Unit selection-based TtS

Spyros Raptis

Institute for Language and Speech Processing / "Athena" Research Center

spy@ilsp.gr

Outline

- Concatenative unit selection: a basic description
- TtS in context – Other processes involved
- The Blizzard Challenge
- Tapping into expressiveness
- A peek into the hands-on/demo session

In defense of unit selection…

- It has large **footprints**:
  Efficient pruning and coding can conveniently fit it to your mobile phone
- It is **cheating**:
  Yes, but that's great! Especially for (limited?) domain dialogue systems.
- It is "**old tech**":
  It may very well be used to tackle new challenging problems (e.g. expressive); it can be hybridized; it has the best vocoder there is (no vocoder); and…

…it still is the best there is in terms of naturalness

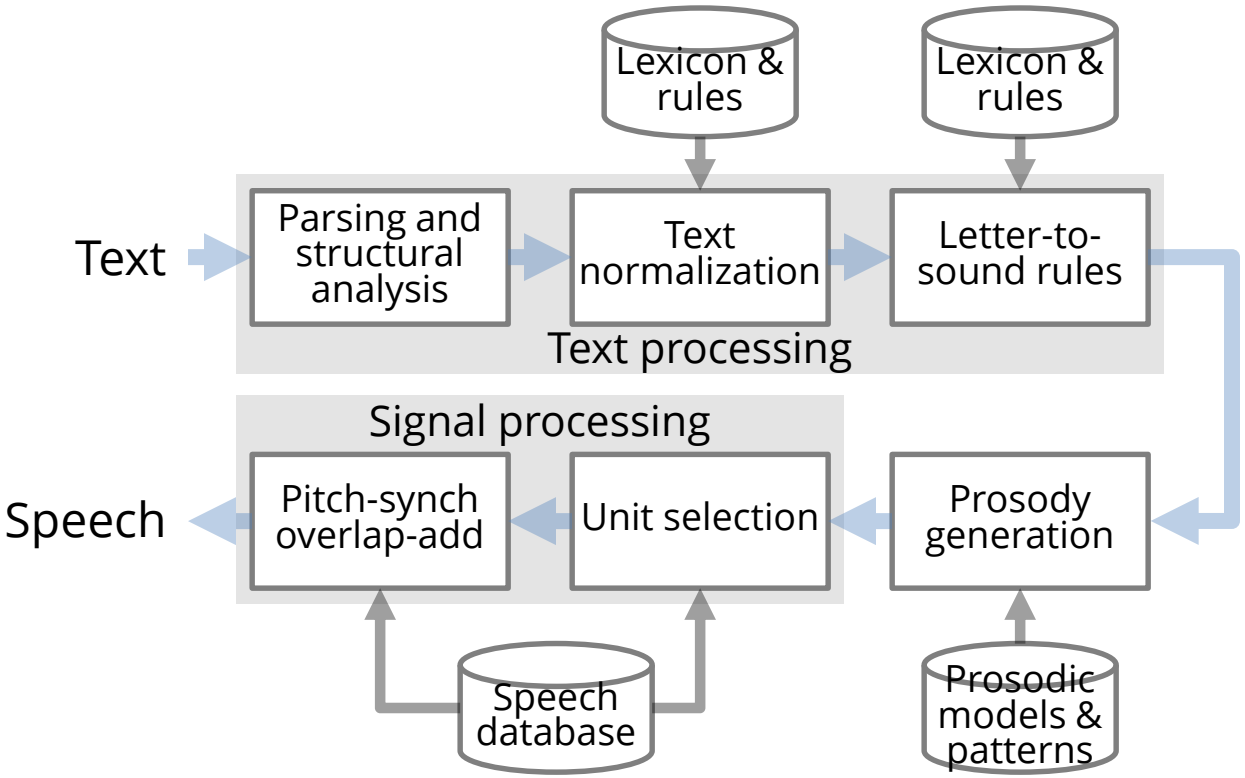
…it still has many challenges

"…and whilst appearing to be fairly simple, requires a great deal of engineering skill to obtain really good results" [*]

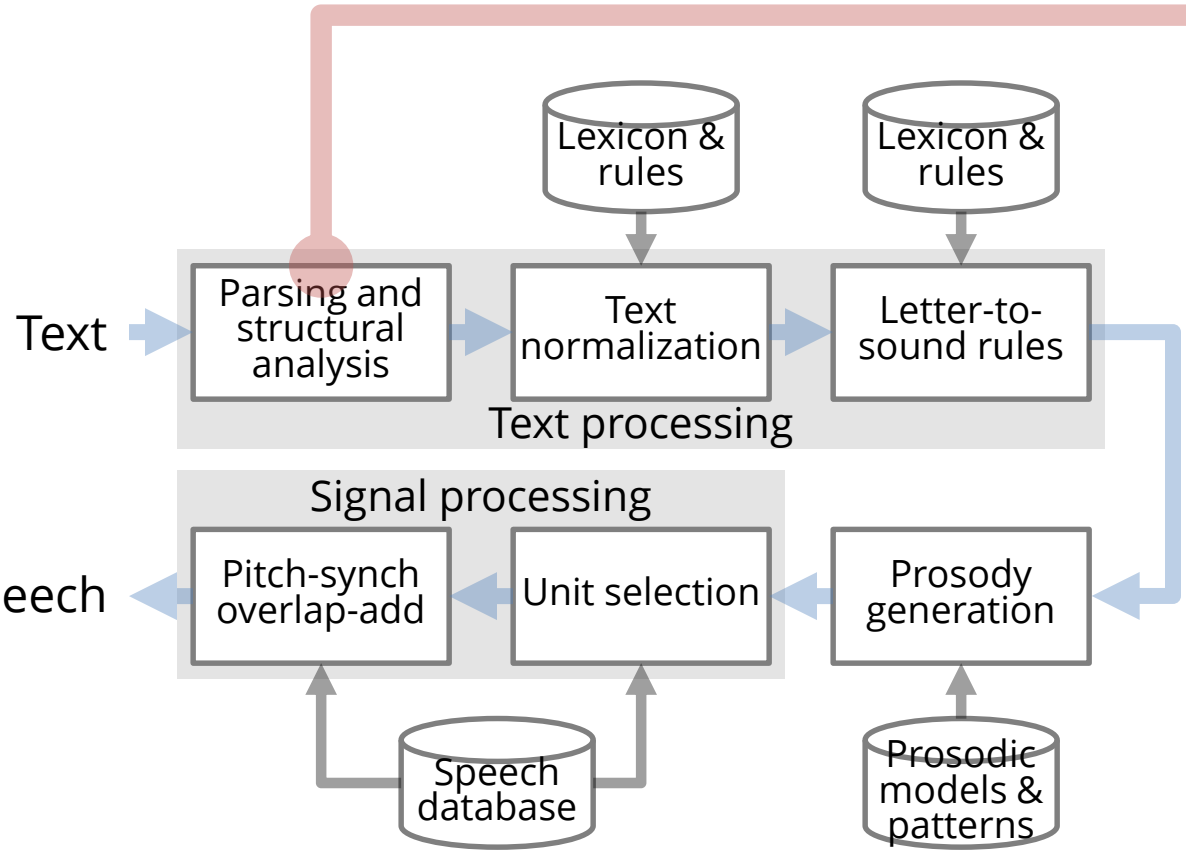
[*] Simon King, Measuring a decade of progress in Text-to-Speech, Loquens, Vol 1, No 1 (2014)

- Text-to-speech synthesis based on **unit selection** and **concatenation**

UNIT SELECTION TTS
The main blocks

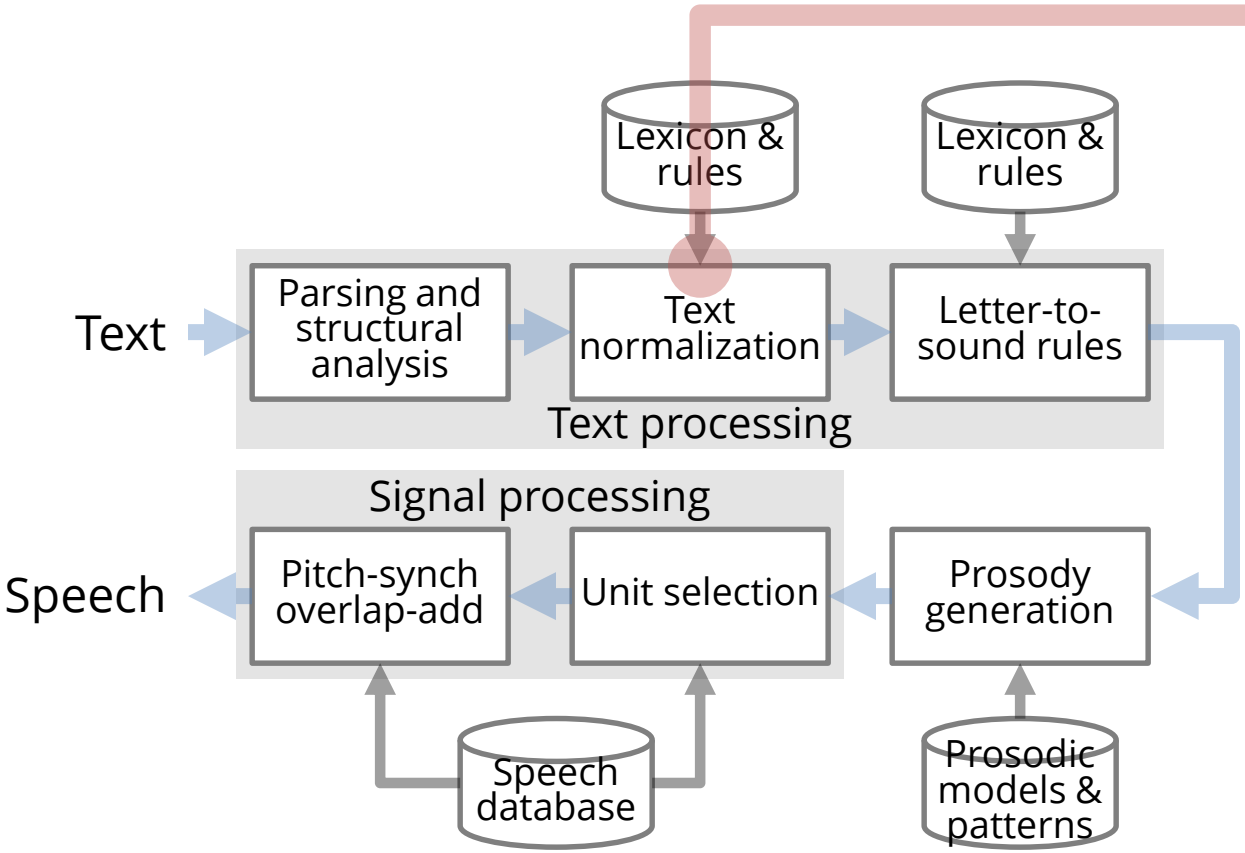


UNIT SELECTION TTS
The main blocks



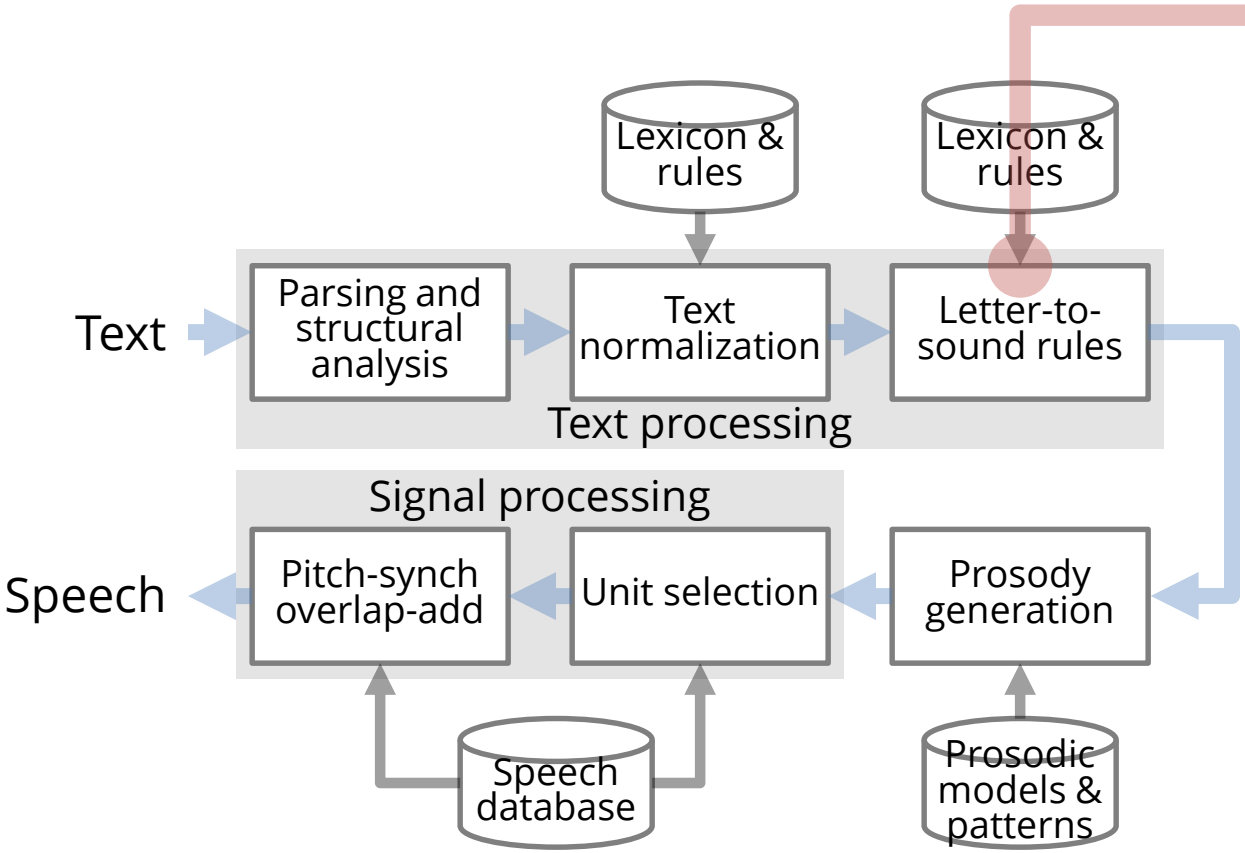
- Tokenization/ sentence breaking
- Surface linguistic parsing (NLP)
- Disambiguation (e.g. homographs)

UNIT SELECTION TTS
The main blocks



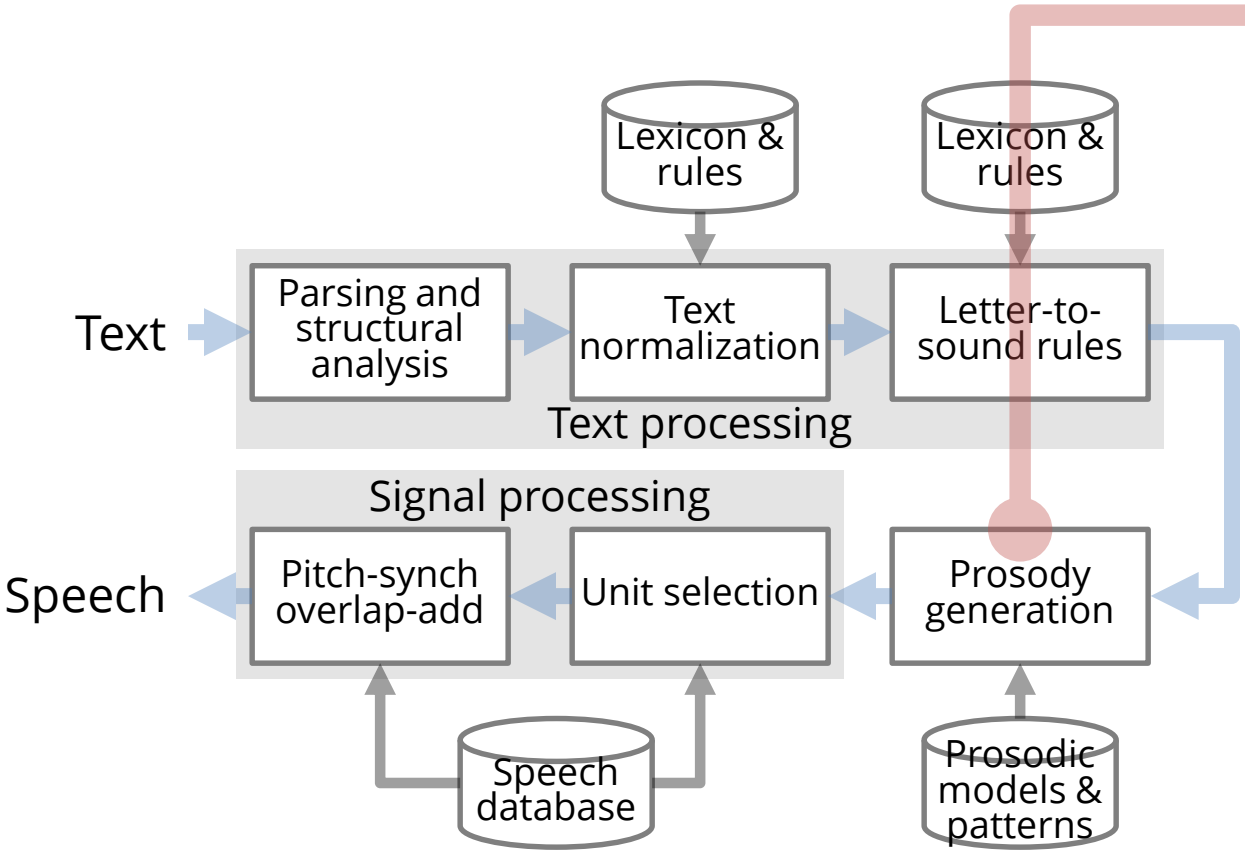
- Properly handle and expand:
  - abbreviations
  - numerals
  - addresses
  - …

UNIT SELECTION TTS
The main blocks



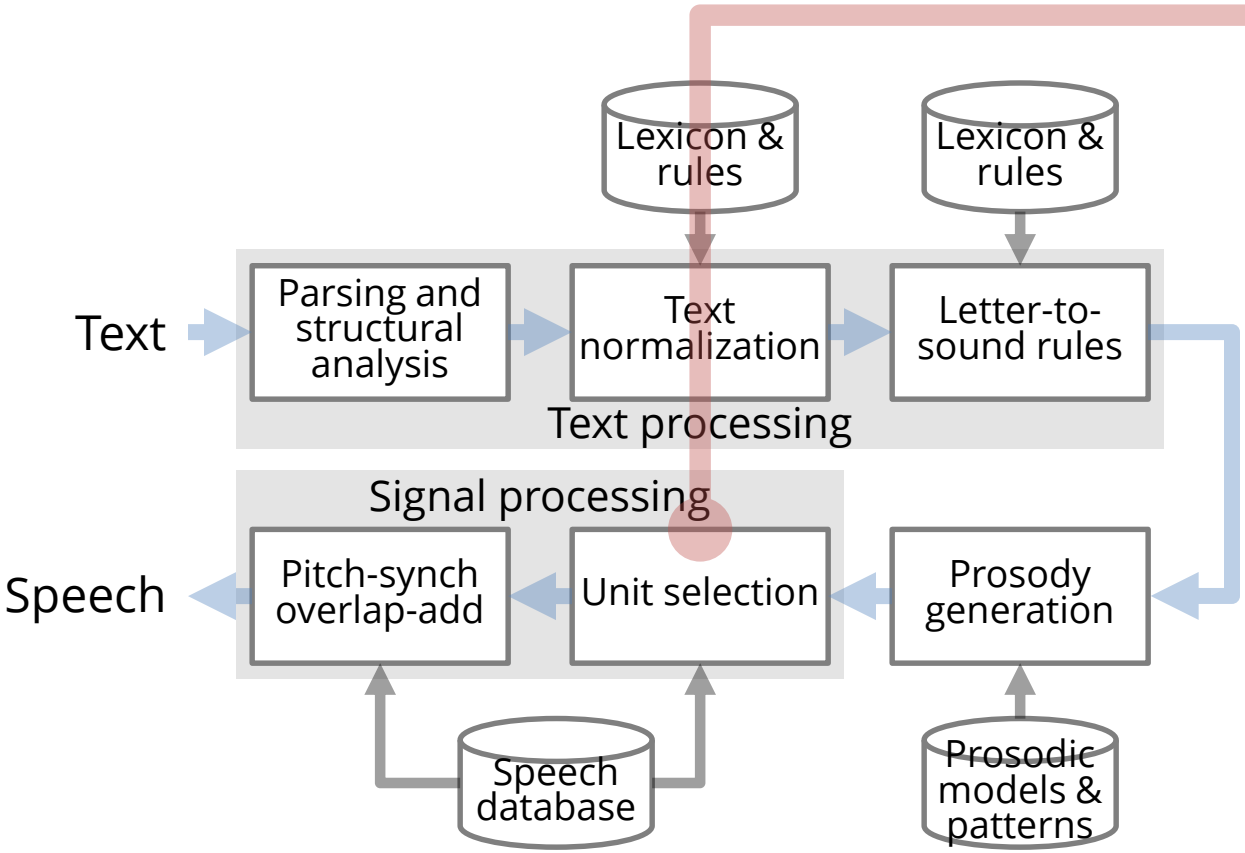
- Convert text to phonemes (and assign stress)
- Easy or hard, depending on the language
- There are always exceptions!

UNIT SELECTION TTS
The main blocks



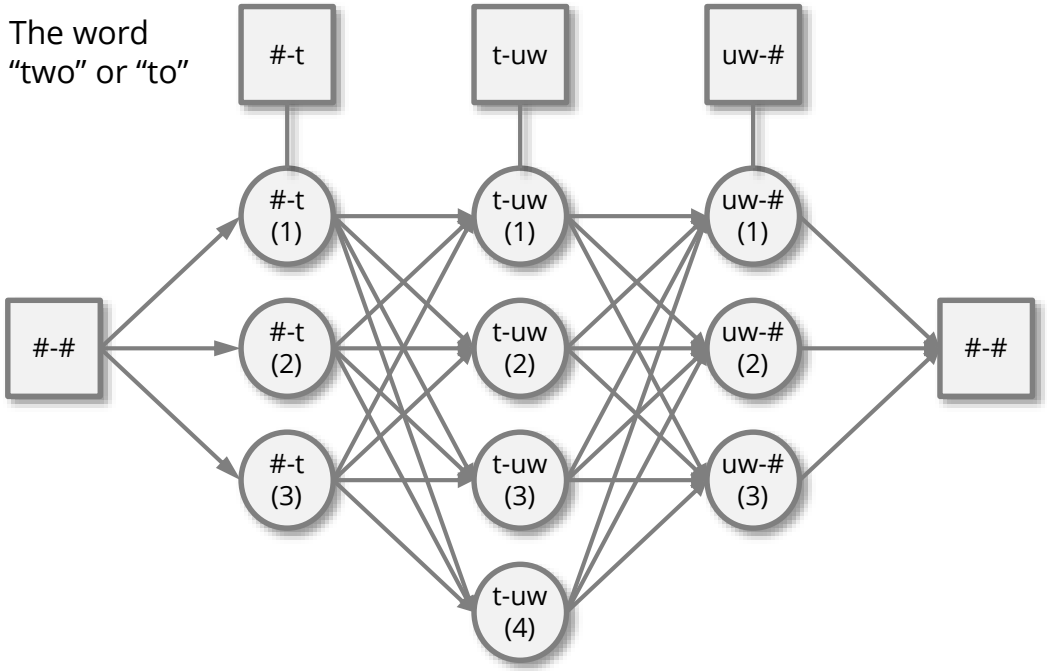
- Generate a prosody "model" for the utterance, e.g.
  - Explicit: generate target values for pitch, duration, intensity
  - Implicit: e.g. performed indirectly by unit selection

UNIT SELECTION TTS
The main blocks



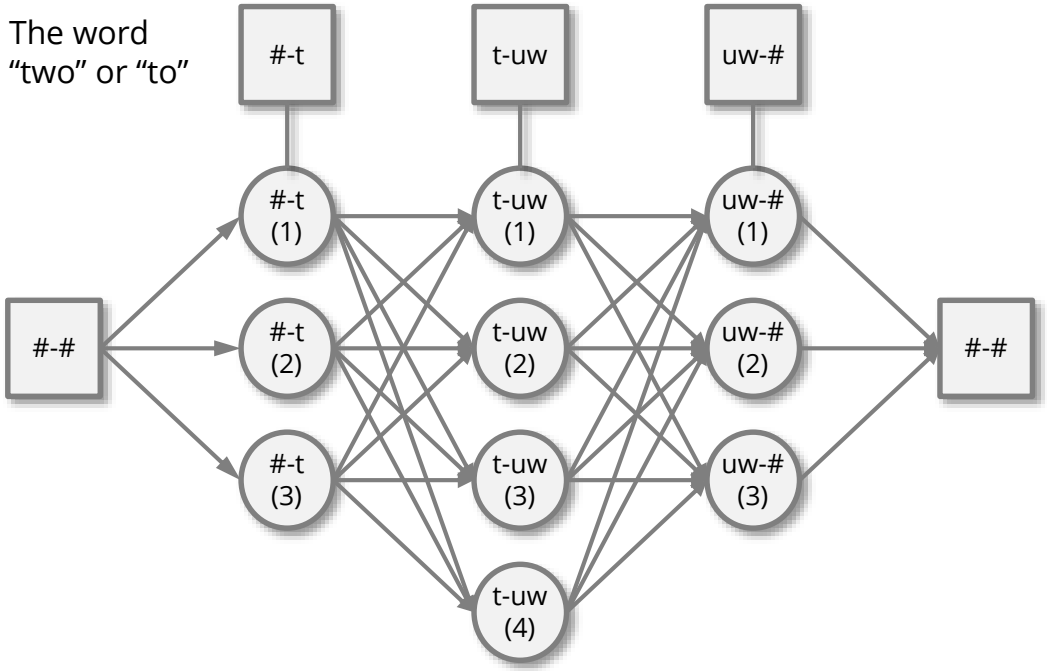
- Usually works at diphone level (i.e. “unit”=“diphone”)
- Select segments from the database to use for synthesizing the target sentence

UNIT SELECTION TTS
Selecting the units



- Select units from the database which are appropriate for the sentence to be synthesized:
  - A “good fit” for the position we want to place them at (target cost); and
  - A “good match” for their neighboring units (join cost)
- Multi-parametric optimization with a combined cost function
- Solve with Viterbi-like algorithms

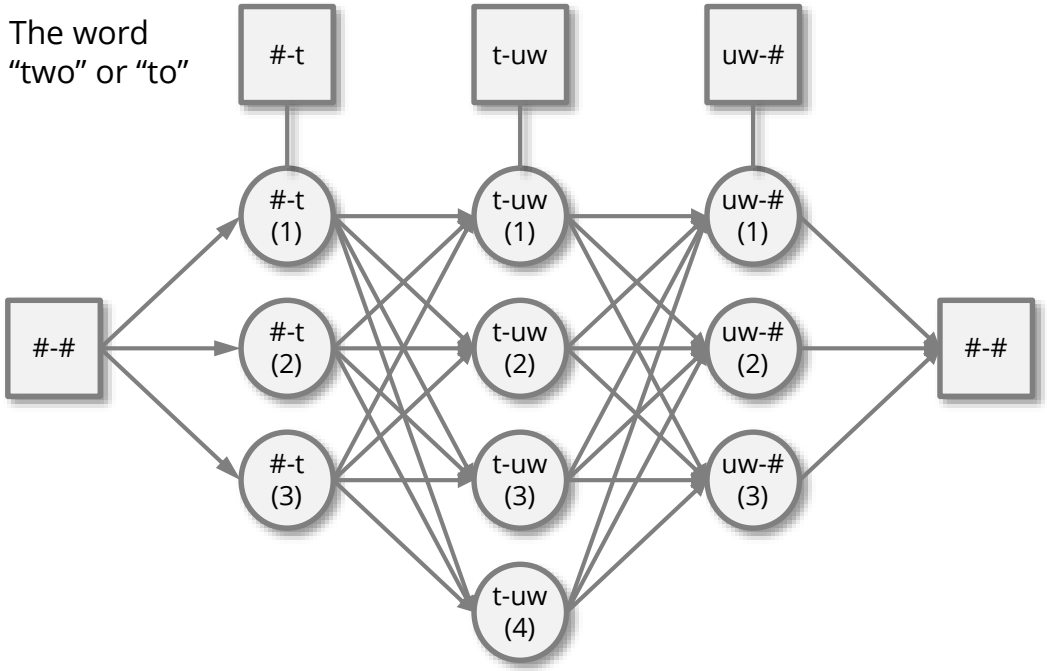
UNIT SELECTION TTS
Selecting the units



Main cost items to consider:

- **Target cost**: Context (phonetic, prosodic, linguistic…):
  - neighboring phoenemes (or families)
  - specification of prosody model (distance from prosodically significant boundaries (stress, word boundaries, sentence boundaries)
  - part-of-speech, phrase type and size, sentence type …

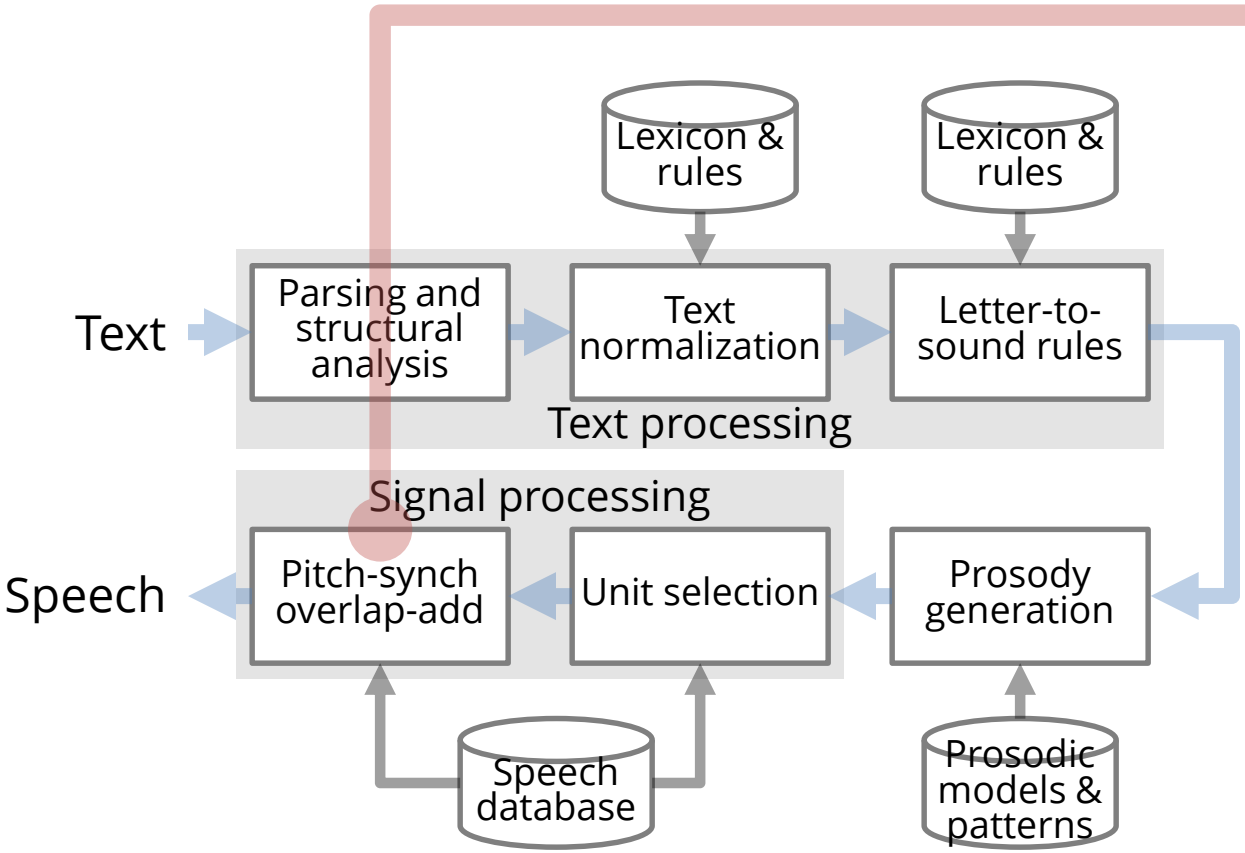
UNIT SELECTION TTS
Selecting the units



Main cost items to consider:

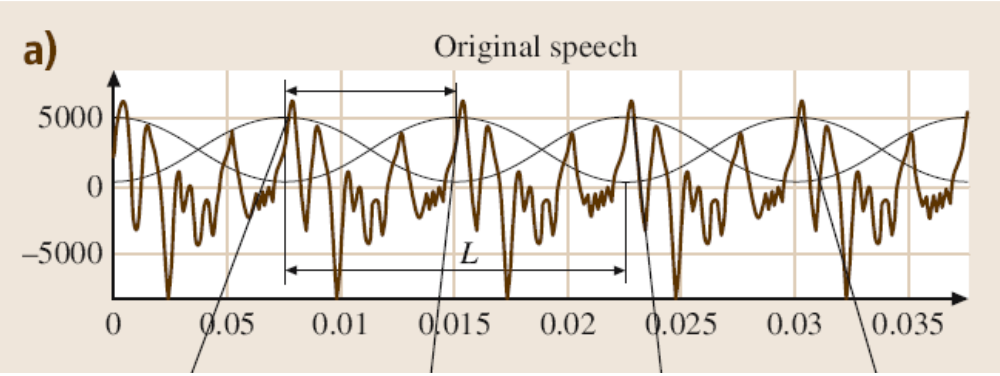
- **Join cost**:
  - acoustic distance (e.g. spectral distance)
  - prosodic distance (e.g. difference in pitch, rate, intensity)
  - distances in other measures of voice quality (e.g. quantities relating to the voice source etc)

UNIT SELECTION TTS
The main blocks



- Glue together (concatenate) the selected segments
- Waveform modification
- Overlap-add (PSOLA)

UNIT SELECTION TTS
Pitch-synchronous overlap-add



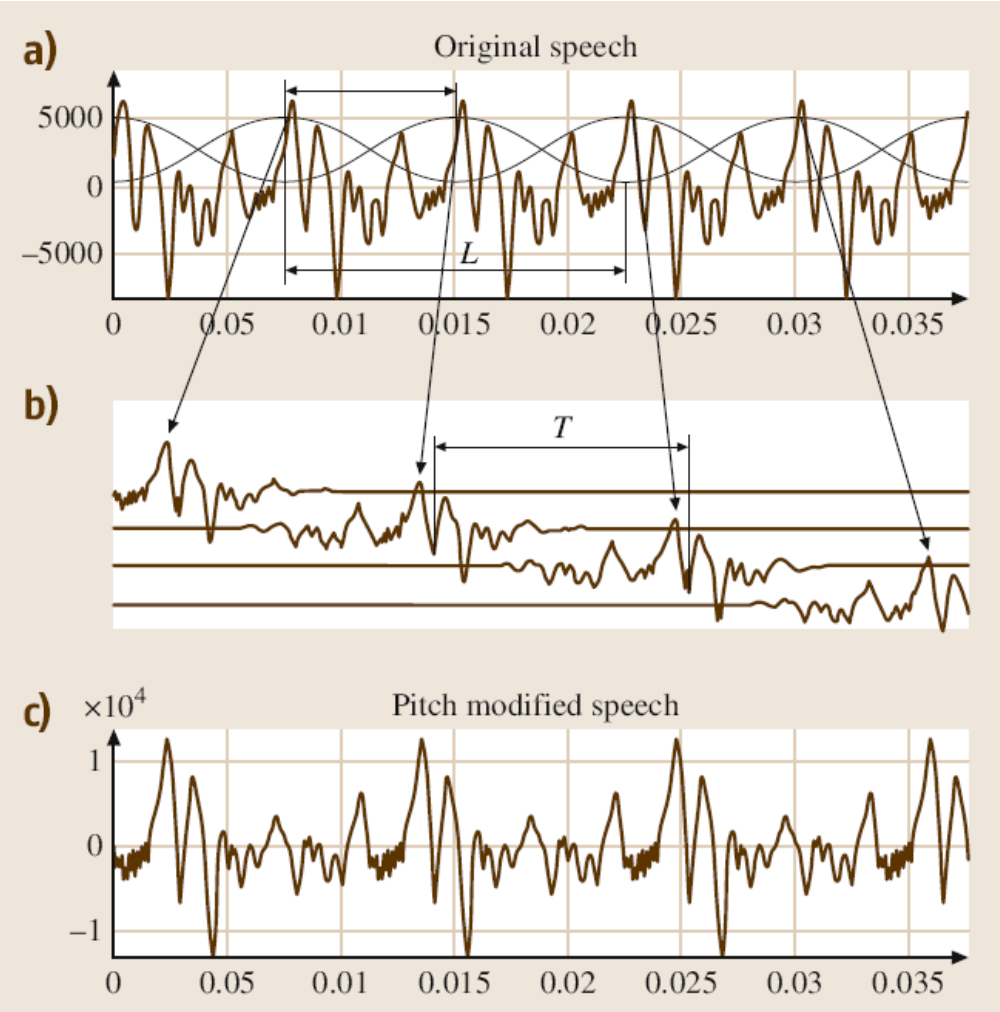
[*]

At **design**-time:

- Identify pitch periods
- Place a pitchmark at a specific place in each period (e.g. glottal closure instant)
- Be consistent!

[*] Figure source: Springer Handbook of Speech Processing, Benesty, Jacob, Sondhi, M. M., Huang, Yiteng (Eds.), Springer, ISBN 978-3-540-49125-5

UNIT SELECTION TTS
Pitch-synchronous overlap-add



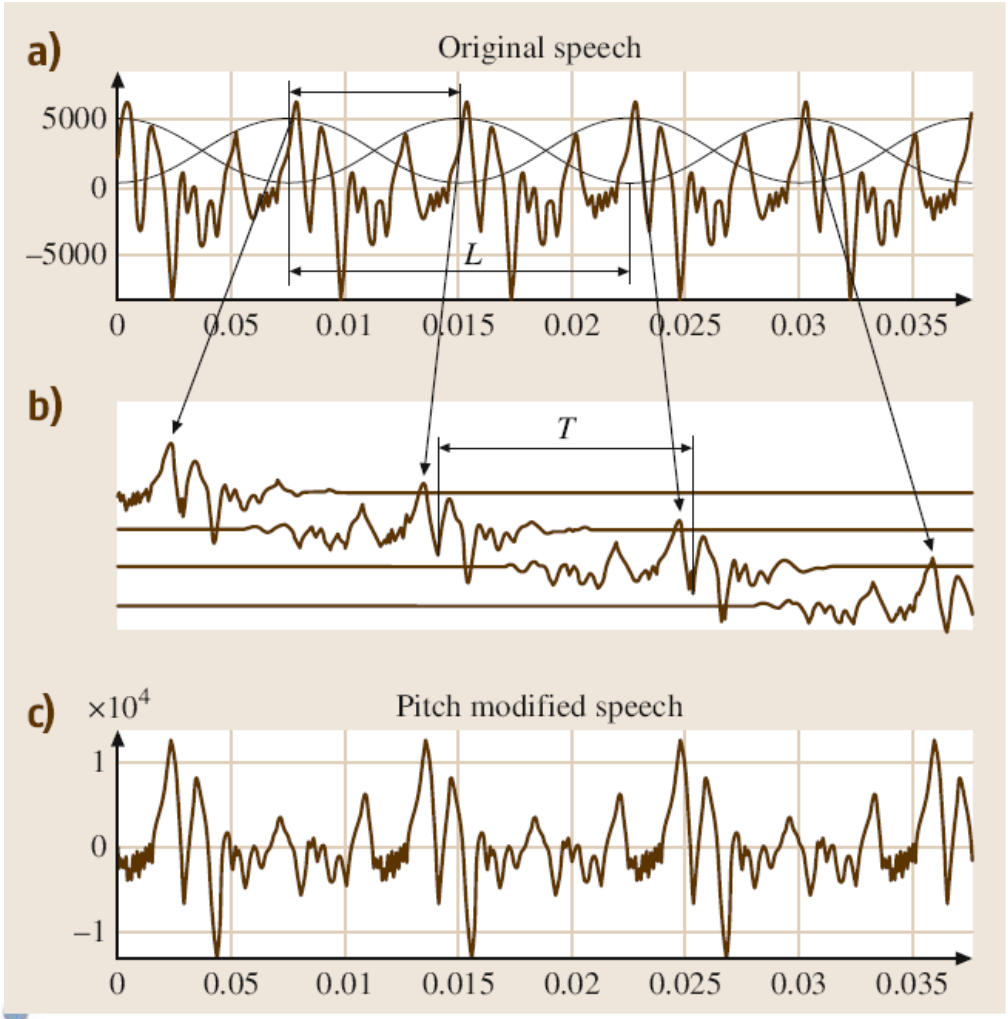
[*]

At **run**-time:

- Apply a Hanning window at the pitchmark
- Shift the frames as needed (stretch or contract them in time) → this will affect the perceived pitch (and duration)
- Repeat/omit frames → this will affect the duration (not the pitch)
- Simply add the resulting frames together

[*] Figure source: Springer Handbook of Speech Processing, Benesty, Jacob, Sondhi, M. M., Huang, Yiteng (Eds.), Springer, ISBN 978-3-540-49125-5

Pitch-synchronous overlap-add



[*]

Things to have in mind:

- Only small waveform modifications go unnoticed
- Only “compatible” units concatenate smoothly

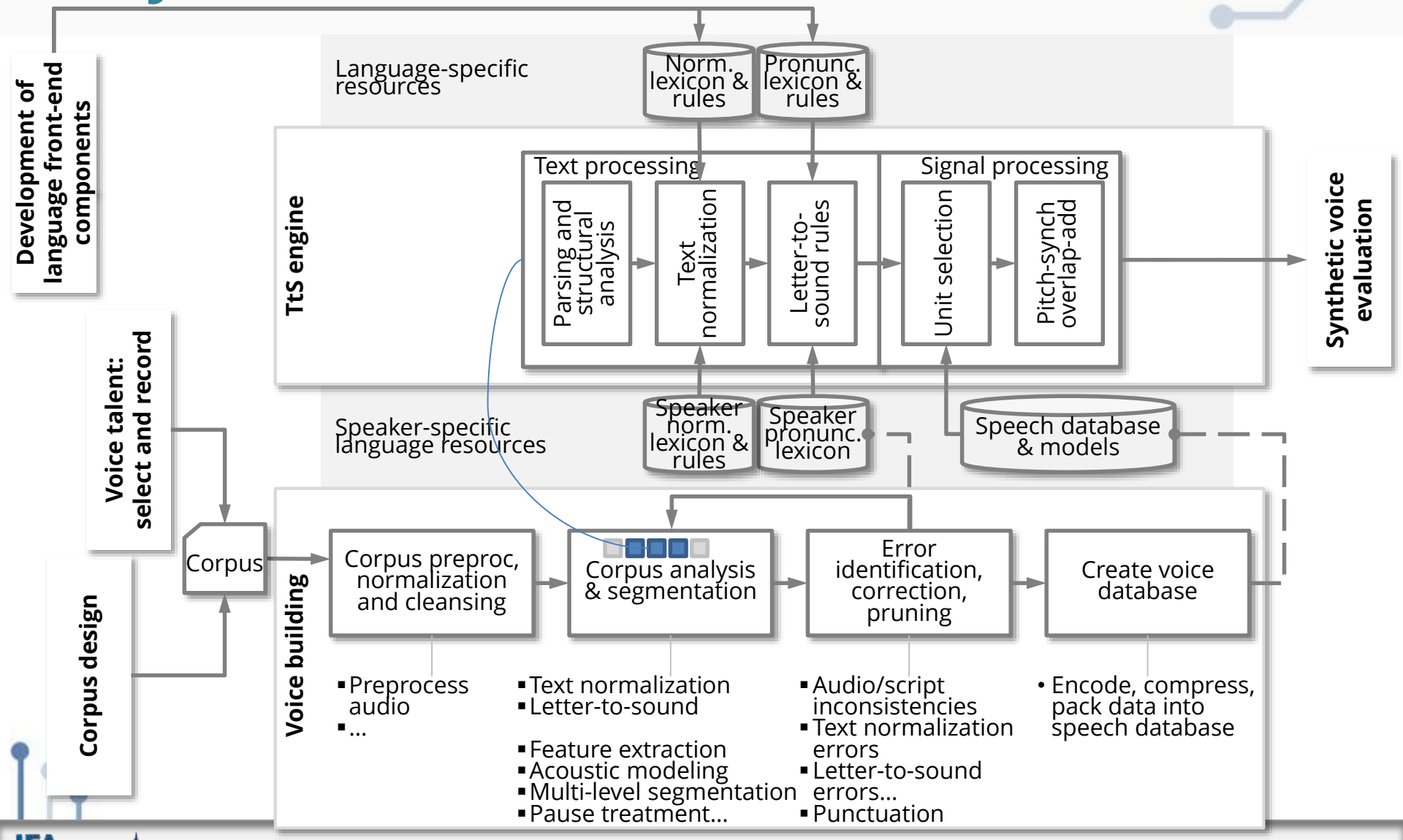
[*] Figure source: Springer Handbook of Speech Processing, Benesty, Jacob, Sondhi, M. M., Huang, Yiteng (Eds.), Springer, ISBN 978-3-540-49125-5

What matters

- Configuring and tuning the unit selection cost function; as much “art” as it is science:
  - need to tune numerous different weights and coefficients and/or devise ways to (automatically?) adapt them to different target voices
  - resort to the data itself instead of using explicit models (and coefficients), e.g. one-class classification
- The speech database:
  - various sources of errors: e.g. segmentation, letter-to-sound, F0 and pitch-marks, …
  - errors have a direct impact on the quality of the synthetic speech (e.g. glitches, discontinuities, robotic quality…)

- TtS in context:
  Other **processes** involved

UNIT SELECTION TTS
TTS synthesis in context



The innoetics/ILSP participation to the **Blizzard Challenge**

The Blizzard Challenge

- An international contest devised to better understand and compare research techniques in building corpus-based speech synthesizers on the same data.
- Organized yearly by Centre for Speech Technology Research (CSTR), University of Edinburgh, UK
- Started in 2005
- ILSP/innoetics participated since 2010
- It has been putting to the test not only the TTS engines, but also the entire voice building pipelines

http://www.synsig.org/index.php/Blizzard_Challenge

From the lab to the wild

Blizzard has been evolving, following what has been happening in the field.

Starting from a carefully crafted speech corpus recorded under controlled conditions with phonetic (and sometimes prosodic!) annotations…

…minus annotations

…minus availability of phonetic transcription:

- Indian languages: no gr2ph available and no data to extract such rules from → segmenting/synthesizing letters(!)

…minus controlled conditions:

- Audiobooks/freestyle → must learn to deal with messy data

…minus corpus:

- Actually a bulk of audio data and (often inconsistent) scripts → must learn to deal with big data

THE BLIZZARD CHALLENGE
The 2013 challenge

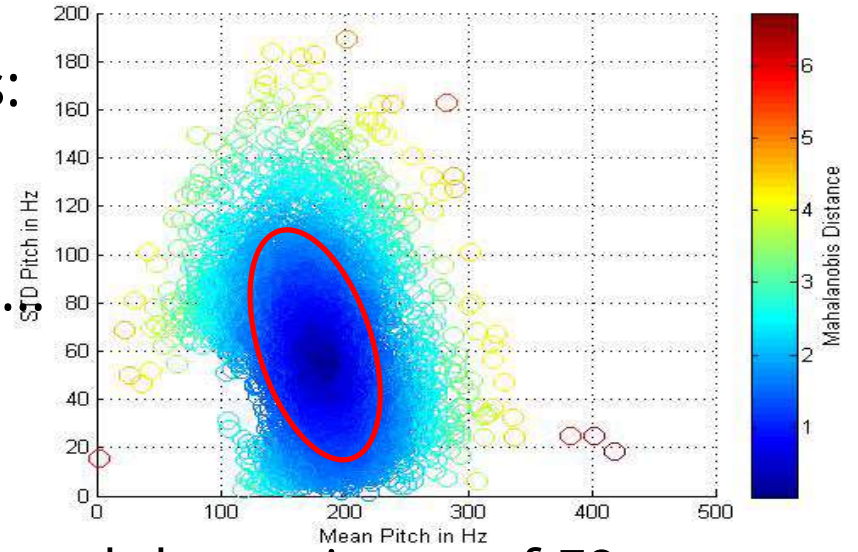
The data:

- English: Audiobook data provided by The Voice Factory. Single female speaker:
  1. Unsegmented: ~300 hours of chapter-sized mp3 files (unsegmented)
  2. ~19 hours of wav files (segmented into sentences and aligned with the text by Lessac Technologies, Inc.)
- Indian languages: About 1 hour of speech data in each of four Indian languages (Hindi, Bengali, Kannada and Tamil)

The tasks:

- Task EH1 – build a voice from the unsegmented audio
- Task EH2 – build a voice from the segmented audio
- Task IH1 – build one voice in each language from the provided data

The 2013 challenge

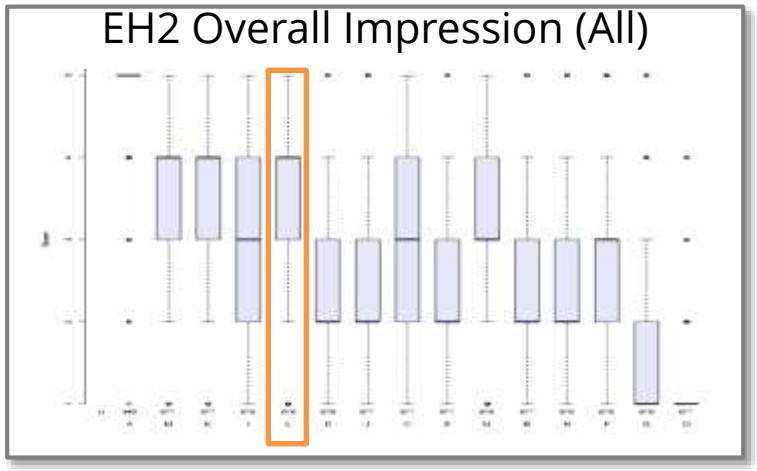


Pruning – throwing away extremists:

- Audiobooks were expressively too rich: extreme voice acting, roles, impersonations, imitations.
- "Acoustic phrases": audio part between two consecutive (recognized) silences
- Simplistic phrase features: mean and the variance of F0 variable on each phrase
- Prune the ones further away from the distribution's centroid
- This was found to be quite successful in keeping "neutral" speech

THE BLIZZARD CHALLENGE
The 2013 challenge

I remember the whole beginning as a succession of flights and drops, a little seesaw of the right throbs and the wrong. After rising, in town, to meet his appeal, I had at all events a couple of very bad days -- found all my doubts bristle again, felt indeed sure I had made a mistake. In this state of mind I spent the long hours of bumping, swinging coach that carried me to the stopping place at which I was to be met by a vehicle from the house. This convenience, I was told, had been ordered, and I found, toward the close of the June afternoon, a commodious fly in waiting for me.



THE BLIZZARD CHALLENGE
The 2013 challenge

So the subject was changed in deference to the children's presence, and we went on talking about other things.

So I set about it, and after great labor and tedious research accomplished my task.

Mrs. Allen was one of that numerous class of females, whose society can raise no other emotion than surprise at there being any men in the world who could like them well enough to marry them. She had neither beauty genius, accomplishment, nor manner. The air of a gentlewoman, a great deal of quiet, inactive good temper, and a trifling turn of mind were all that could account for her being the choice of a sensible, intelligent man like Mr. Allen. In one respect she was admirably fitted to introduce a young lady into public, being as fond of going everywhere and seeing everything herself as any young lady could be. Dress was her passion.

- Tapping into **expressiveness**

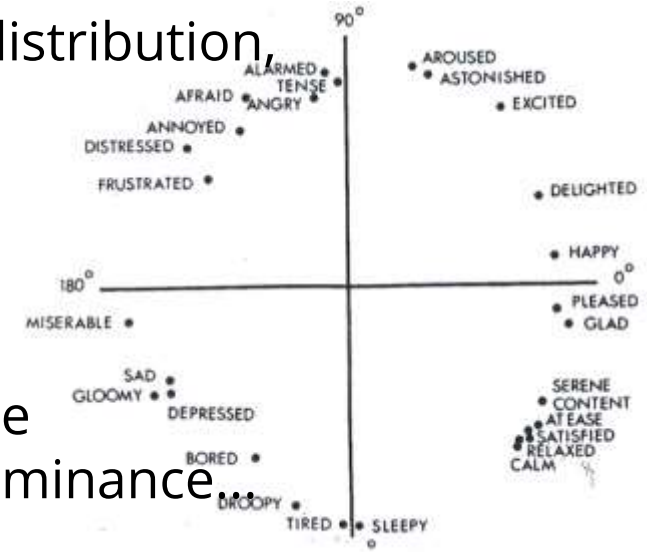
Expressive speech synthesis?

So, we can mimic expressiveness…

…but can we imitate it?

Speech emotion recognition

- Extract affective or emotional content from speech.
  - wide variability of expressivity patterns in human speech
  - still rather limited understanding of how these are linked to various qualities of speech
- Speech characteristics involved:
  - pitch, duration/rate, intensity, energy distribution, voice source characteristics…
- Dominant approaches:
  - categorical: e.g. big-six
  - dimensional: finite set of underlying dimensions into which emotions can be decomposed, e.g. pleasure/arousal/dominance…



Speech emotion recognition

But…

- cannot capture the rich variability and subtle nuances of expression in human speech
- fully-blown emotions cannot be expected in most domains and applications
- many speech applications involve other expressive speaking styles in addition to, or instead of, the expression of emotions…

e.g. child-directed speech and storytelling

When **expressive speech synthesis** is the target, the task goals can be somehow different

In the outset

The approach:

- Does not assume (nor impose) a specific emotion theory

  Instead, it employs a data-driven approach
- Does not attempt to directly link measurable speech features to specific high-level emotions…

  Instead, it seeks to reveal underlying structure and latent components
- Does not confine itself to a fixed set of pre-defined, fully-blown emotions

  Instead, it seeks to explore the expressive patterns employed (to convey emotions, style and other information)
- Ultimate goal is to model and imitate expressive speaking styles for expressive speech synthesis

Storytelling (synthetic)

- Deliver, through synthetic speech, a listening experience that is equally engaging as that provided by a human storyteller
- To achieve that we need to address certain questions:
  - is expressiveness **distinguishable** in natural speech in terms of measurable features?
  - which are those **feature patterns** that serve well for this purpose?
  - explore the wide **repertoire** of expressive patterns employed by human speakers in the course of rich narratives, which are perceived to carry an expressive load and are employed to fulfill a (short-term) narrative goal
  - is there an "**expressive typology**" that naturally emerges from the data itself (not imposed by our theory)

Steps for a pilot investigation

1. Corpus selection and pre-processing
2. Corpus annotation
3. Analysis and feature extraction
4. Extracting latent features
5. Predicting RoEIs
6. Hidden order?

1. Corpus selection and pre-processing

- Recordings from a professional speaker of a tale, used for the production of an educational software for children (in Greek)
- It was recorded using professional equipment at an anechoic studio at ILSP, resampled at 16KHz
- Only narrative, single-speaker utterances have been selected
- For the purposes of this investigation, a small subset of ~100 sentences was extracted
- Segmentation through speech recognition

The ultimate target is expressive speech synthesis.
NOT emotion recognition, or modeling expressive speech in general,
or identifying common features or universalities, or…!
So, single speaker is OK (actually a must), limited size can be OK

2. Corpus annotation

- Manual annotation of regions of expressive interest (RoEIs):
  Speech segments which, according to the subjective judgment of the listener/researcher, bared some salient expressive load or were uttered in a way meant to invoke affect to the audience.
- No explicit annotations or expressive labels have been assigned, but a plain on/off flag indicating whether a word is perceived as expressively loaded.

3. Analysis and feature extraction [1/4]

Words have been chosen as the basic analysis unit; i.e. they are assumed to be the basic 'atomic' unit of expressiveness

- working at sentence level, may introduce difficulties, especially when expressive speech synthesis is also desired. This may be performed in a much more straightforward manner when working at word level.
- employing a word-based level of analysis permits to fuse in linguistic knowledge at the analysis or generation stages, i.e. to associate expressive patterns applied to a word, to its grammatical properties or its place in the syntactic structure of the sentence.

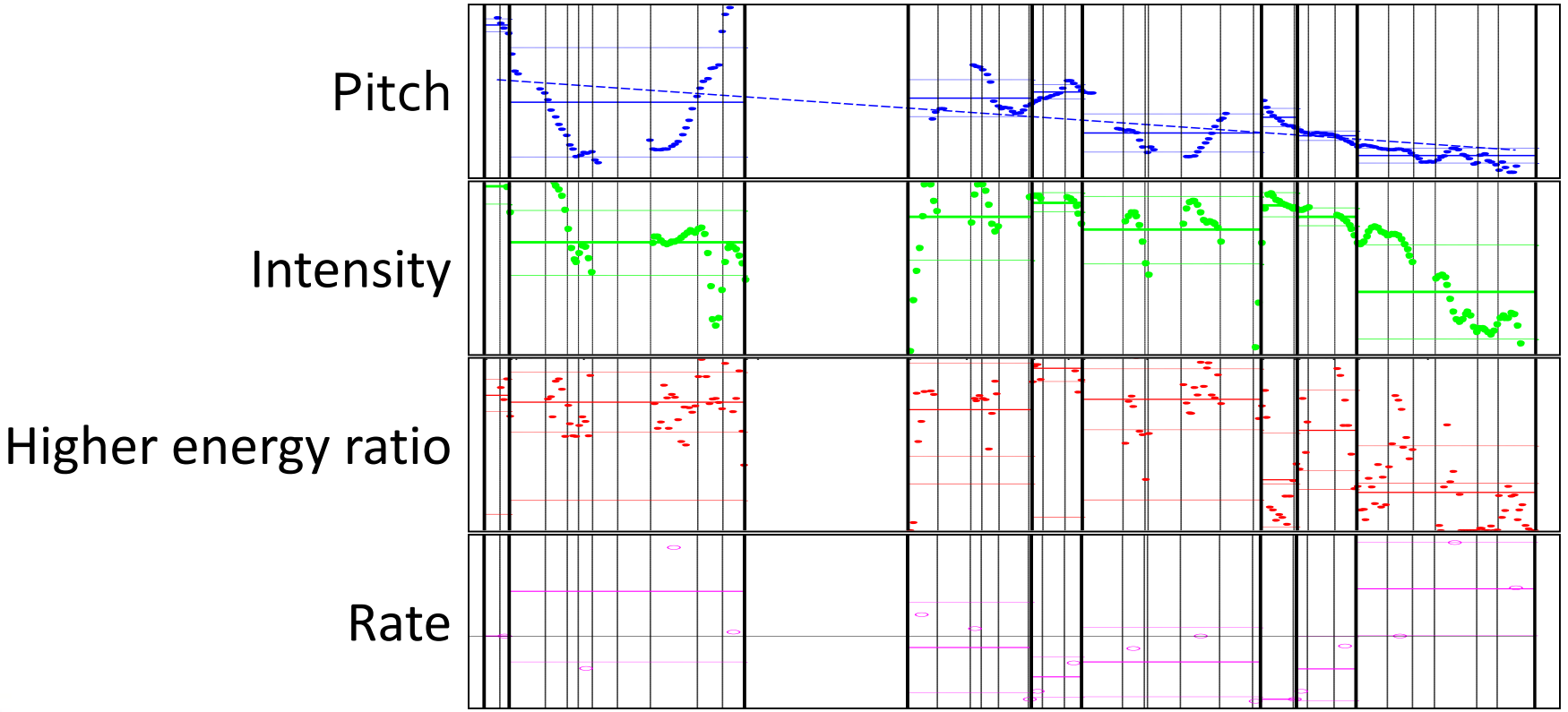
A higher unit level would also be valid, e.g. phrases

3. Analysis and feature extraction [2/4]

- The features employed relate to four main speech quantities:
  - **pitch** and **intensity**, measured over the voiced frames of the word and averaged
  - **energy**, as the ratio of the energy at the higher 1/5 of the frequency range over the total energy of the spectrum, normalized to 100
  - **rate**, normalized phoneme durations averaged over the entire word
- To obtain values for these features at the **word level**, their mean values (_MEAN) and standard deviations (_STD) have been calculated over the entire word

3. Analysis and feature extraction [3/4]

"Το περιστέρι, έφερνε ένα χαρτάκι με ένα μήνυμα"
(The dove was carrying a piece of paper with a message)
[to peristéri <pause> éferne éna xartáci me éna mínima].



3. Analysis and feature extraction [4/4]

Context

- PREV_PAUSE [bool]:
- NEXT_PAUSE [bool]: true when the previous/next word is a pause

Acoustic features

- Fo_MEAN
- Fo_STD
- INTENSITY_MEAN
- INTENSITY_STD
- ENERGY_MEAN
- ENERGY_STD
- RATE_MEAN
- RATE_STD

Acoustic feature deltas

- PREV_DELTA_Fo_MEAN
- NEXT_DELTA_Fo_MEAN
- PREV_DELTA_Fo_STD
- NEXT_DELTA_Fo_STD
- PREV_DELTA_INTENSITY_MEAN
- NEXT_DELTA_INTENSITY_MEAN
- PREV_DELTA_INTENSITY_STD
- NEXT_DELTA_INTENSITY_STD
- PREV_DELTA_ENERGY_MEAN
- NEXT_DELTA_ENERGY_MEAN
- PREV_DELTA_ENERGY_STD
- NEXT_DELTA_ENERGY_STD
- PREV_DELTA_RATE_MEAN
- NEXT_DELTA_RATE_MEAN
- PREV_DELTA_RATE_STD
- NEXT_DELTA_RATE_STD

4. Extracting latent features [1/4]

- From correlation analysis, it becomes clear that the feature set has a large degree of **redundancy**
- Part of it can be discarded without losing any of its interpretational ability, for example by **reducing the dimensionality** of the set.
- Working with a feature space of reduced dimension can simplify the investigation and provide further intuition on the contribution of each feature in what listeners perceive as expressively prominent regions in speech and, thus, in a more effective way of detecting them.

4. Extracting latent features [2/4]

- Applied Principal Component Analysis (PCA)
- Kept only the components that seem to capture significant part of the variance (values larger than 1,00)

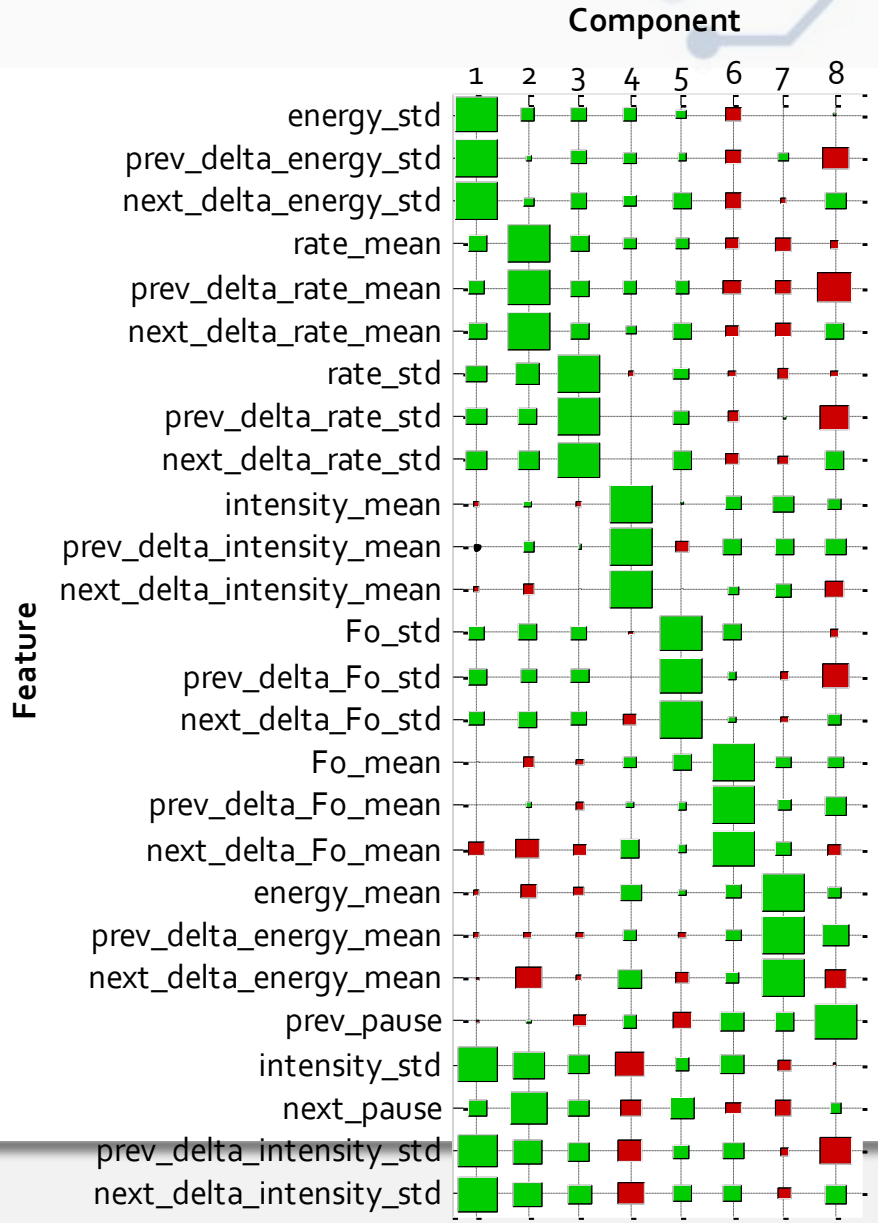
8 components
~74% of variability

	Initial Eigenvalues		
	Var	% of Variance	Cum. %
	1	2	3
1	6.24	24.01	24.01
2	3.18	12.23	36.25
3	2.20	8.47	44.72
4	2.02	7.77	52.48
5	1.68	6.47	58.95
6	1.48	5.70	64.65
7	1.31	5.03	69.68
8	1.11	4.26	73.94
9	0.94	3.61	77.55
10	0.77	2.97	80.52
...			
26	0.11	0.41	100.00
	26.00	100.00	

(1) the variance captured by each component,
(2) the percentage of the total variance in the data captured by the respective component,
(3) the cumulative variance captured

4. Extracting latent features [3/4]

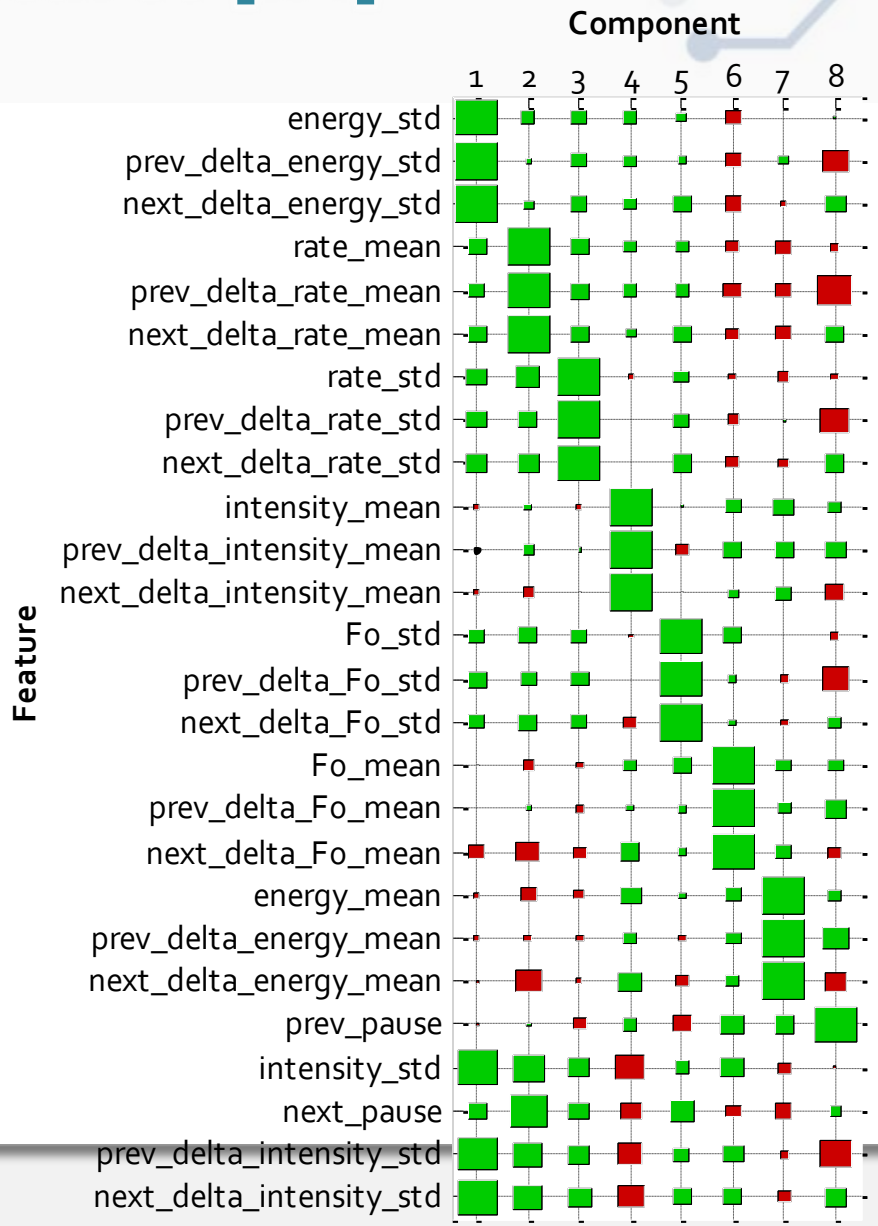
- Association of original features and extracted components (features have been reordered so as to group together the ones that are more strongly associated with each of the components)
- Insight into the role of extracted components → reveals some structure



4. Extracting latent features [3/4]

- **Component 1 – "energy variance component"**: strongly related to energy features of words and weakly related to any other features, with the exception of some that relate to intensity.
- **Component 2 – "mean rate component"**: strongly related to mean rate variables and weakly related to any other variable.
- **Component 3 – "rate variance component"**…

Components: good "concepts" (?)



5. Predicting RoEIs [1/2]

- **Question**: How reliably can RoEIs be predicted based on the extracted components ?
  I.e. *can we, based on a word's surface acoustic features, determine whether it has expressive load (or, more precisely, whether the specific researcher would annotate it as such).*
- This is formulated as a typical binary classification problem

5. Predicting RoEIs [2/2]

- Logistic regression results

Observed			Predicted		
			marked		Percentage Correct
			,00	1,00	
Step 7	marked	,00	576	133	81,2
		1,00	195	375	65,8
		Overall Percentage			**74,4**

- This is fairly decent, given the arbitrary manner that the words have been originally annotated
- So, we could afford to use this classifier to automatically mark RoEIs in a much larger corpus

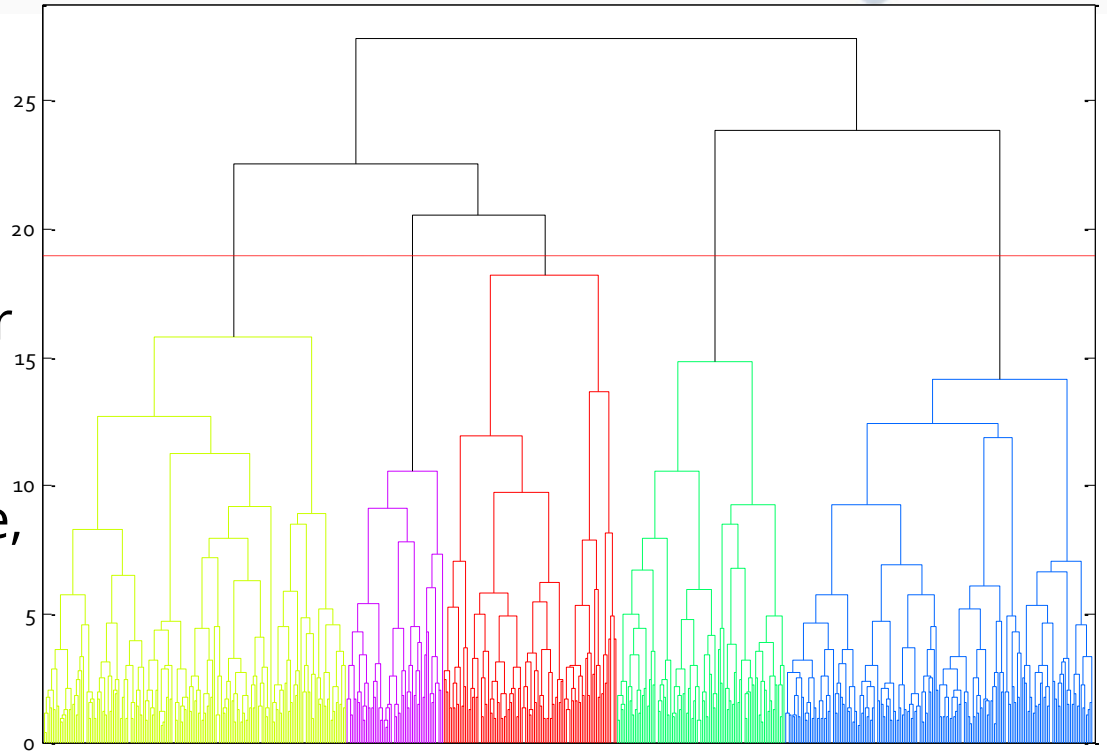
6. Hidden order? [1/4]

- **Question**: Is there a "natural" division of the expressively loaded words into subcategories based on their surface acoustic features (or, rather, the latent components previously extracted)?
  - This would be a first step for revealing any latent structure, patterns or regularities that are observed in the expressive characteristics employed in storytelling, thus leading the way to an "expressive typology" for storytelling.
  - This can be formulated as a typical clustering problem
  - Hierarchical clustering preferred over, e.g. *k*-means clustering, as the desired number of categories cannot be estimated in advance

6. Hidden order? [2/4]

- **Normalization**: z-scores
- **Metric**: Euclidean distance
- **Linkage criterion**: Ward's minimum variance method (the decrease in variance for the cluster being merged)

- Results in a "clean" structure, with sufficiently distinct and well separated clusters
- Cutting off at the level displayed, leads to 5 categories



Height represents the distance between two clusters being joined

Category	Cardinality	Color
1	52	
2	94	
3	165	
4	167	
5	92	

6. Hidden order? [3/4]

- **Discriminant analysis** employed to formulate discriminant functions for the 5 categories based on the 8 components
- Basic **interpretation** of categories can be based on the coefficient of each component in each discriminative function
- The component that is most significant for a category is shown in black; other significant components in gray

Comp.		Category				
		1	2	3	4	5
1	Energy variance	-,009	2,033	,906	-,087	,232
2	Mean rate	,860	1,460	1,560	-,422	,257
3	Rate variance	2,460	,399	,592	,323	-,368
4	Mean intensity	-,998	2,167	-1,009	1,218	1,175
5	Fo variance	-,406	,203	2,308	-,056	-,318
6	Mean Fo	-,232	,124	-,687	,002	,662
7	Mean energy	-,926	-,918	,012	-,388	,730
8	Preceding pause	-1,161	-1,245	-,411	1,716	3,156
	Constant	*-4,896*	*-4,572*	*-3,949*	*-2,623*	*-3,831*

6. Hidden order? [4/4]

- Evaluation of a classifier built to separate different categories
- Classification/confusion table (*with* cross-validation)
- An overall 80,7% was categorized correctly, i.e. the model provides a sufficient description of the data.

		Predicted category					Total
	Cat.	1	2	3	4	5	
Count	1	37	4	1	6	4	52
	2	7	67	11	5	4	94
	3	6	17	129	8	5	165
	4	8	3	5	144	7	167
	5	0	1	3	5	83	92
%	1	**71,2**	7,7	1,9	11,5	7,7	100,0
	2	7,4	**71,3**	11,7	5,3	4,3	100,0
	3	3,6	10,3	**78,2**	4,8	3,0	100,0
	4	4,8	1,8	3,0	**86,2**	4,2	100,0
	5	,0	1,1	3,3	5,4	**90,2**	100,0

Remarks

It’s all about speech synthesis…

- we don’t seek to detect what is the speakers’ state; we merely want to imitate their speaking style
- we don’t need to understand (or even assign labels) to the speaking styles that emerge; we just need to gracefully reproduce them
- we don’t need to find THE “right” expressive style to assign to a synthetic word/phrase/sentence/…; a plausible one would do just fine… *and there should be quite a few in storytelling!*

Contact

Spyros Raptis
Institute for Language and Speech Processing /
"Athena" Research Center
spy@ilsp.gr
http://www.ilsp.gr