Ludwig
Maximilians
Universität
München
LMU

IPS
INSTITUTE OF PHONETICS
AND SPEECH PROCESSING

Speaker Adaptation in Speech Synthesis

Uwe Reichel

Institute of Phonetics und Speech Processing

University of Munich

reichelu@phonetik.uni-muenchen.de

22nd October 2007

Contents

Definition and Motivation

Speaker Adaptation:

Transformation of symbolic and/ or signal aspects of a **source** utterance to derive a **target** utterance which differs from the source in terms of **speaking style** and/ or **speaker identity**

Motivation for speaking style modification:

- increasing variability and therefore also naturality of synthesised speech
- adapting synthesised speech to environmental needs (e.g. evoke hyperarticulation in noisy environments)
- evaluating influences of acoustic parameters on speaking style (by perception experiments with synthesised stimuli)

Motivation for speaker identity modification:

- commercially: enhance voice availability for e.g. navigation system customers
- evaluating influences of acoustic parameters on speaker identity (perception experiments)

Influence on Speaker and Speaking Style Characteristics

speaker-related influences:

- gender, age, body size, dialect, sociolect, health constitution, etc.

influences related to speaking style:

- occasion of the utterance, addressed hearer, emotion, importance of the conveyed message, etc.

Domains of Speaker adaptation in speech synthesis

Symbolic level

- word sequence (in **concept-to-speech**-synthesis)
- phoneme sequence
- prosodic structure: position and types of accents and phrase boundaries

Signal level

- f0 contour
- glottal excitation (voice quality)
- intensity
- vocal tract: formant frequencies, bandwidths, trajectories
- speech rate, segment duration
- most of these domains encode segmental as well as suprasegmental information

Example: Acoustics of emotions (excerpt of a collection by Schroeder, 2001)

Emotion	**Parameter Settings**
Joy	F0 mean: +50 %
	F0 range: +100 %
	speech rate: +30 %
	voice quality: modal or tense
Fear	F0 mean: +150 %
	F0 range: +20 %
	speech rate: +30 %
	voice quality: falsetto

Is the expression of fear an increased expression of joy?

Contents

- **adaptation**
  - symbolic level
    - – word sequence
    - – phoneme sequence
    - – prosodic structure
  - signal level
    - voice morphing
    - voice conversion
      - data lookup
      - mapping
      - modification
        - – glottis-filter decomposition
        - – glottal parameters
        - – vocal tract parameters
        - – glottis-filter re-composition

Contents

Contents

- adaptation
  - **symbolic level**
    - **– word sequence**
    - – phoneme sequence
    - – prosodic structure
  - signal level
    - voice morphing
    - voice conversion
      - data lookup
      - mapping
      - modification
        - – glottis-filter decomposition
        - – glottal parameters
        - – vocal tract parameters
        - – glottis-filter re-composition

Word sequence (not addressed yet)

- **Interlingua** (rule based)
  1. translation of the source word sequence into an abstract semantic (Interlingua) representation
  2. translation of this representation into the target word sequence
  - **example:** transformation into colloquial speaking style
    source: Frank trinkt drei Bier
    Interlingua: $\textsc{trinken}(\textsc{frank},\ \textsc{bier}) \wedge \textsc{anzahl}(\textsc{bier},\ 3)$
    target: Frank pfeift sich drei Bier rein
  - translation between source, Interlingua and target by means of **Categorial Grammar** (Steedman, 1998)

- **Statistical machine translation**
  1. **Training: Phrase alignment** of parallel texts in order to collect phrase co-occurrence probabilities. Further word sequence (**n-gram**) probabilities are collected.
  2. **Application:**
     – transformation of the source text S into a target text T that maximises $P(T|S)$
     – in general $P(T|S)$ cannot be estimated directly, since T and S are usually not entirely given as parallel texts in the training data. So T and S need to be decomposed, which can be achieved by re-formulation of $P(T|S)$ (**Bayes' rule**):

$$P(T|S) = \frac{P(S|T)P(T)}{P(S)}$$

     – $P(S|T)$ is called the **translation model**, and $P(T)$ is called the **language model** of T

Example:

- **Training:** phrase alignment in parallel texts; calculation of co-occurrence probabilities ($P(S|T)$) and word sequence probabilities ($P(T)$); here: *maximum likelihoods*.

Text A	Text B
Frank trinkt drei Bier	Frank pfeift sich drei Bier rein

$P(S|T)$: $P(\text{Frank trinkt drei Bier}|\text{Frank pfeift sich drei Bier rein}) = 1$
$P(T)$: $P(\text{pfeift}|\text{Frank}) = 1$, $P(\text{sich}|\text{pfeift}) = 1, \ldots$

- **Application:**

$$\begin{aligned} \hat{T} &= \arg\max_T \left[P(T|\text{Frank trinkt drei Bier})\right] \\ &= \arg\max_T \left[P(\text{Frank trinkt drei Bier}|T) \cdot P(T)\right] \\ &= \text{Frank pfeift sich drei Bier rein} \end{aligned}$$

Contents

- adaptation
  - **symbolic level**
    - – word sequence
    - **– phoneme sequence**
    - – prosodic structure
  - signal level
    - voice morphing
    - voice conversion
      - data lookup
      - mapping
      - modification
        - – glottis-filter decomposition
        - – glottal parameters
        - – vocal tract parameters
        - – glottis-filter re-composition

Phoneme sequence

- Speaking style or speaker dependent **grapheme-to-phoneme** conversion, or
- **phoneme-to-phoneme** conversion e.g. from canonical pronunciation to a dialectal variation
- **Rule-based** conversion (Kipp, 1999, including knowledge of phonotactics)
- **Statistic classifiers:**
  1. **Training:** Phoneme alignment of parallel pronunciation dictionaries; let some classifier (decision tree, neural net, etc.) learn the relations
  2. **Application:** transformation guided by co-occurrence knowledge learned in the training phase

Example:

$$\begin{array}{c|c|c|c} \text{k} & \text{a} & \text{l} & \text{t} \\ \text{k} & \text{OY} & \_ & \text{d} \end{array}$$

$\longrightarrow$ excerpt of derived co-occurrence knowledge in a 3-phoneme window:

k **a** l $\longrightarrow$ **OY**
a **l** t $\longrightarrow$ **_**
l **t** # $\longrightarrow$ **d**

Contents

- adaptation
  - **symbolic level**
    - word sequence
    - phoneme sequence
    - **prosodic structure**
  - signal level
    - voice morphing
    - voice conversion
      - data lookup
      - mapping
      - modification
        - glottis-filter decomposition
        - glottal parameters
        - vocal tract parameters
        - glottis-filter re-composition

Prosodic Structure

- **task:** sequence of syllables $\longrightarrow$ sequence of stress and boundary levels
- Text-based prediction of accent and phrase boundary location guided e.g. by:
  - syntax (e.g. Chomsky et al., 1968; Gee et al., 1983)
  - phonology (e.g. metrical phonology, Liberman, 1977)
  - semantics, statistical predictability (Bolinger, 1972; Pan et al., 2000)
  - information structure (focus–background, given–new, theme–rheme; Vallduví, 1993)
  - speaking style: hyperspeech connected with density of accents and phrase boundaries (Lindblom, 1990)
- **rule based prediction** (Van Deemter, 1998) or training of **statistical classifiers** (Veilleux, 1994)

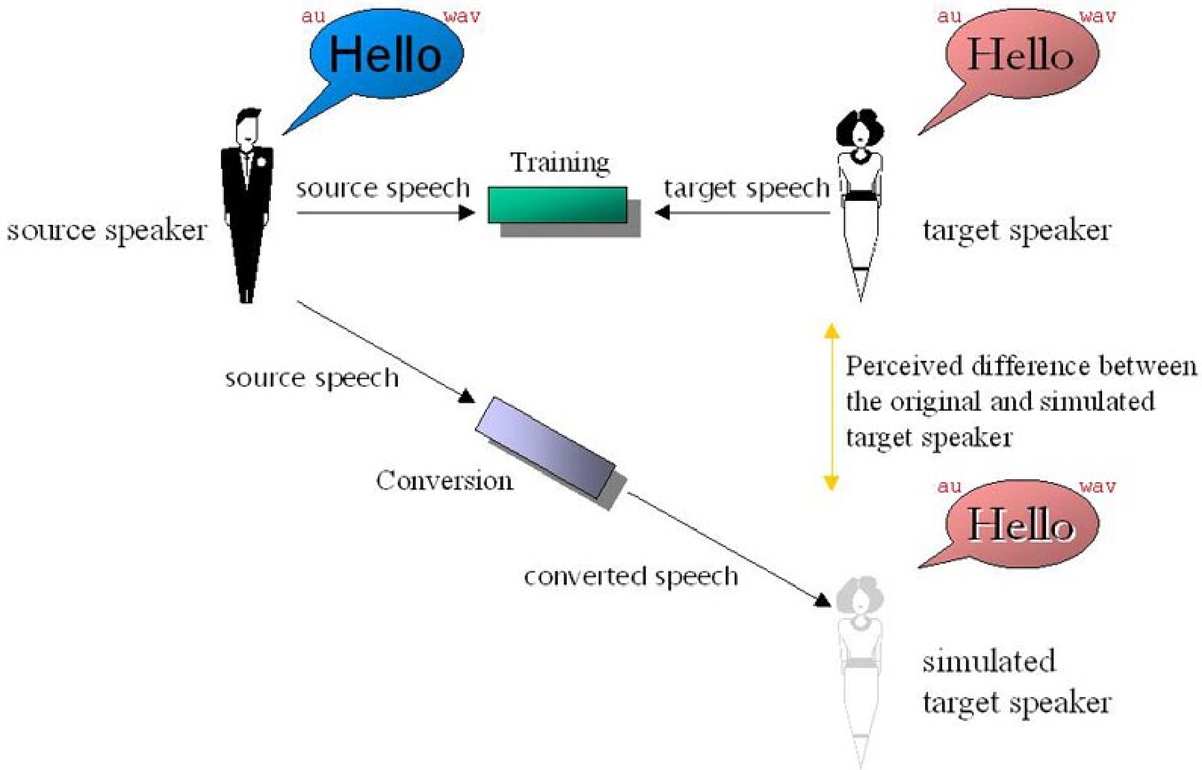
Contents

- adaptation
  - symbolic level
    - – word sequence
    - – phoneme sequence
    - – prosodic structure
  - **signal level**
    - **voice morphing**
    - **voice conversion**
      - **data lookup**
      - **mapping**
      - **modification**
        - – glottis-filter decomposition
        - – glottal parameters
        - – vocal tract parameters
        - – glottis-filter re-composition

Signal level

- **voice morphing:** continuous interpolation between two voices (e.g. Pfitzinger, 2004)
- **voice conversion:** changing a voice to a specified target
- **data lookup:** Selection of symbol and signal segments from huge labelled databases
- **mapping: replacement** of source entities by stored targets
- **modification: transformation** of source entities' features to target values

Voice Conversion



CSLU

Contents

- adaptation
  - symbolic level
    - – word sequence
    - – phoneme sequence
    - – prosodic structure
  - **signal level**
    - voice morphing
    - **voice conversion**
      - **data lookup**
      - mapping
      - modification
        - – glottis-filter decomposition
        - – glottal parameters
        - – vocal tract parameters
        - – glottis-filter re-composition

Data lookup

- Speech signal selection from huge databases (e.g. Campbell et al., 1997)
- **Advantage:**
  - no artefacts arising from signal processing
- **Disadvantages:**
  - expensive and time consuming effort to record and label data
  - much less generic than other approaches (e.g. add new emotion $\longrightarrow$ new recordings needed)
  - problem of real-time signal retrieval (huge search space)
  - black box: no phonetic knowledge acquisition

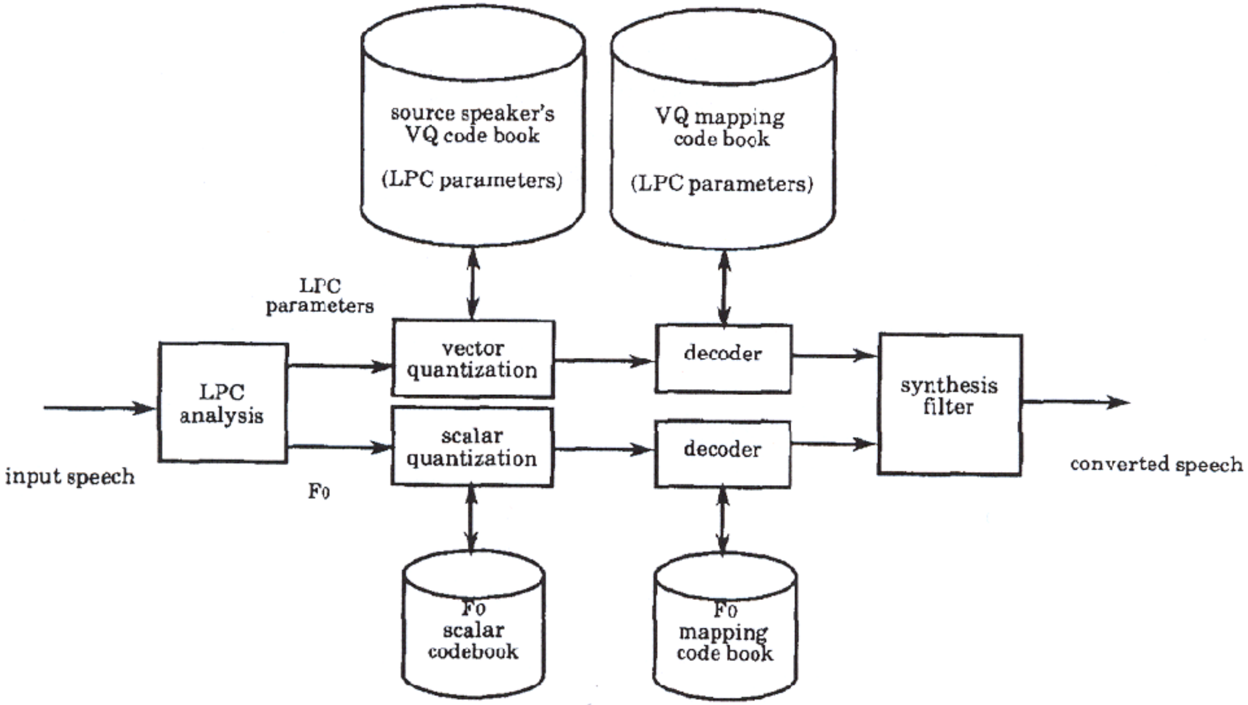
Contents

- adaptation
  - symbolic level
    - word sequence
    - phoneme sequence
    - prosodic structure
  - **signal level**
    - voice morphing
    - **voice conversion**
      - data lookup
      - **mapping**
      - modification
        - glottis-filter decomposition
        - glottal parameters
        - vocal tract parameters
        - glottis-filter re-composition

Mapping

- needed: a) an acoustic characteristics representation, b) a training corpus, and c) a mapping algorithm
- **Characteristics representation:**
  - segments (e.g. 20 ms frames) in the training data are represented as feature vectors
  - vectors contain e.g. f0, representation of glottal spectrum and transfer function of the vocal tract in form of Mel-Cepstral, DFT or LPC coefficients
- **Training corpus:**
  - contains signals of source and target voice
  - phonetically segmented and labelled
  - **vector quantisation** of the feature vectors in a smaller number of prototype vectors (e.g. centroids of derived vector partitions) a) to get reliable co-occurrence counts of source and target vectors, and b) to be able during application to assign new unseen vectors to existing (most similar) prototypes.

- **Code book mapping algorithm:** (e.g. Abe et al., 1990)



Kuwabara et al. (1995)

- **application task:** generate for each segment of the source voice an appropriate segment of the target voice, which is derived from the target voice database.
- Let S be the actual feature vector of the source voice to be mapped, which is assigned to the source prototype vector P_i^s. The corresponding target vector T is then calculated the following way:

$$T \quad = \quad \frac{\sum_j h_{ij} \cdot P_j^t}{\sum_j h_{ij}},$$

where h_{ij} is a weight reflecting the number of co-occurrence between source prototype vector P_i^s and target prototype vector P_j^t in the training data. Thus T is the normalised sum of all target prototype vectors in which the influence of each vector depends on its number of co-occurrence with P_i^s.

- mapping can be carried out independently for glottis and filter
- **Improvements:**
  - vector mapping at key points $+$ interpolation (reduces data sparseness problem)
  - spectral smoothing vs. discontinuities of target vectors chosen independently of each other
  - smoothing via context dependent mapping (e.g. use also the neighbouring source vectors of S, use T history)

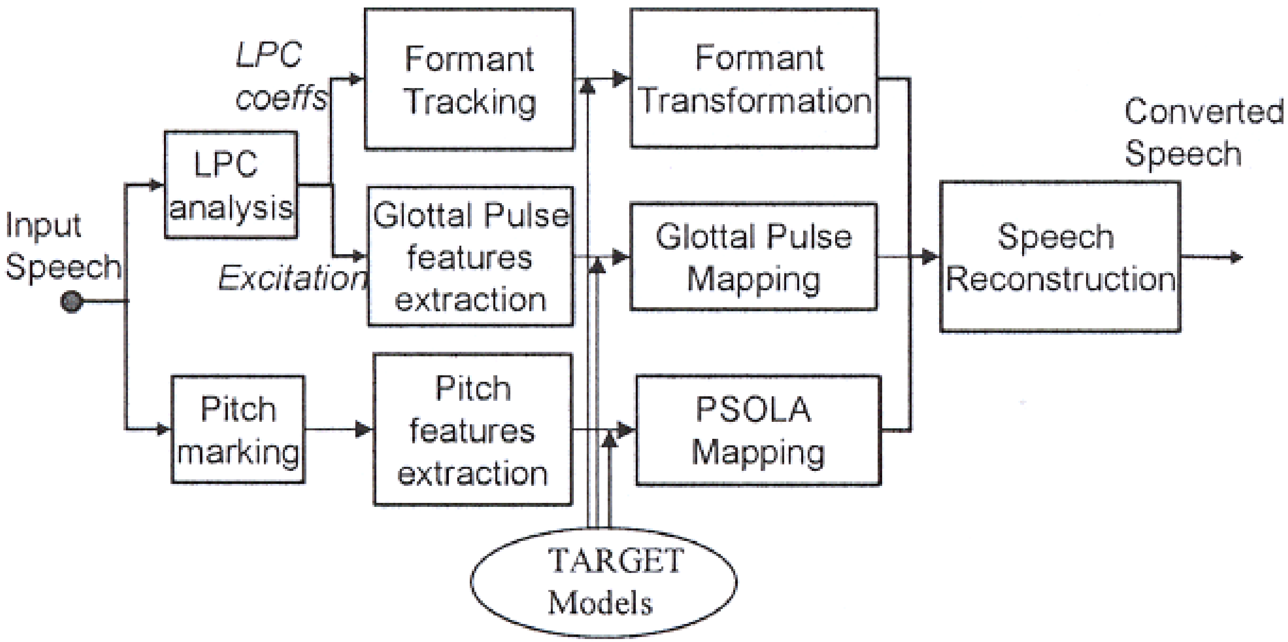
Contents

- adaptation
  - symbolic level
    - – word sequence
    - – phoneme sequence
    - – prosodic structure
  - **signal level**
    - voice morphing
    - **voice conversion**
      - data lookup
      - mapping
      - **modification**
        - glottis-filter decomposition
        - – glottal parameters
        - – vocal tract parameters
        - – glottis-filter re-composition

Modification

- e.g. Rentzos et. al. (2003)
- **Advantages:**
  - work on small databases $\longrightarrow$ fast data acquisition, low footprint applications
  - highly generic
  - acquisition and evaluation of phonetic knowledge
- **Disadvantages:**
  - artefacts arising from signal processing
  - so far less natural than previous approaches

- **Procedure:**

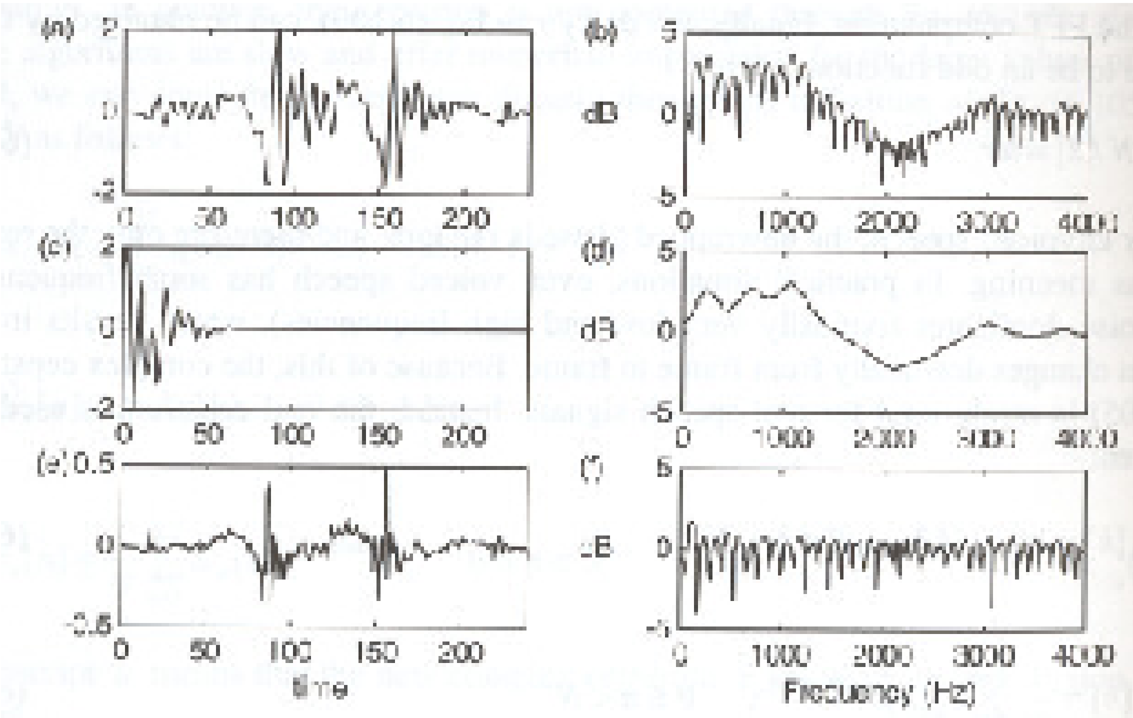


Rentzos et al. (2003)

Contents

- adaptation
  - symbolic level
    - – word sequence
    - – phoneme sequence
    - – prosodic structure
  - **signal level**
    - voice morphing
    - **voice conversion**
      - data lookup
      - mapping
      - **modification**
        - **– glottis-filter decomposition**
        - – glottal parameters
        - – vocal tract parameters
        - – glottis-filter re-composition

Excitation-filter decomposition



Huang et al. (2001)

Decomposing a speech signal (a) with spectrum (b) into vocal tracts impulse response (c, d) and glottal excitation (e, f) by **cepstral analysis** or **linear prediction**. Not needed for prosody modification with TD-PSOLA (see below).

Cepstral Analysis

- DFT of a time signal $\longrightarrow$ spectrum[1]
- macrostructure of the envelope corresponds to filter characteristics, microstructure to the excitation
- reapply DFT on the spectrum treating the frequency axis as a time axis
- excitation found in high frequency components, filter characteristics in low frequency components
- low pass filtering to separate excitation and filter

[1] **log** spectrum to transform the multiplicative composition of excitation and filter into an **additive** one, needed by the subsequent steps.

Linear prediction LP

- the n-th sample in a sequence can be predicted by the p previous samples

$$\hat{s}[n] \quad = \quad \sum_{k=1}^{p} a_k s[n-k]$$

- the weights a_k are to be chosen in order to minimise the error $(=$ residual$)$ $e[n]$ between the real sample value $s[n]$ and the predicted value $\hat{s}[n]$

$$e[n] = \arg \min_{a_1 \ldots a_p} \left[s[n] - \sum_{k=1}^{p} a_k s[n-k] \right]$$

- by z-transform the filter transfer function is derived from the coefficients $a_1 \ldots a_p$. the glottal signal is derived from the residual $e[n]$

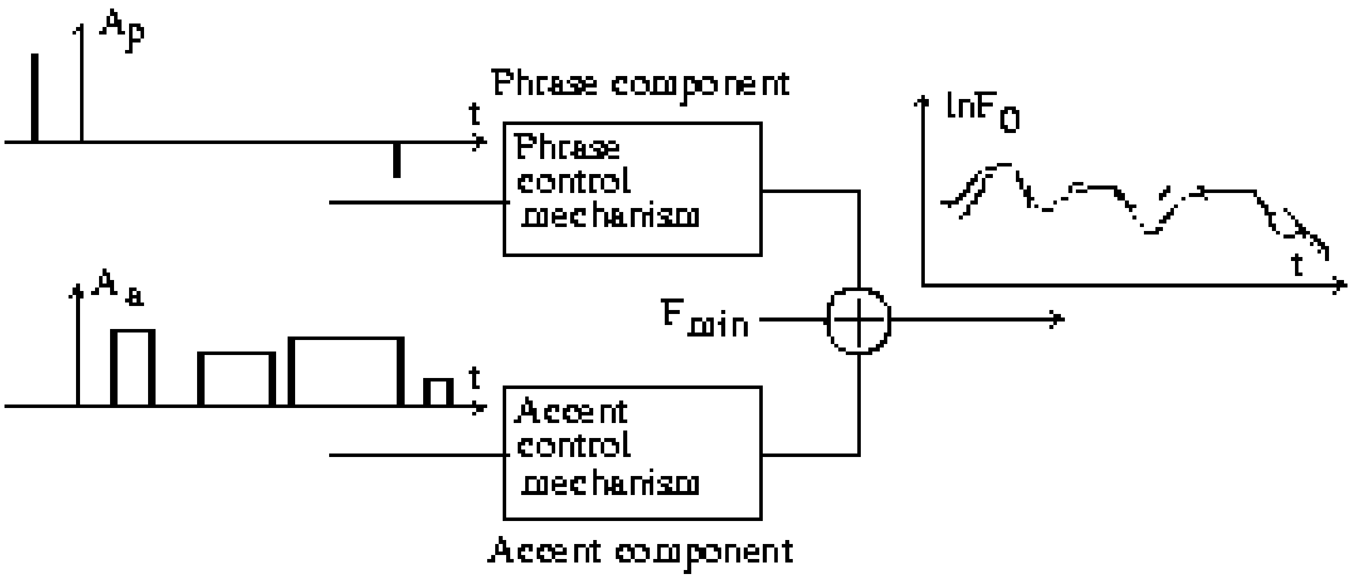
Contents

- adaptation
  - symbolic level
    - – word sequence
    - – phoneme sequence
    - – prosodic structure
  - **signal level**
    - voice morphing
    - **voice conversion**
      - data lookup
      - mapping
      - **modification**
        - – glottis-filter decomposition
        - **– glottal parameters**
        - – vocal tract parameters
        - – glottis-filter re-composition

Glottal parameters, Prosody

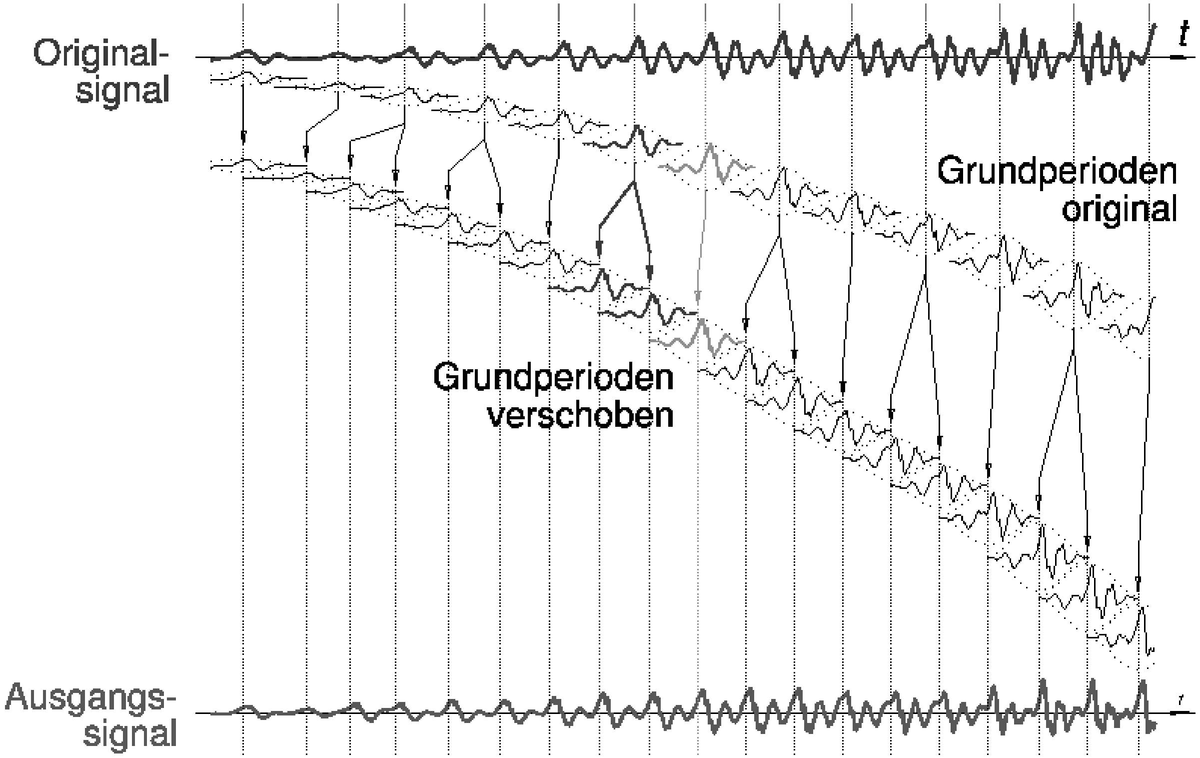
- pitch, duration, intensity, voice quality
- **Pitch measurement: e.g. by autocorrelation**
  - time domain algorithm, no need for source filter decomposition
  - signal is correlated with a version of itself, which is moved along the time axis
  - the correlation reaches its first maximum when the signal maximally ressembles its displaced version
  - this takes place as soon as the displaced version has been moved exactly 1 period T of the signal, which is $\frac{1}{f0}$
- simple: **Pitch rescaling:**
  - $f0_T = a + b \cdot f0_S$
  - moving f0 average and pitch span

- more elaborated: **Transforming prosodic structure to intonation**
  - Parameterisation of intonation e.g. by the **Fujisaki model** (Fujisaki, 1987)



- super-position of phrase component, accent component and baseline f0
- components $C_p(t)$ and $C_a(t)$ realised as critically damped systems (just positive oscillation values)
- systems are fed by phrase commando A_p (dirac impulse) and accent commando A_a (rectangle impulse) respectively.
- **phrase component:** global intonation contour of intonation phrase
- **accent component:** local f0 movements tied to accentuated syllables
- text and speaker based prediction of parameter values (Möbius, 1993)
- estimating parameter values for each intonation phrase by minimising the error between original contour and Fujisaki model output (**analysis by synthesis; but: no bi-uniqueness given**)

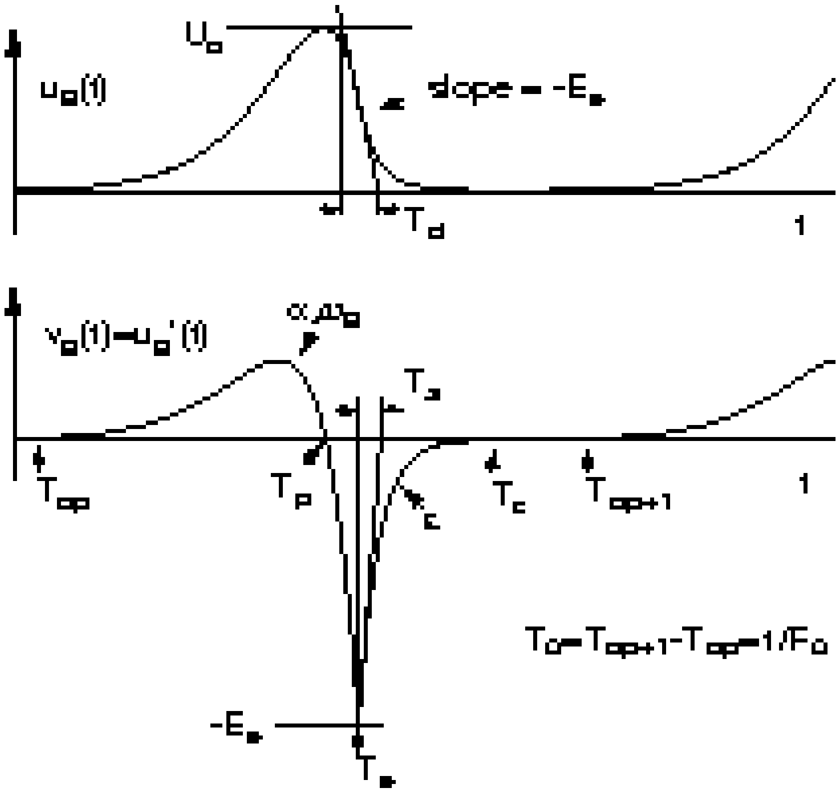
- **Applying the new pitch information; manipulation of pitch, duration and intensity (prosody): TD-PSOLA**
  - Moulines et al. (1990)
  - TD: manipulation in the **time domain**, no excitation-filter decomposition needed
  - PSOL: elementary building blocks are overlapping (OL) windows spanning about 2 f0 periods of the signal, and being centered on glottal pulses (PS: pitch synchronous)
  - A: manipulation by moving the windows and adding (A) the signals
  - **manipulating f0: increasing** by moving the windows closer to each other, **lowering** by moving the windows away from each other (+ replication or deletion of windows to preserve duration)
  - **manipulating duration:** replication of windows
  - **manipulating intensity:** sum copies of a window



Original-
signal
t
Grundperioden
original
Grundperioden
verschoben
Ausgangs-
signal

Hess (2004)

- **Manipulating not just pitch but also the glottal excitation pattern: Liljencrants-Fant parameterisation**



Iseli, et al. (2000)

– model of glottal flow $u_g(t)$ and its derivate $v_g(t)$ (representing flow changes)

LF Parameter	Description
T_{op}	instant of glottal opening
T_e	instant of maximum flow decrease (short before glottal closure)
T_p	instant of maximum glottal flow
T_a	effective duration of glottal flow decay
...	

– estimating parameter values for each glottal cycle by minimising the error between original excitation signal and LF modelled signal (**analysis by synthesis; but: no bi-uniqueness given**)

– **Relation between the parameters and voice quality:**

(Derived) Parameter	Calculation	Effect
Open Quotient	$\frac{T_e - T_{op}}{T_0}$	high $\longrightarrow$ breathy low $\longrightarrow$ creaky
T_a	$\frac{1}{\text{cut-off frequency}}$	spectral tilt shorter closing phase $\longrightarrow$ steeper upper falling flank of spectral envelope (word stress marker)

Contents

- adaptation
  - symbolic level
    - – word sequence
    - – phoneme sequence
    - – prosodic structure
  - **signal level**
    - voice morphing
    - **voice conversion**
      - data lookup
      - mapping
      - **modification**
        - – glottis-filter decomposition
        - – glottal parameters
        - **– vocal tract parameters**
        - – glottis-filter re-composition

Manipulation of vocal tract parameters

- long term average spectrum, spectral envelope, formant frequencies, formant trajectories, formant bandwidths
- LP coefficients can approximately be related to vocal tract geometry (sequence of log areas; Markel et al., 1976)
- global re-scaling of coefficients to simulate vocal tract shape
- local modifications to treat speaker dependent articulatory/ acoustic trajectories
- calculate coefficients connected to a desired vocal tract shape and movements $\longrightarrow$ new time varying filter transfer function (Childers, 1989)

Contents

- adaptation
  - symbolic level
    - – word sequence
    - – phoneme sequence
    - – prosodic structure
  - **signal level**
    - voice morphing
    - **voice conversion**
      - data lookup
      - mapping
      - **modification**
        - – glottis-filter decomposition
        - – glottal parameters
        - vocal tract parameters
        - **– glottis-filter re-composition**

Excitation-filter re-composition:

- **convolution** of an excitation signal ($\longleftarrow$ e.g. LF model) and a time varying filter ($\longleftarrow$ e.g. LPC coefficients)

Thank you for listening!