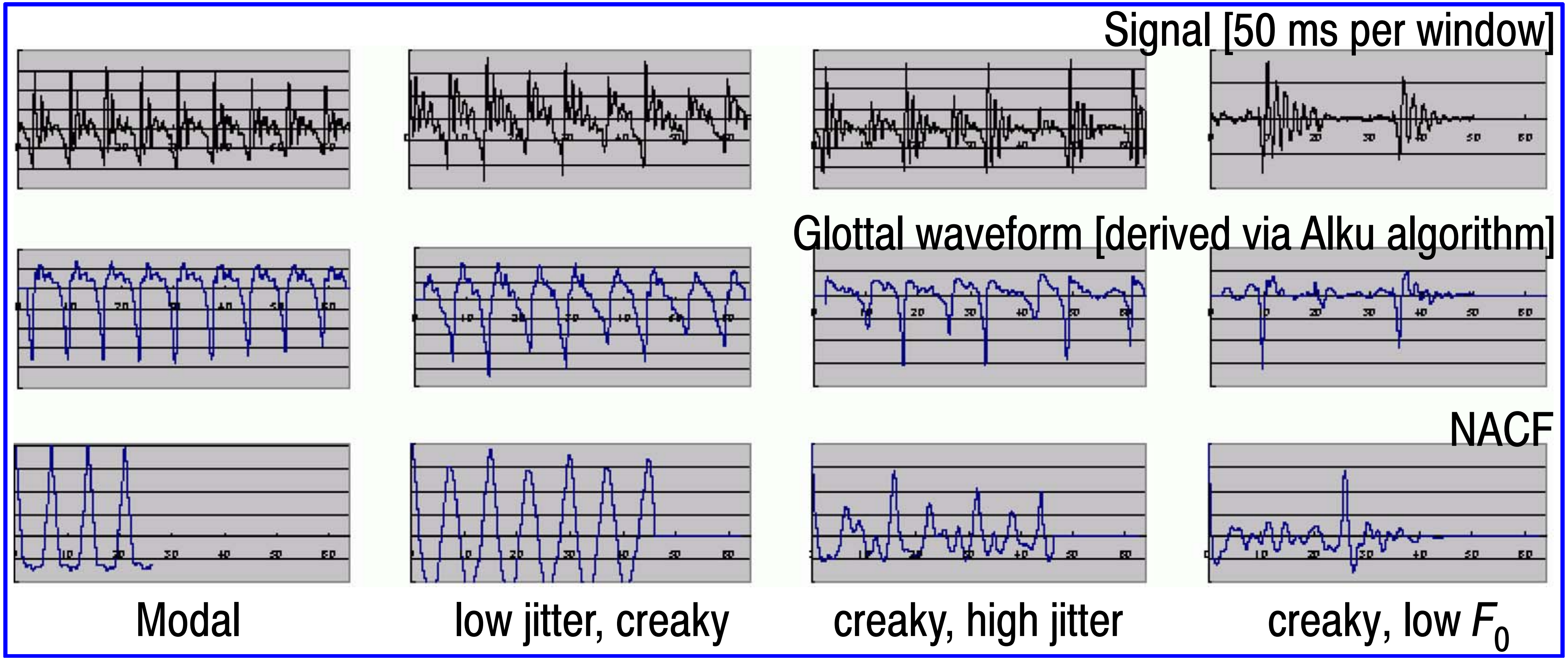




Analyzing Creaky Voice [Ishi, 643-646]

Employs a normalized autocorrelation function (NACF) derived from the reconstructed glottal waveform

- Creaky voice detected from 6 parameters derived from the first two peaks of the NACF
- 5619 frames from phrase-final utterance parts by 1 female speaker
- Decision tree on three parameters yields 91.5% correct



Signal [50 ms per window]

Glottal waveform [derived via Alku algorithm]

NACF

Modal | low jitter, creaky | creaky, high jitter | creaky, low F_0

Outline of this Talk

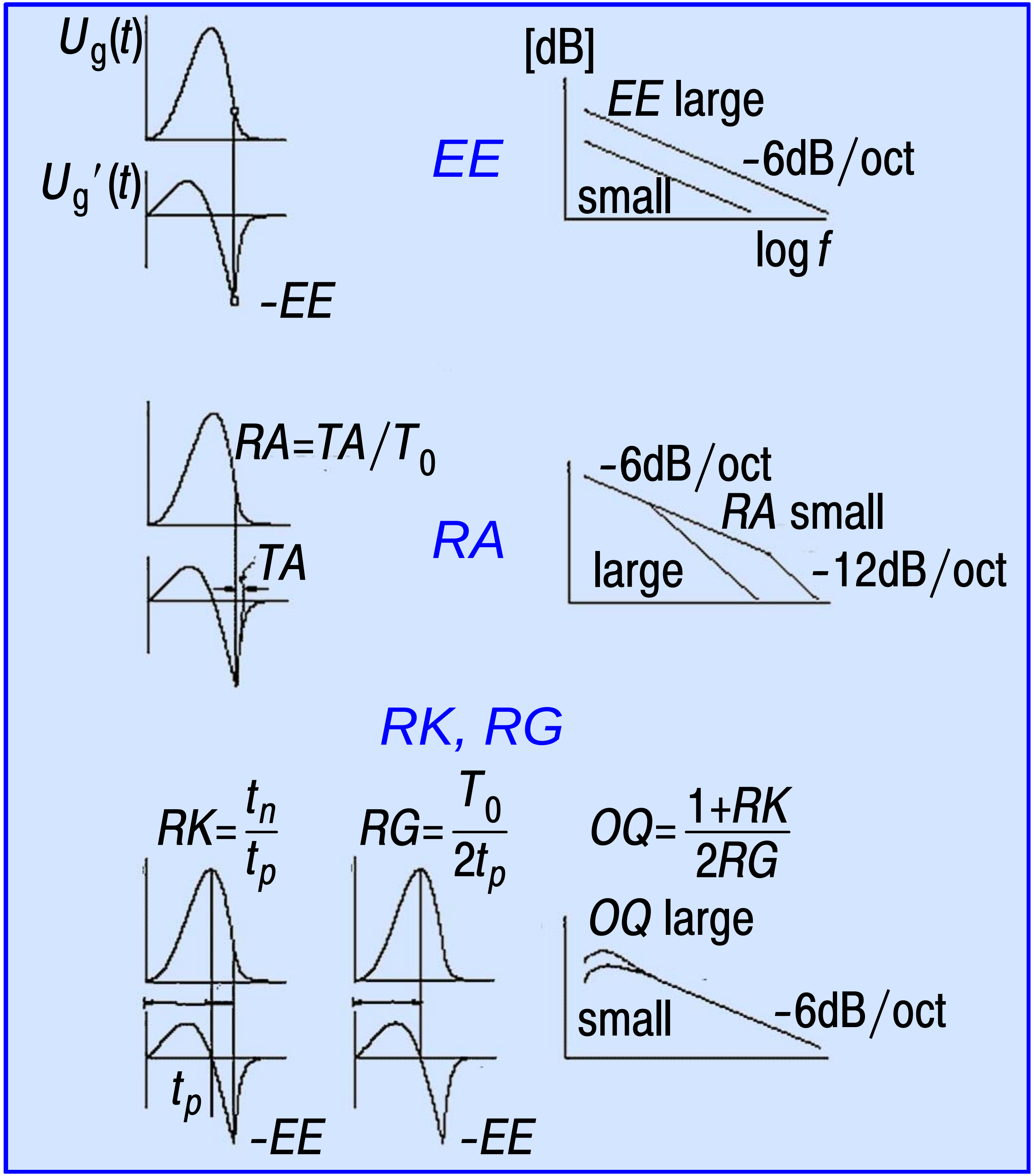
- General Topics and Numbers
- A Glimpse at Individual Languages
- Modelling
- Prosody and Music
- Measurement
- **Prosody and Voice Quality**
- Corpora and Speech Technology
- Some Open Questions to Take Home

Prosody and Voice Quality

[Ní Chasaide/Gobl, 189-196]

Parameters of the Liljencrants-Fant Model to describe voice source quality

EE	excitation strength
TA	duration of return phase
RA	normalized ...
RK	degree of skewing
OQ	open quotient

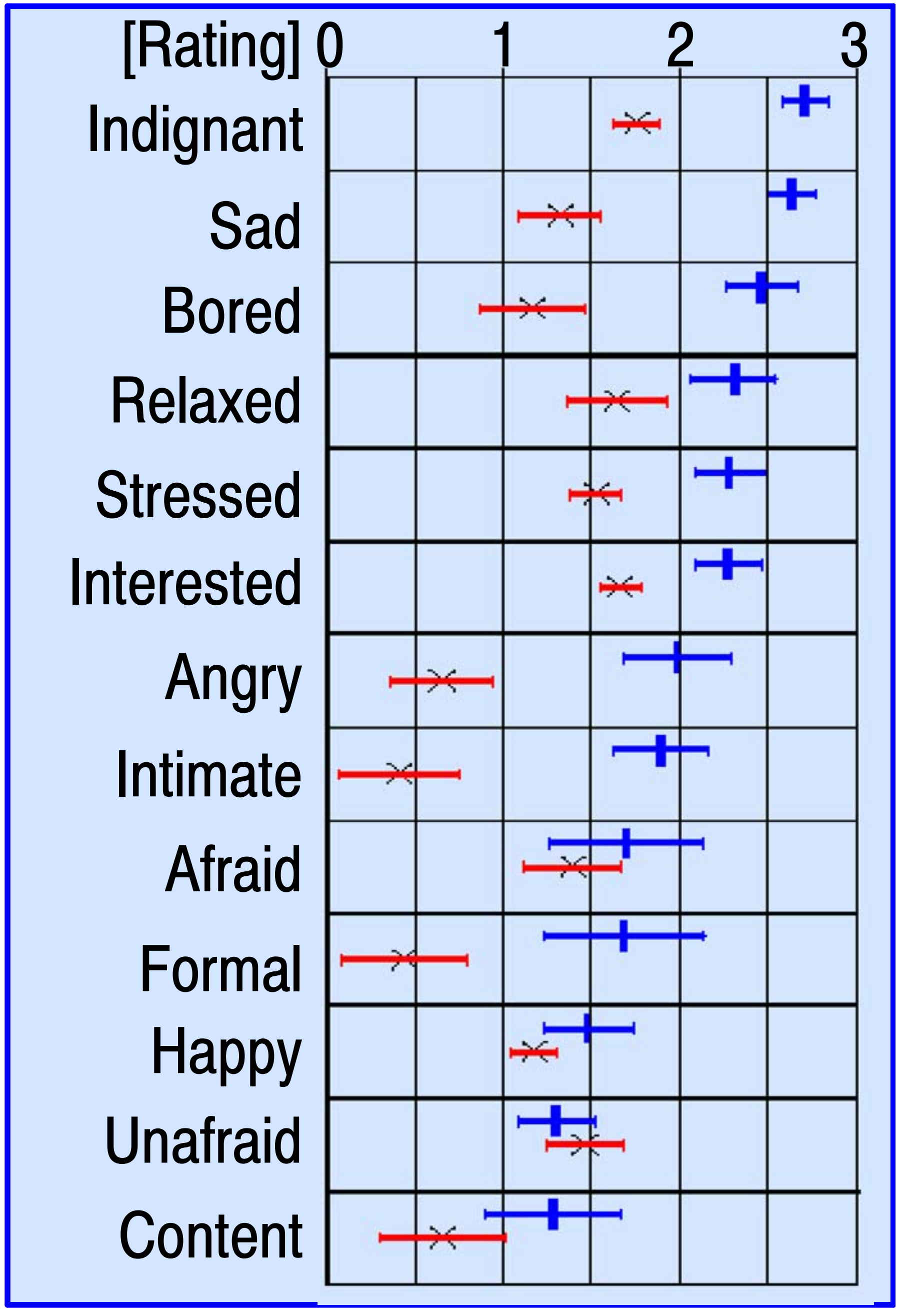


$$RK=\frac{t_n}{t_p} \qquad RG=\frac{T_0}{2t_p} \qquad OQ=\frac{1+RK}{2RG}$$

Voice Quality and Affective State

[Ní Chasaide/Gobl, 189-196]

- Any change of acoustic prosodic parameters also influences voice quality in a rather strong way.
- Voice quality and F0 together are much more powerful in conveying information about emotion and attitude than F0 alone.

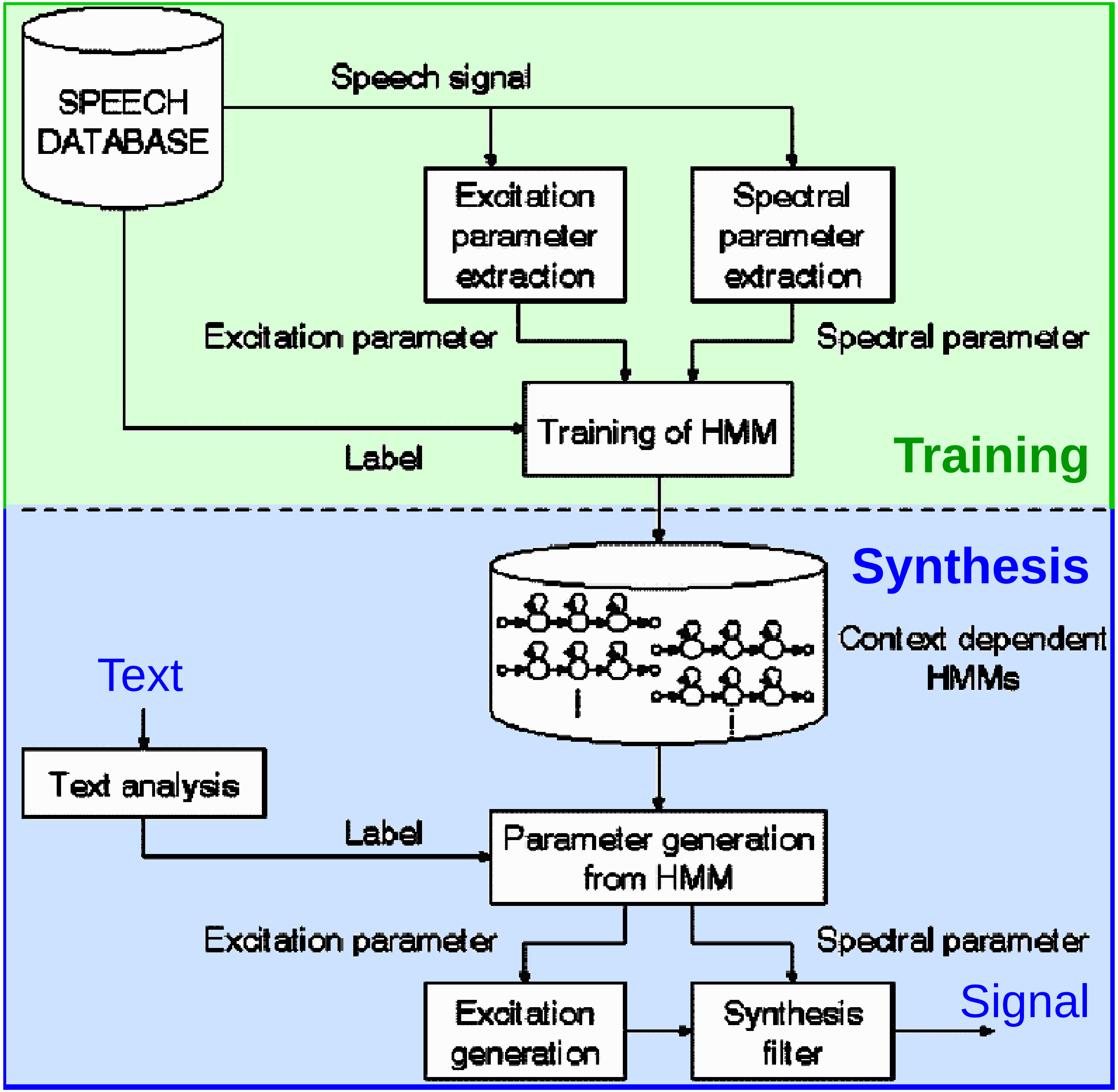


Voice Quality, Prosody, and Emotion

- Voice quality is a parameter that must not be neglected when trying to achieve naturalness in speech synthesis. It is even more important in expressive speech synthesis.
- A strong influence factor of emotion on voice quality is whether the emotion leads to increased or decreased activation level of the speaker [SCHERER].
- Para- and extralinguistic features tend to be gradual, whereas linguistic features are regarded as categorical.
- Voice quality, above all, is a multidimensional entity. It primarily affects the voice source, but vocal tract parameters are relevant as well.
- Voice quality is also a time-variant entity which is deliberately varied by the speaker during an utterance.

HMM Speech Synthesis [Tokuda, 2004] as a Tool

- Easy to use for experiments due to high flexibility and high adaptability
- Can be used for voice conversion, synthesis of expressive speech, etc.
- Current drawback: synthesis is parametric (vocoder)



Outline of this Talk

- General Topics and Numbers
- A Glimpse at Individual Languages
- Modelling
- Prosody and Music
- Measurement
- Prosody and Voice Quality
- **Corpora and Speech Technology**
- Some Open Questions to Take Home

“Evolutive” (Prosodic) Corpus: MARSEC [Auran/Bouzon/Hirst, 561-564]

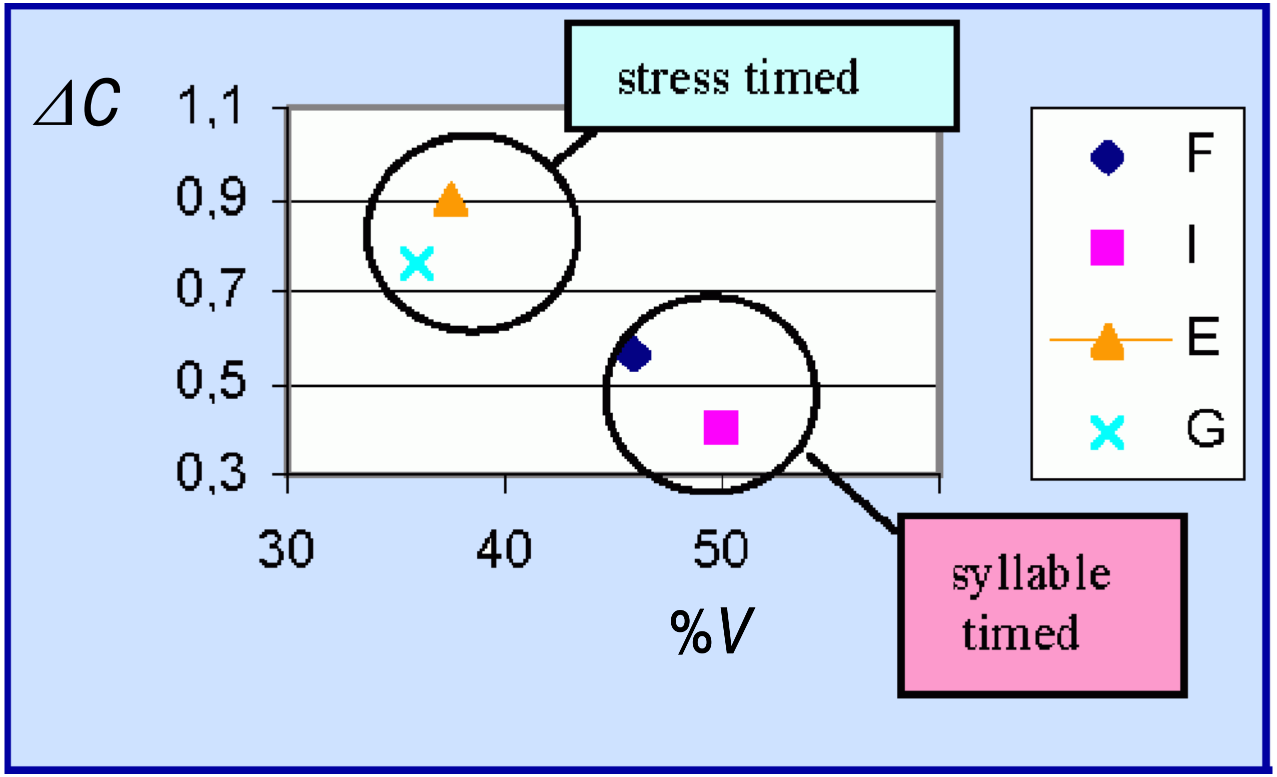
- MARSEC (MAchine-Readable Spoken English Corpus)
- Authentic speech: BBC recordings; 11 radio speech styles, mostly read; some interviews; 53 speakers; 55000 word tokens; >5 hours
- Word and intonation phrase boundaries manually aligned
- Manual and semiautomatic transcription using lexicon lookup and special rules for reduced forms and elisions
- Forced phonetic alignment by HMM segmentation with subsequent manual inspection (still under progress)
- Syllabification following maximum-onset principle and phonotactics; grouping into rhythmic units following both Abercrombie’s definition of stress foot and Jassem’s model of anacrusis and narrow rhythm unit
- Intonation coding by INTSINT and MOMEL
- To be further refined ...

The *Bonn Tempo Database* and Stress-Timed vs. Syllable-Timed Languages [Wagner/Dellwo, 227-230]

Database: short story with many city names, read in 4 languages (German, English, French, Italian) by several native speakers each, with five different speeds from *very slow* over *normal* to *as fast as possible*. The database has been labeled on the phone level.

Applications:

- Check *stress-timed* vs. *syllable-timed* language category using Ramus's measure
- Development of a new measure which is resistant to changes in articulation rate



Consistency of Annotation

[Gut/Bayerl, 565-568]

The LeaP [LEArn Prosody] Corpus: 12–hour corpus of German and English utterances by learners of these languages at different levels

- Corpus labeled on several tiers: phrase [speech vs. non-speech], word, syllable, segment [vocalic vs. consonantal regions], tone (ToBI), pitch [initial high, final low, in-between peaks and valleys]
- 6 annotators with different degrees of experience
- Test with a subset of the corpus
- Only fair agreements on the syllable and tone tiers
- Very high agreement for words ($\varkappa$=0.95), segments (0.99), and pitch (0.87); these tiers also had highest intra-annotator agreement; intermediate (0.73) for the phrase tier
- Almost perfect agreement between annotators is possible, but depends (1) on the complexity of the task and the units to be annotated, (2) on the experience of the annotator.

Outline of this Talk

- General Topics and Numbers
- A Glimpse at Individual Languages
- Modelling
- Prosody and Music
- Measurement
- Prosody and Voice Quality
- Corpora and Speech Technology
- **Some Open Questions to Take Home**

Some Open Questions to Take Home

- "There is very little empirical data about the voice source contribution to basic intonational elements. Even less is known about its role in perception." [A. Ní Chasaide, 192]
- "[...] we feel that a research priority is to find a way to analyse prosody in a way that encompasses both the linguistic and paralinguistic: to illuminate the latter we need to understand how those features that constitute the linguistic elements are modified to express affect." [A. Ní Chasaide, 194]
- The statement "although phonetics [...] is interested in all aspects of speech, the focus of phonetic notation is on the linguistically relevant aspects" [Handbook of the IPA] misses a fundamental goal of the discipline. It seems [...] that phonetics [...] should be extended so that it is capable of describing all aspects of information involved in spoken communication [Maekawa, 367].

Some Open Questions to Take Home [2]

- There seems to be quite a lot of convergence on some major issues (gradual vs. categorical coding of pragmatic information; role of affective paralinguistics) and some of the paradigms to study the issues (controlled experiments with systematic variation; synthesis to check hypotheses). [K. Scherer] - Yet further work is needed along this line.
- Linguistic information tends to be categorical, paralinguistic features tend to be continuous. How can we reconcile the two approaches?
- Voice quality has been shown to have a major influence in prosody in both the linguistic and para/extralinguistic domains. Integrated approaches must take account of this multidimensional parameter.
- A systematically constructed, multilingual and cross-cultural database for emotional speech is highly desirable [K. Scherer].

Last But Not Least

Thanks for a Wonderful Conference

- to the Scientific Committee for putting together this program
- to all the speakers and poster presenters
- and, last but not least, to the local organization for this magnificent site, the excellent organization, and all the hospitality

Dōmo arigatō gozaimashita!