Text-to-Speech Synthesis (TTS)

TAL-IDMC-OS-UL-2021-22

1

Acoustic: Text ➔ Speech

Audiovisual: Text ➔ Speech + Face

TAL-IDMC-OS-UL-2021-22

4

What for?

- Communication/Messaging
- Multimedia
- Dialog (TTS+ASR)
- Human-Machine Interface
- Speech impairment, sight and hearing impairments

Used Technologies

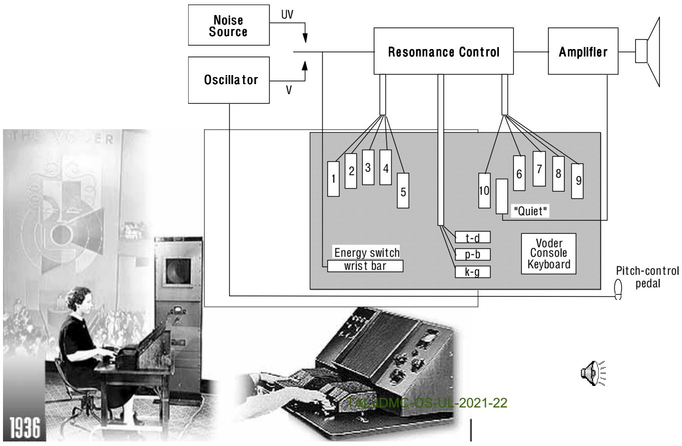
- Rule-based synthesis
- Diphone-based synthesis
- Unit selection-based synthesis
- DNN

Challenges

- Intelligibility
  - Speech is not a raw concatenation of speech chunks
  - We are **very** sensitive to **naturalness**

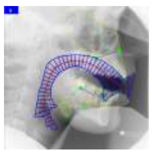
History of TTS

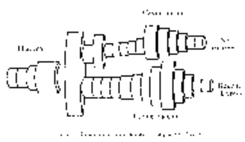
(The Voder, 1936)

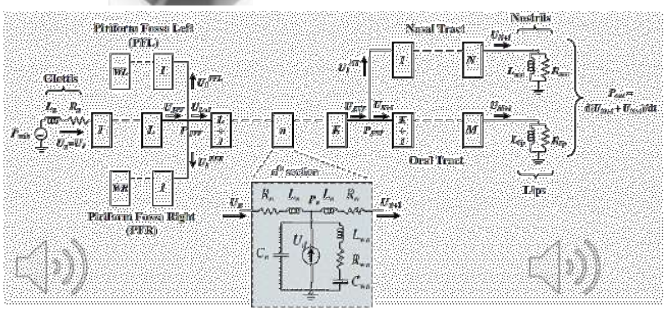


7

Articulatory Synthesis







Rules to control a low-dimensionality articulatory model, by Cecil Coker, 1968.

Sentences produced by an articulatory model, James Flanagan and Kenzo Ishizaka, 1976.

9

Formant synthesis technology

- The synthesized speech output is created using additive synthesis and an acoustic model (physical modelling synthesis).
- Parameters such as fundamental frequency, voicing, and noise levels are varied over time to create a waveform of artificial speech.
- This method is sometimes called **rule-based** synthesis.
- Formant synthesis generates artificial, robotic-sounding speech that would never be mistaken for human speech.

Output from the first computer-based phonemic synthesis-by-rule program, created by John Kelly and Louis Gerstman, 1961.

Domain-specific synthesis

- Domain-specific synthesis concatenates prerecorded words and phrases to create complete utterances.
- It is used in applications where the variety of texts the system will output is limited to a particular domain:
  - transit schedule announcements
  - weather reports.
  - Etc.

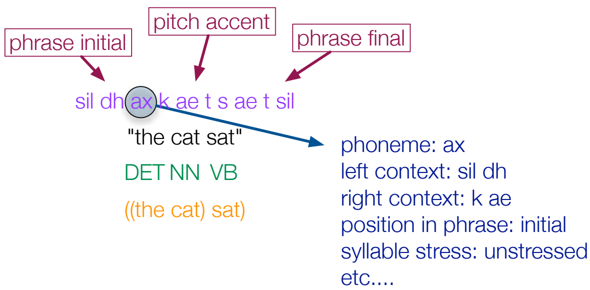
Text processing

- Text processing breaks the original input text into units suitable for further processing:
  - expanding abbreviations
  - part-of-speech (POS) tagging
  - letter-to-sound rules
  - prosody prediction
- We end up with a '**linguistic specification**', all the information required to generate a speech waveform, such as
  - phone sequence
  - phone durations
  - pitch contour

12

From text...

NLP: Natural Language Processing



Extracted from Simon King, University of Edinburgh

13

From text... NLP: Natural Language Processing

«Le grand président est arrivé à 8h10.»

1. Pre-processing:

Le	[MOT]
Grand	[MOT]
Président	[MOT]
Est	[MOT]
Arrivé	[MOT]
À	[MOT]

8h10 ➔huit heure dix. [TIME]

TAL-IDMC-OS-UL-2021-22

14

From text... NLP: Natural Language Processing

«Le grand président est arrivé à 8h10.»

2. Lemmatization :

Le	[**ARTICLE** | **PRONOUN**]
Grand	[**ADJECTIVE**| **NOUN**]
Président	[**NOUN** | **VERB**]
Est	[**NOUN** | **AUXILIARY** | **VERB**]
Arrivé	[**PAST PARTICIPLE** | **ADJECTIVE**]
À	[**PREPOSITION**]

8h10 ➔huit heure dix. [**TIME**]

[NUM] [NOUN] [NUM]

Lemmatization is the algorithmic process of determining the lemma of a word based on its intended meaning.

TAL-IDMC-OS-UL-2021-22

15

From text...

NLP: Natural Language Processing

«Le grand président est arrivé à 8h10.»

2. Lemmatization :

Le	[**ARTICLE**]
Grand	[**ADJECTIVE**]
Président	[**NOUN**]
Est	[**AUXILIARY**]
Arrivé	[**PAST PARTICIPLE**]
À	[**PREPOSITION**]

8h10 ➔huit heure dix. [**TIME**]
[NUM] [NOUN] [NUM]

From text...

NLP: Natural Language Processing

«Le grand président est arrivé à 8h10.»

3. Phonetization:

Le	[**ARTICLE**]	l @
Grand	[**ADJECTIVE**]	g R a~
Président	[**NOUN**]	p R e z i d a~
Est	[**AUXILIARY**]	e
Arrivé	[**PAST PARTICIPLE**]	a R i v e
À	[**PREPOSITION**]	a

8h10 ➔huit heure dix. [**TIME**] H i t 9 R d i s
[NUM] [NOUN] [NUM]

Phonetization is the process of encoding language sounds using phonetic symbols.

From text...

NLP: Natural Language Processing

«Le grand président est arrivé à 8h10.»

4. Post-phonetization :

Le	[**ARTICLE**]	l @
Grand	[**ADJECTIVE**]	g R a~
Président	[**NOUN**]	p R e z i d a~
Est	[**AUXILIARY**]	e **t**
Arrivé	[**PAST PARTICIPLE**]	a R i v e
À	[**PREPOSITION**]	a
8h10 ➔huit heure dix. [**TIME**] [NUM] [NOUN] [NUM]		H i t 9 R d i s

From text...

NLP: Natural Language Processing

- Access to an exception dictionary:
  - Some Acronyms
  - Foreign words
  - Proper names
- Post-phonetization
  - By rules

From text...

NLP: Natural Language Processing

«Le grand président est arrivé à 8h10.»

5. *Intonation/Prosody* :

l@ gRa~ pRezida~ et aRive a Hit 9R dis

l@ gRa~ pRezida~ et aRive a Hit 9R dis

TAL-IDMC-OS-UL-2021-22

20

From text...

NLP: Natural Language Processing

- Prosody model
  - Model giving acoustic values:
    - Duration
    - F0 (pitch)
    - Learning
  - Model giving tones:
    - And duration
    - Pre-labelling
    - Learning

TAL-IDMC-OS-UL-2021-22

21

From text...

NLP: Natural Language Processing

Exeample : MBROLA .pho
[Phoneme] [Duration] [FO_values_per_segment]

l	58	0 116 36 116 74 118
@	51	14 119 57 118
g	69	0 118 70 113
R	84	33 113 90 115
a~	90	46 120
p	72	
R	66	0 130 42 125 86 120
e	78	26 120 63 115
z	74	0 114 39 115 78 116
i	71	18 115 59 111
d	84	0 108 40 106 82 106
a~	90	22 102 61 97

22

From text... to speech

«Le grand président est arrivé à 8h10.»

[ARTICLE]	l @
[ADJECTIF]	g R a~
[NOM]	p R e z i d a~
[AUXILIAIRE]	e t
[PARTICPE PASSE]	a R i v e
[PREPOSITION]	a
[HORAIRE]	H i t 9 R d i s

l @ g R a~ p R e z i d a~ e t a R i v e a H i t 9 R d i s

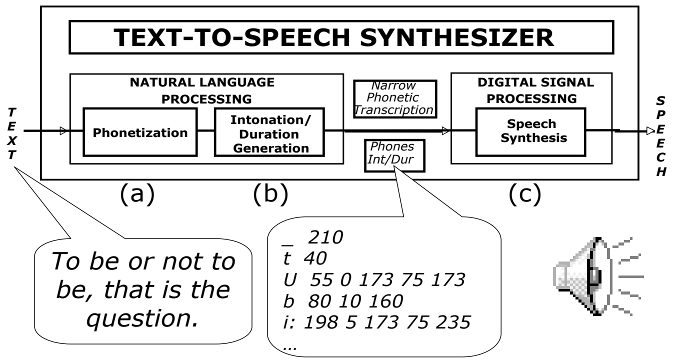
l	58	0 116 36 116 74 118
@	51	14 119 57 118
g	69	0 118 70 113
R	84	33 113 90 115
a~	90	46 120
p	72	
R	66	0 130 42 125 86 120
e	78	26 120 63 115
z	74	0 114 39 115 78 116
i	71	18 115 59 111
d	84	0 108 40 106 82 106
a~	90	22 102 61 97

- input is text
- output is a linguistic specification

- Waveform generation
  - **concatenation**, or
  - generate from a model

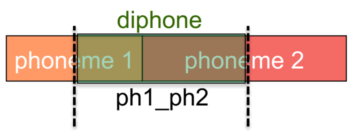
23

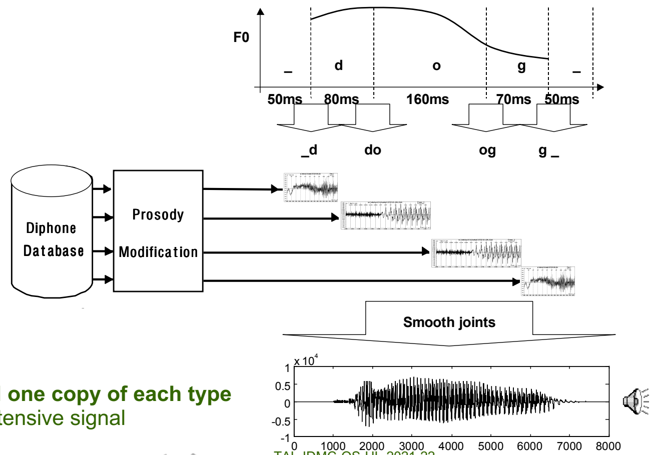
TTS = NLP + DSP



24

Diphone synthesis





Diphone units : record **one copy of each type**
Synthesis involves extensive signal manipulation

25

Problems with diphone synthesis

Signal processing is required to manipulate:

- **F0** & **duration**: fairly easy, within a limited range
- **Spectrum**: not so easy, can do simple smoothing at the joins but otherwise it's not obvious what aspects to modify

But, this extensive signal processing

- introduces artefacts and **degrades the signal**
- cannot faithfully replicate every detail of natural variation in speech
  - **what** to replicate
  - No powerful enough techniques to **manipulate** every aspect of speech

Solutions ?

- *Reduce* the need for manipulation by *increasing* the number of unit types?
- Cannot record and store versions of every speech sound in every possible context :
  - there are far too many
  - some will sound almost identical, so recording all of them is not necessary
- But we **can** have each speech sound in a **sufficient variety** of different contexts

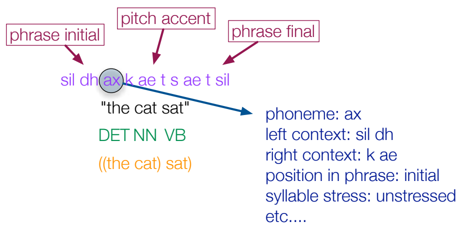
Unit-selection synthesis

- The key concept of unit selection speech synthesis:
  - record naturally-varying units, occurring in complete utterances
  - synthesis involves careful **selection** of appropriate units

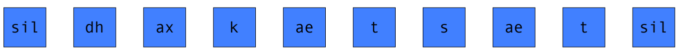
30

Unit-selection synthesis

From text .. Linguistic information (NLP)



The target unit sequence

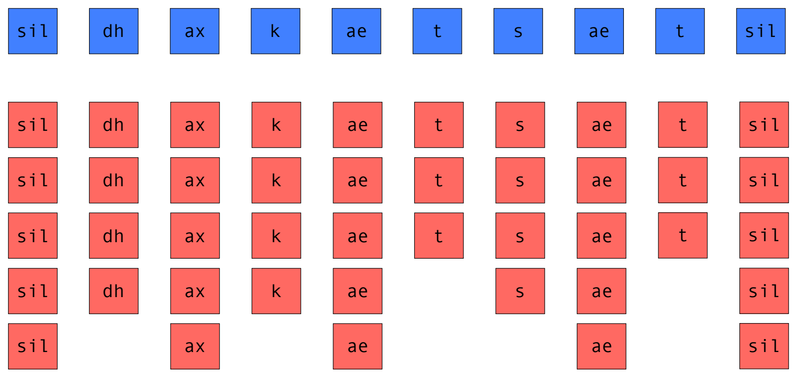


A sequence of **target** units : each unit is annotated with linguistic features for each target unit

31

Unit-selection synthesis

Retrieve candidate units from the pre-recorded database



Several **candidates** (incl. waveforms) : each candidate is annotated with linguistic features

Find the best-sounding sequence of candidates
- measuring **similarity** using the **target cost function**
- measuring **concatenation** quality using the **join cost function**

Similarity

- assume that units from similar linguistic contexts will sound similar
  - target cost function measures linguistic feature **mismatch**
  - target cost function measures acoustic **distance** between candidates and targets

Concatenation

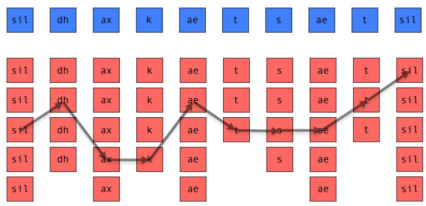
- The join cost measures the **acoustic mismatch** between two candidate units
- A typical join cost quantifies the acoustic mismatch across the concatenation point

Which candidate sequence will sound best? The one with lowest cost !

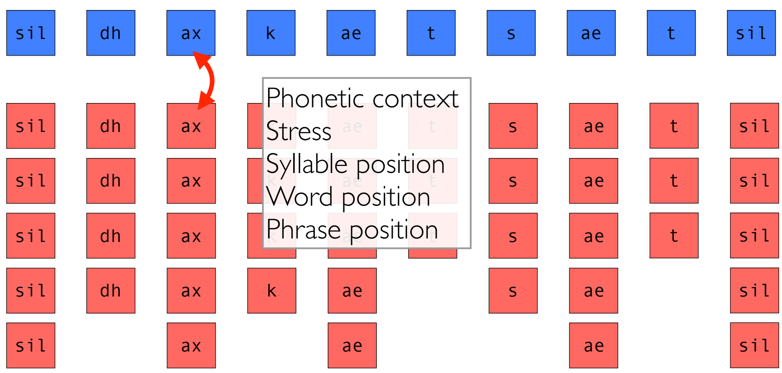
$$C = w_{tc}\ TC + w_{ajc}\ JC$$

- TC = target cost*
- JC = join cost

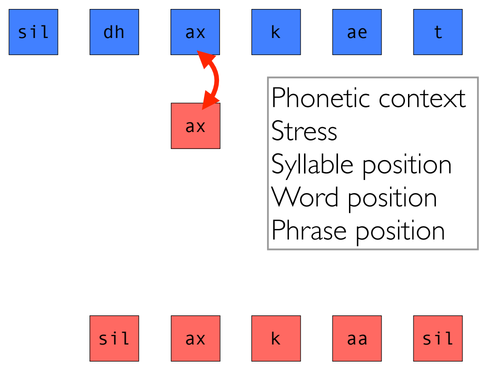
- **Dynamic programming search**



Target cost



Target cost



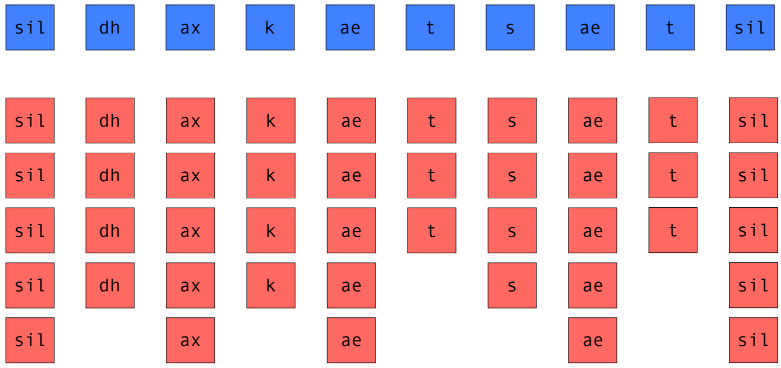
feature	weight
stress	10
syllable position	5
word position	5
POS	6
phrase position	7
left phonetic context	4
right phonetic context	3
bad F0	25
duration outlier	10

36

feature	weight	*target*	candidate 1	candidate 2
stress	10	*primary*	primary	none
syllable position	5	*coda*	onset	coda
word position	5	*final*	final	final
POS	6	*noun*	noun	verb
phrase position	7	*initial*	*initial*	*initial*
left context	4	*[b]*	[b]	[v]
right context	3	*[s]*	[w]	[s]
		target cost =		

37

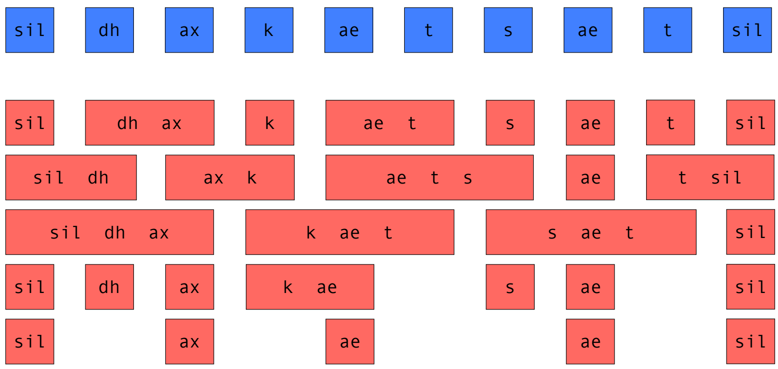
Homogeneous unit



TAL-IDMC-OS-UL-2021-22

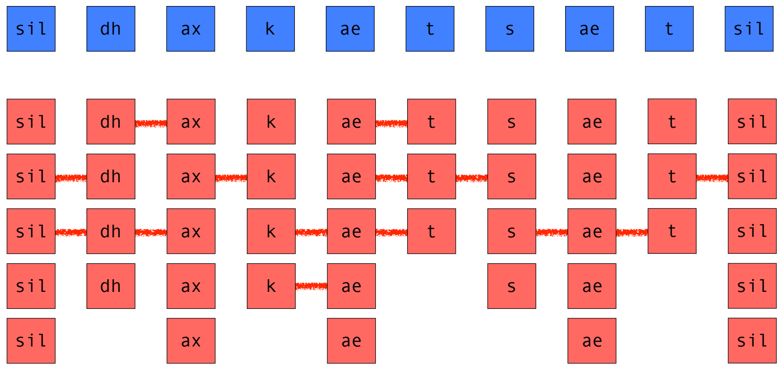
38

Heterogeneous units



TAL-IDMC-OS-UL-2021-22

39



Homogeneous unit type with "JC=0" = heterogeneous units !

40

Concatenative TTS

Advantages: intelligible

Limitations:

- require huge databases
- emotionless, not natural
- difficult to modify the voice (e.g., switching to a different speaker, or altering the emphasis or emotion) without recording a whole new database

42

Parametric TTS

How does it work?

– using learning based parametric models, e.g., HMM
– all the information required to generate speech is stored in the parameters of the model

Advantages: lower data cost and more flexible
Limitations: less intelligible than concatenative TTS

Neural TTS

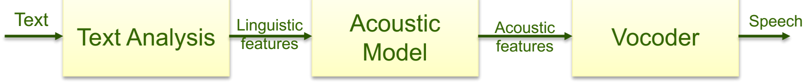
How does it work?

- a special kind of parametric models
- text to waveform mapping is modeled by (deep) neural networks

Advantages

– huge quality improvement, in terms of both intelligibility and naturalness
– less human preprocessing and feature engineering

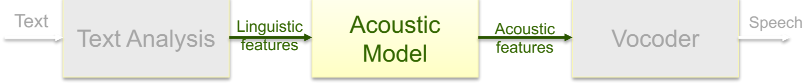
Components of parametric/neural TTS



- Text analysis: text→linguistic features *(identical to what we have seen before)*
- Acoustic model: linguistic features→acoustic features
- Vocoder: acoustic features→speech

Acoustic model

Generate acoustic features from linguistic features

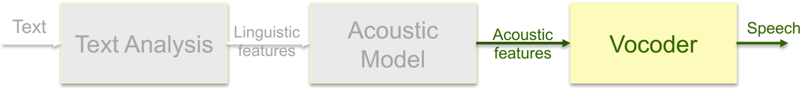


- F0, V/UV, energy
- Mel-scale Frequency Cepstral Coefficients (MFCC), Bark-Frequency Cepstral Coefficients (BFCC)
- Mel-generalized coefficients (MGC), band aperiodicity (BAP),
- Linear prediction coefficients (LPC),
- Mel-spectrograms
  - Pre-emphasis, Framing, Windowing, Short-Time Fourier Transform (STFT), Mel filter

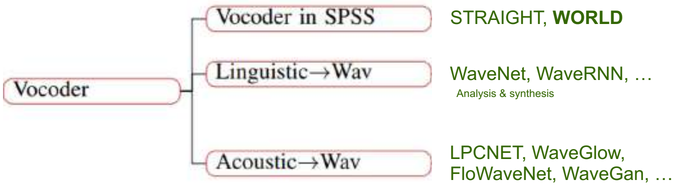
➔Better intelligibility when converting to speech

Vocoder

Generates speech waveform from acoustic features.



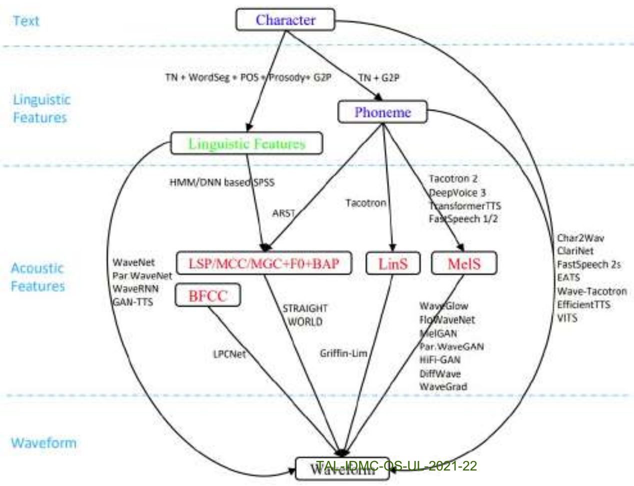
- Mel-spectrogram→ Vocoder → Speech



Mel Spectrogram ➔Better intelligibility when converting to speech

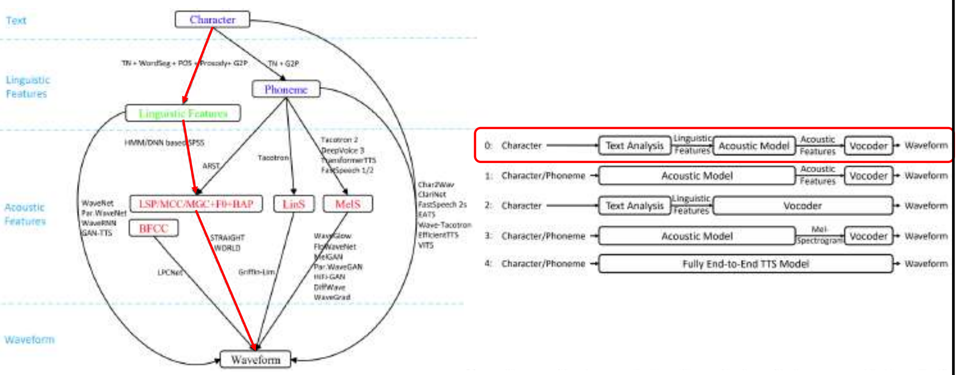
SPSS: statistical parametric speech synthesis

Neural TTSs – General scheme



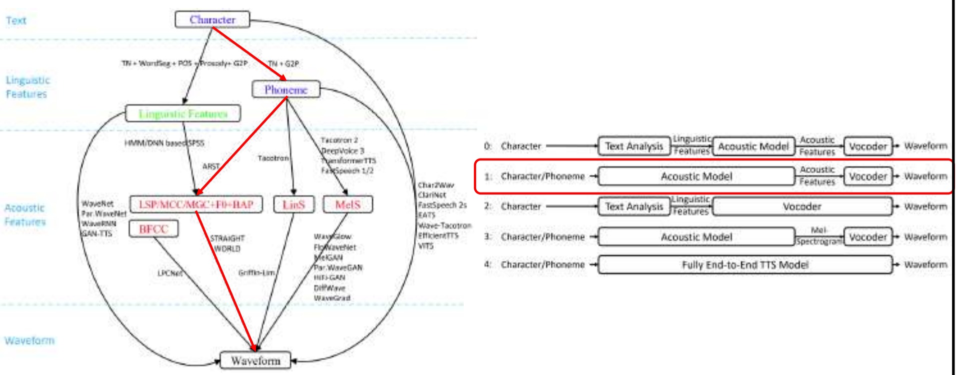
TAL-IDMC-CS-UL-2021-22

Neural TTSs – General scheme



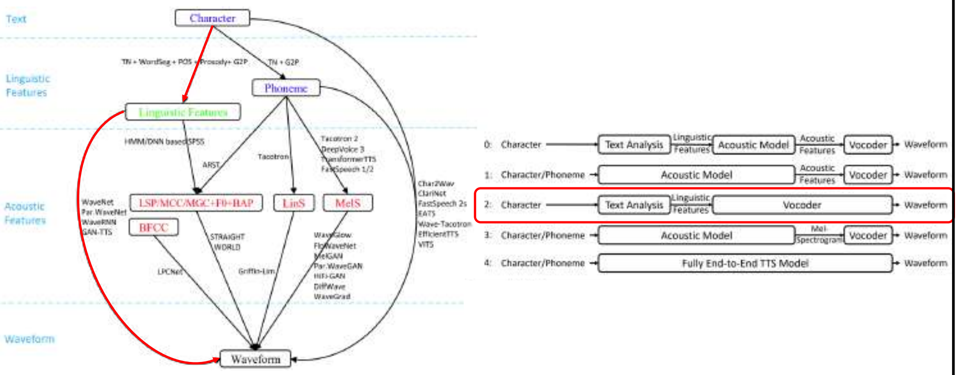
52

Neural TTSs – General scheme



53

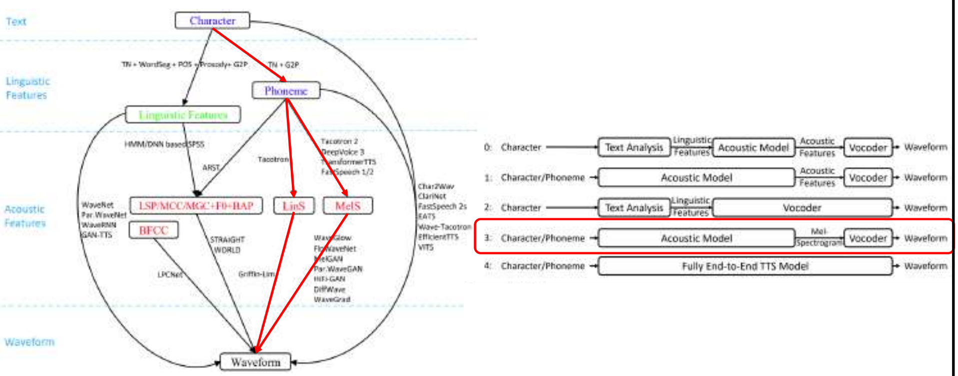
Neural TTSs – General scheme



TAL-IDMC-OS-UL-2021-22

54

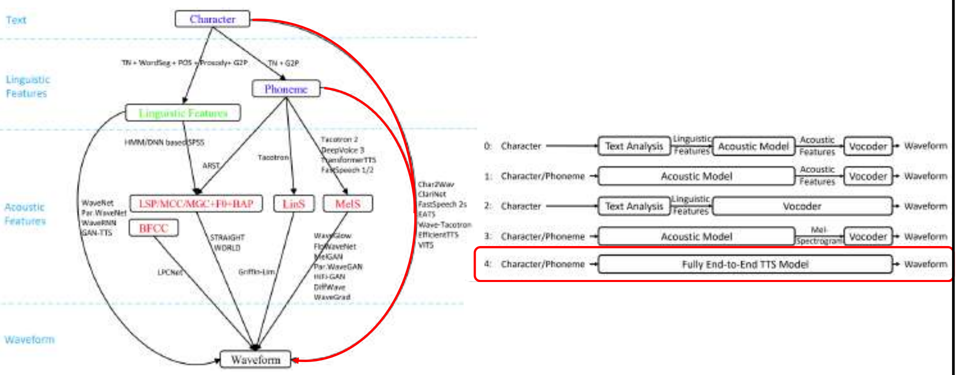
Neural TTSs – General scheme



TAL-IDMC-OS-UL-2021-22

55

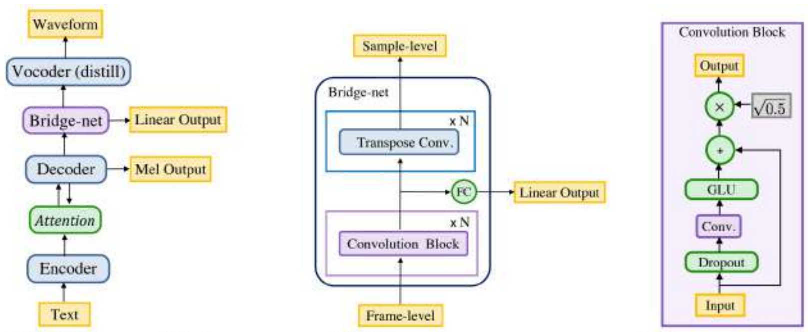
Neural TTSs – General scheme



Fully End-to-End TTS

ClariNet : an end-to-end text-to-speech architecture.

– It is a text-to-wave architecture that is fully convolutional and can be trained from scratch.



(a) Text-to-wave architecture (b) Bridge-net (c) Convolution block

Text-to-wave model converts textual features into waveform. All components feed their hidden representation to others directly. (b) Bridge-net maps frame-level hidden representation to sample-level through several convolution blocks and transposed convolution layers interleaved with softsign non-linearities. (c) Convolution block is based on gated linear unit.

Ping et al. (2019). ClariNet: Parallel Wave Generation in End-to-End Text-to-Speech – ICLR 2019

Audiovisual speech

- Is it important?

Audiovisual Intelligibility

- Importance of the ***visual*** channel when ***auditory*** channel is deteriorated. Sumby & Pollack (1954)

- Influence of visual on perception:
  Mc Gurk (1976) effect : audio **ba** + visual **ga** = audio-visual **da**

ba

ga

Audio-visual Perception **da**

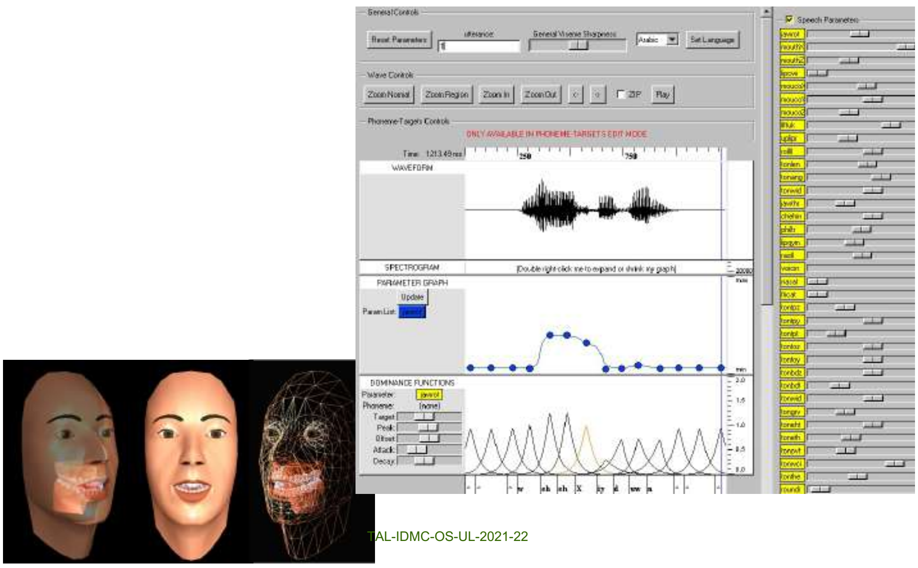
Audiovisual speech synthesis

PARAMATERIC TALKING HEAD

TAL-IDMC-OS-UL-2021-22

66

Articulation and coarticulation Prototyping Software



TAL-IDMC-OS-UL-2021-22

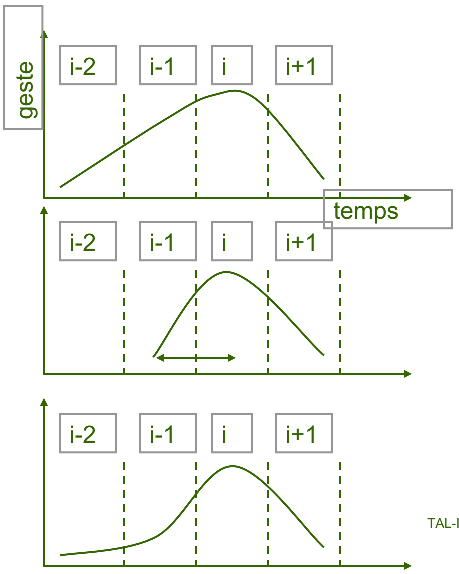
67

Coarticulation

- **Coarticulation** expresses the fact that a speech sound is influenced by a preceding or following speech sound.
- There are two types of coarticulation:
  - ***anticipatory coarticulation***, when a feature or characteristic of a speech sound is anticipated (assumed) during the production of a preceding speech sound;
  - ***carryover*** or *perseverative coarticulation*, when the effects of a sound are seen during the production of sound(s) that follow.

68

3 bilabial coarticulation models



1. Look-ahead
2. Time-locked
3. Hybrid

69

Modeling coarticulation

TAL-IDMC-OS-UL-2021-22

70

Coarticulation
Dominance functions

• A good example of coarticulation is the word “stew”. Initial consonants take on the lip protrusion of the upcoming vowel.

• A speech segment is a gesture with a certain dominance which increases and then decreases over time.

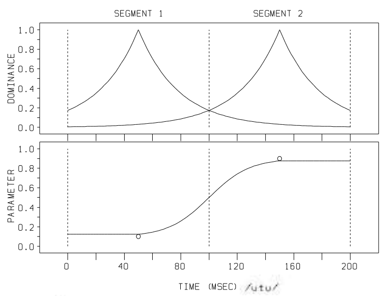
• **Dominance functions** of nearby segments overlap, and control parameters are calculated as a combination of segment target values, weighted by segment dominance functions.

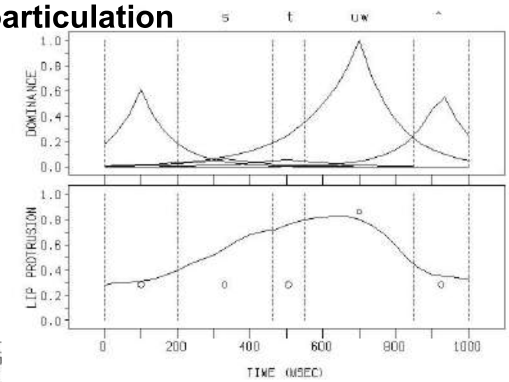
• Different dominance functions exist for different articulators, such as jaw rotation, tongue tip height, and lip rounding.

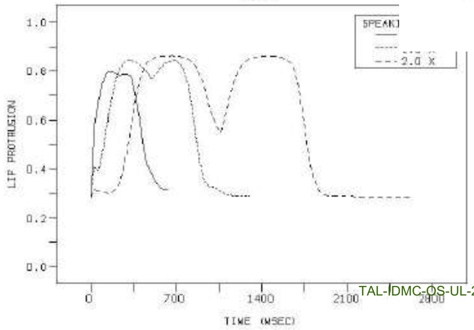
TAL-IDMC-OS-UL-2021-22

71

Coarticulation





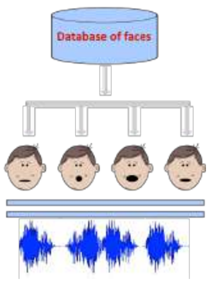


72

Acoustic-visual synthesis using bimodal unit-selection

Classical approach

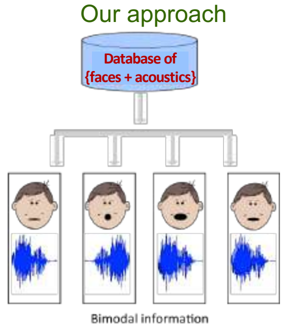
Animated faces generated from a database



Animation synchronized with acoustic speech

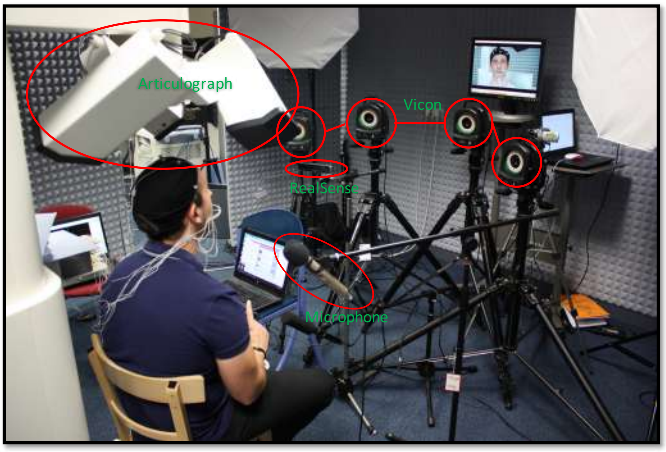
Audiovisual speech synthesis

Our approach



73

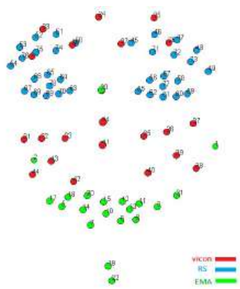
Multimodal acquisition system



TAL-IDMC-OS-UL-2021-22

75

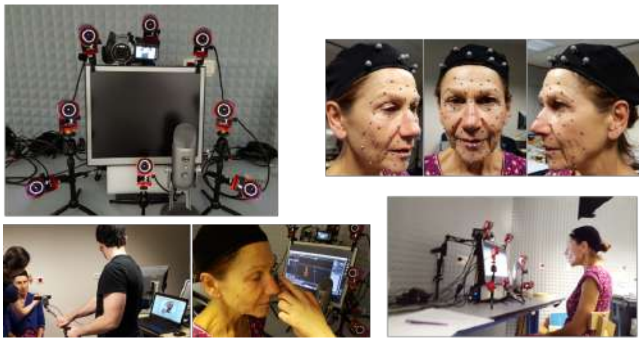
Multimodal acquisition system



TAL-IDMC-OS-UL-2021-22

76

Multimodal acquisition system



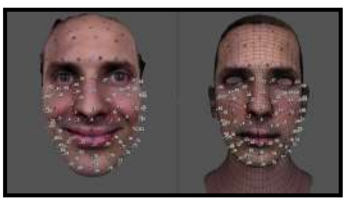
TAL-IDMC-OS-UL-2021-22

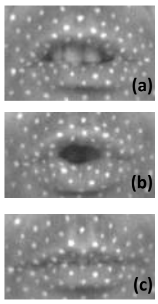
77

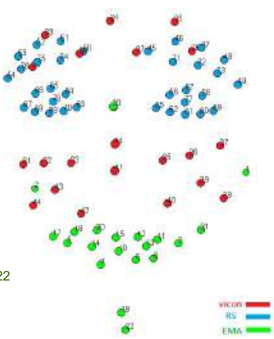
What visual data

wrt speech

- **All about lips..**

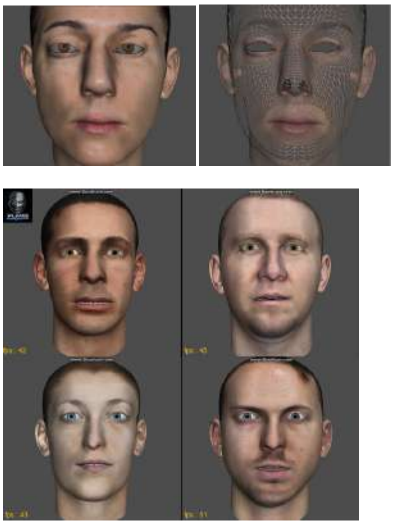


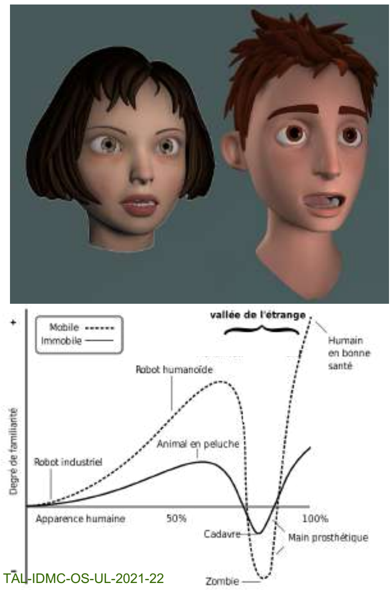




TAL-IDMC-OS-UL-2021-22

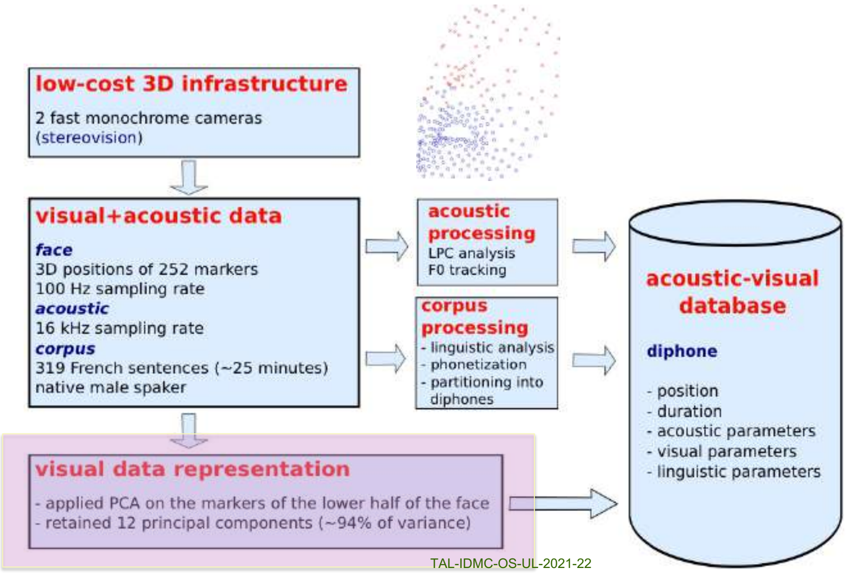
78





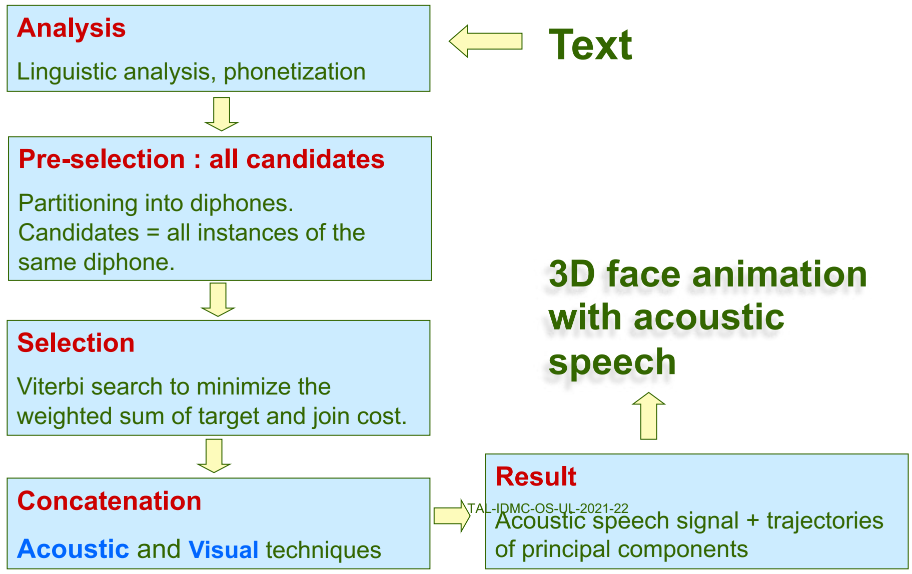
82

Data Acquisition and Database Preparation



83

Acoustic-Visual Unit-Selection Synthesis



85

Selection

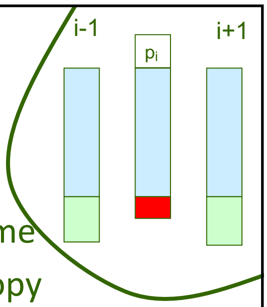
- Dynamic programming
  - Minimize **target** and **join** cost
- Acoustic-visual units

$$C = w_{tc}\, TC + (w_{ajc}\, AJC + w_{jvc}\, JVC)$$

  - TC = target cost*
  - AJC = acoustic join cost
  - VJC = visual join cost

*without *visual* information in the target cost (until now)

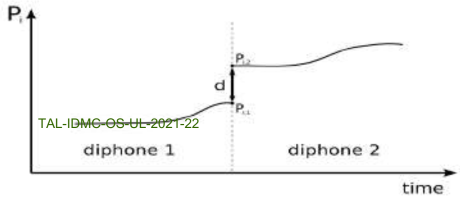
86



Selection (2)

- **Target** Cost
  - Weights for each feature and phoneme
  - Automatic learning : clustering, entropy (information gain)
  - Duration *w.r.t.* the position (intra-corpus)
- **Acoustic join** cost
  - F0, spectral distance, energy, dF0,…
- **Visual join** cost

$$VJC = \sum_{i=1}^{12} w_i \left(P_{i,1} - P_{i,2}\right)^2$$



87

Audiovisual speech synthesis using DNN

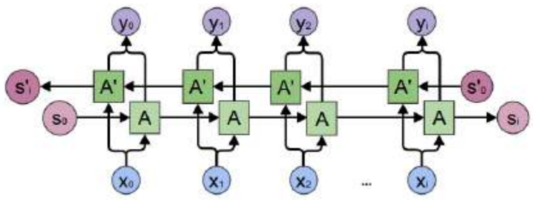


Figure 1: Bidirectional RNN processing a sequence x, where A is the forward layer and A' the backward layer.

text ➔ linguistic specification

linguistic specification ➔ acoustic features

linguistic specification ➔ visual features

acoustic features ➔ waveform

visual features ➔ animation

92

Neural Architecture

Duration fully connected network :

Inputs:
- 1 parameter, the duration of the current phone
- 41 size phone labels vector
- 7 size emotion labels vector

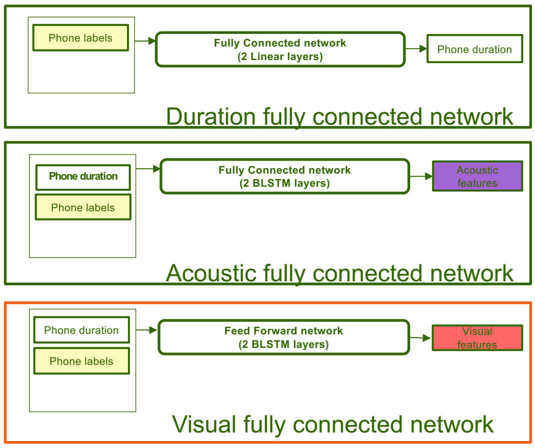
Acoustic fully connected network:

Inputs:
- 187 acoustic parameters (default Merlin parameters):
60 mel-cepstra coefficients, 1 band aperiodicities, log-f0, and a binary voicing decision with delta and delta-delta features

- 41 size phone labels vector
- 7 size emotion labels vector

Visual fully connected network:

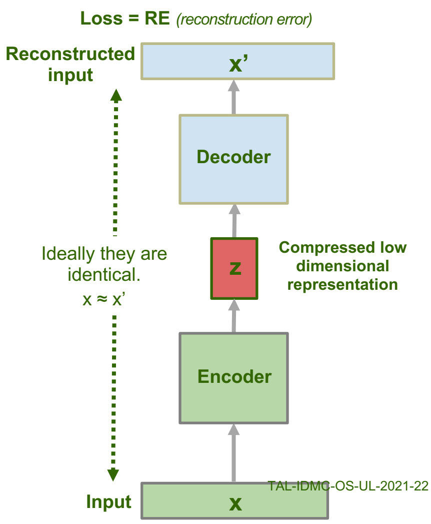
Inputs:
- **132 visual parameters (3 (x,y,z) x 44 markers)**
- **41 size phone labels vector**
- **7 size emotion labels vector**



93

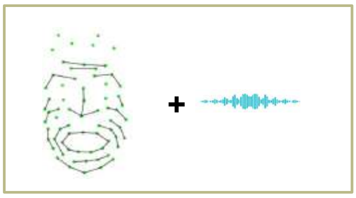
Audiovisual speech synthesis using DNN

Auto Encoder



95

Expressive audiovisual TTS



	sentences	Duration
Neutral	2000	4h 02min
Joy	500	58min
Sadness	500	1h 10min
Disgust	500	1h 53min
Surprise	500	1h 04min
Anger	500	55min
Fear	500	1h 11min

Post-processing:

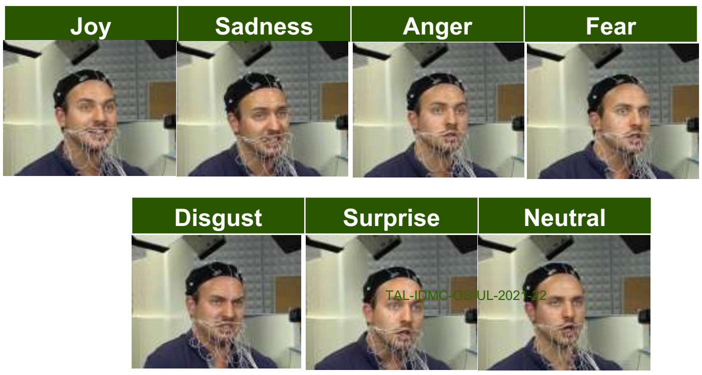
Remove head mvt → Spatial alignment → Text-Audio alignment →

Expressive Corpus acquisition

Technique *exercise in style*

- The same set of sentences uttered in different emotions (joy, sadness, anger, fear, disgust, surprise) + neutral.

➔The linguistic content of the sentences does not help to identify the expressed emotion (dissociation between semantics and syntax).



Neural Architecture

Duration fully connected network :

Inputs:

- 1 parameter, the duration of the current phone
- 41 size phone labels vector
- 7 size emotion labels vector

Acoustic fully connected network:

Inputs:

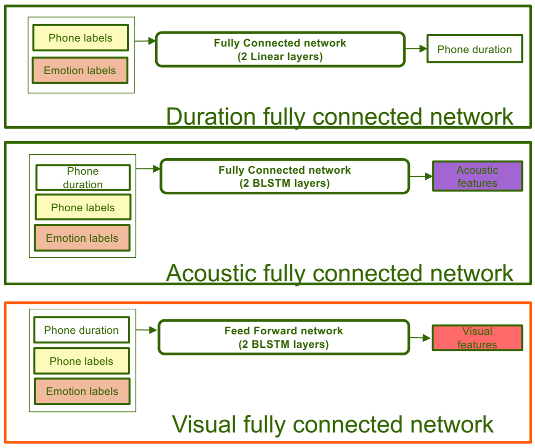
- 187 acoustic parameters (default Merlin parameters):
60 mel-cepstra coefficients, 1 band aperiodicities, log-f0, and a binary voicing decision with delta and delta-delta features

- 41 size phone labels vector
- 7 size emotion labels vector

Visual fully connected network:

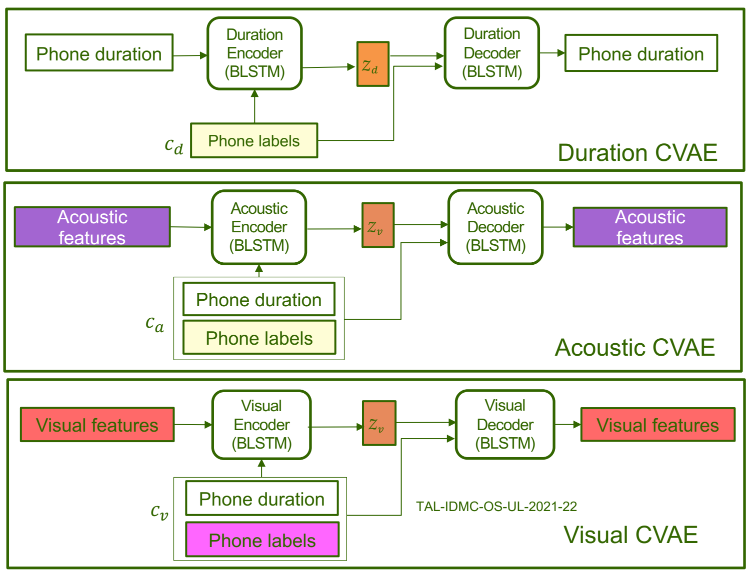
Inputs:

- **132 visual parameters (3 (x,y,z) x 44 markers)**
- **41 size phone labels vector**
- **7 size emotion labels vector**



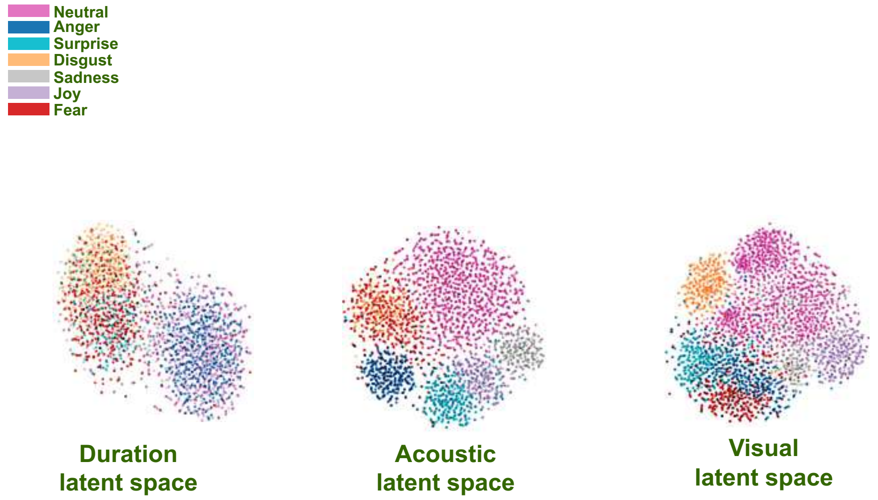
TAL-IDMC-OS-UL-2021-22

100



101

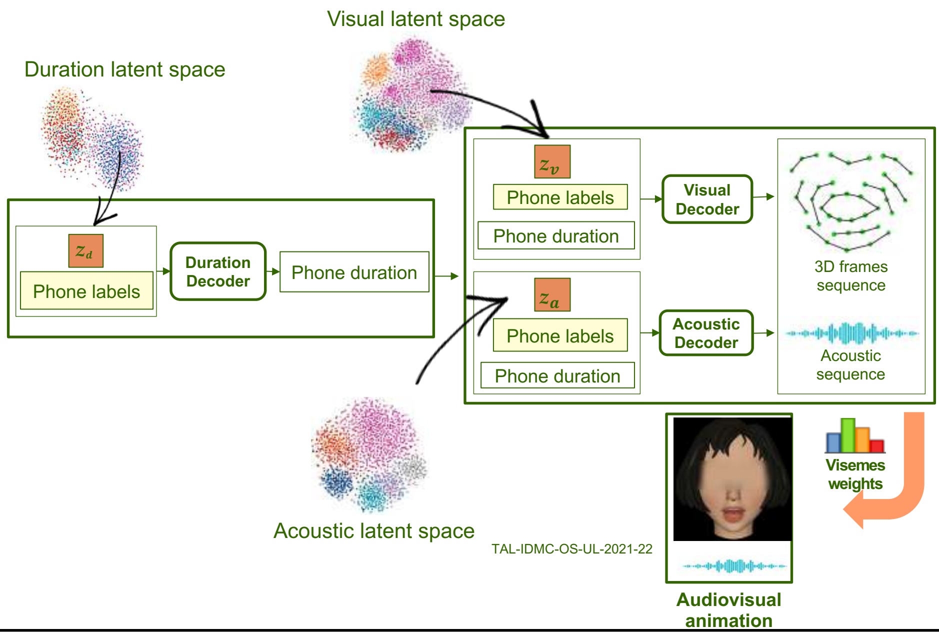
Neural architecture – t-SNE plot of latent spaces



TAL-IDMC-OS-UL-2021-22

102

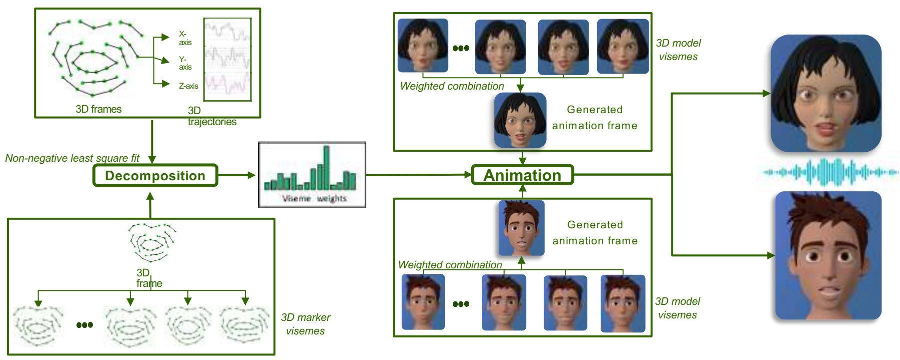
Neural architecture – Expressive Synthesis process



TAL-IDMC-OS-UL-2021-22

103

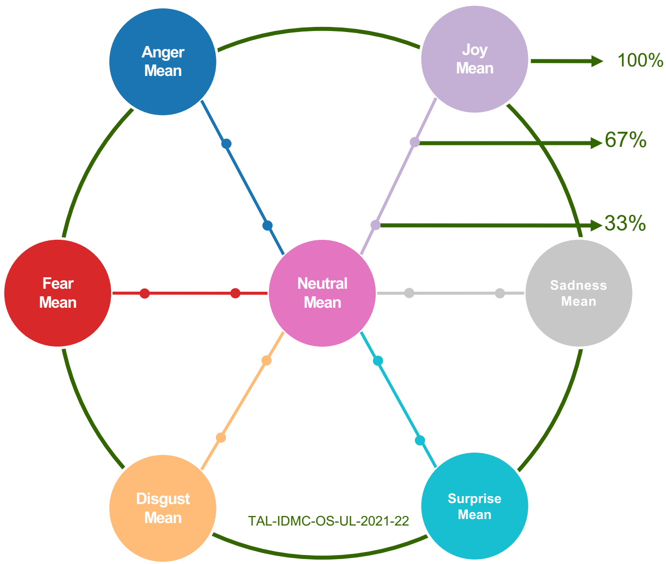
3D animation: from sparse space to blendshapes



TAL-IDMC-OS-UL-2021-22

104

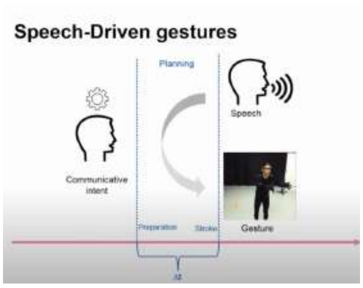
Varying the intensity of the generated emotion

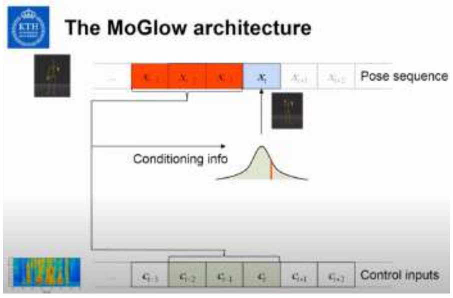


TAL-IDMC-OS-UL-2021-22

105

Towards spoken, visual, expressive and gestural communication





Trinity Gesture Dataset
- One male actor
- 244 minutes mocap & audio

https://www.youtube.com/watch?v=4_Gq9rU_yWg

TAL-IDMC-OS-UL-2021-22

Alexanderson, S., Székely, É., Henter, G. E., Kucherenko, T., & Beskow, J. (2020, October). Generating coherent spontaneous speech and gesture from text. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents* (pp. 1-3).

109

Towards spoken, visual, expressive and gestural communication

When gesture is the main means of communication

➔ Sign Language

N. Camgoz et al., "Neural sign language translation" 2018.

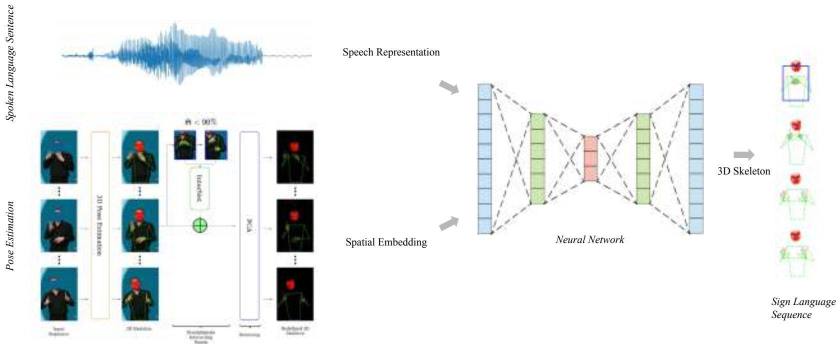
TAL-IDMC-OS-UL-2021-22

110

Towards spoken, visual, expressive and gestural communication

Sign language synthesis

Our current work (multispeech/LORIA)



TAL-IDMC-OS-UL-2021-22

111