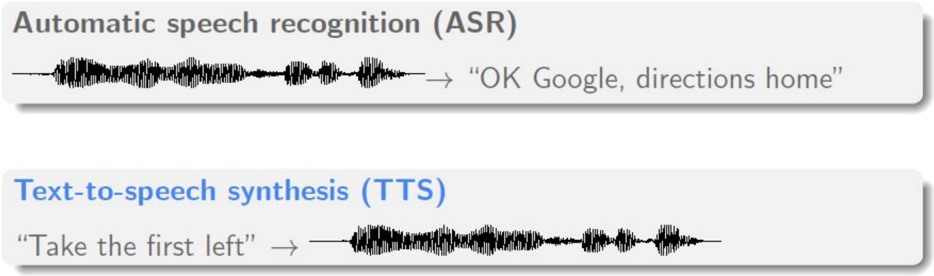
Speech synthesis

Chloé Clavel, Telecom-Paris

Speech synthesis or TTS (Text To Speech)

Objective :

- be able to read any written text



Automatic speech recognition (ASR)

→ “OK Google, directions home”

Text-to-speech synthesis (TTS)

“Take the first left” →

Speech synthesis or TTS (Text To Speech)

Purpose different from talking machines

- Talking machine = concatenation of word/phrase records
- Ex: Pre-recorded voices from the metro or talking clocks

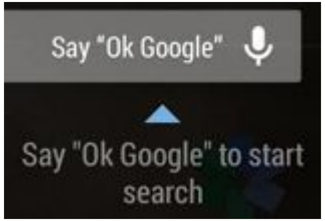
> “au quatrième top il sera exactement :
>
> (nombre inséré) heure (nombre inséré) minutes (nombre inséré) secondes

- Constraints :
  - limited vocabulary,
  - sentence with fixed structures,
  - reasonable size of the record base

Examples of applications

- Telecommunications services
- Cinema, traffic and bank account information
- Reading of SMS, emails for blind people
- Learning foreign languages
- Oral conversation with an animated conversational agent (open source software: Open Mary)
- Virtual Assistants
- Companion robots
- Assistants virtuels
- Robots compagnons

What can I help you with?



Say "Ok Google"

Say "Ok Google" to start search

Virtual assistant

Production of the voice signal: the ancestors

Von Kempelen (1791)

Manual voice synthesizer

Reproduction of the human vocal tract

Production of vocal sounds

Ex : "nostrils" (narines) for making the '
"m” sounds

levers (leviers) and tubes dedicated to '
"sh" sounds

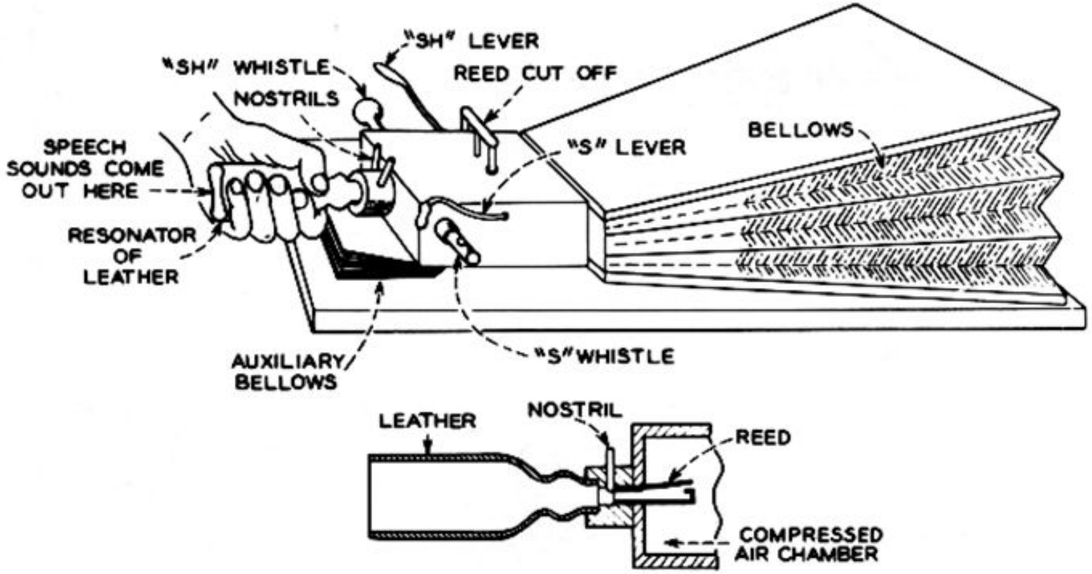


Fig. 10. Wheatstone's reconstruction of von Kempelen's speaking machine.[1]

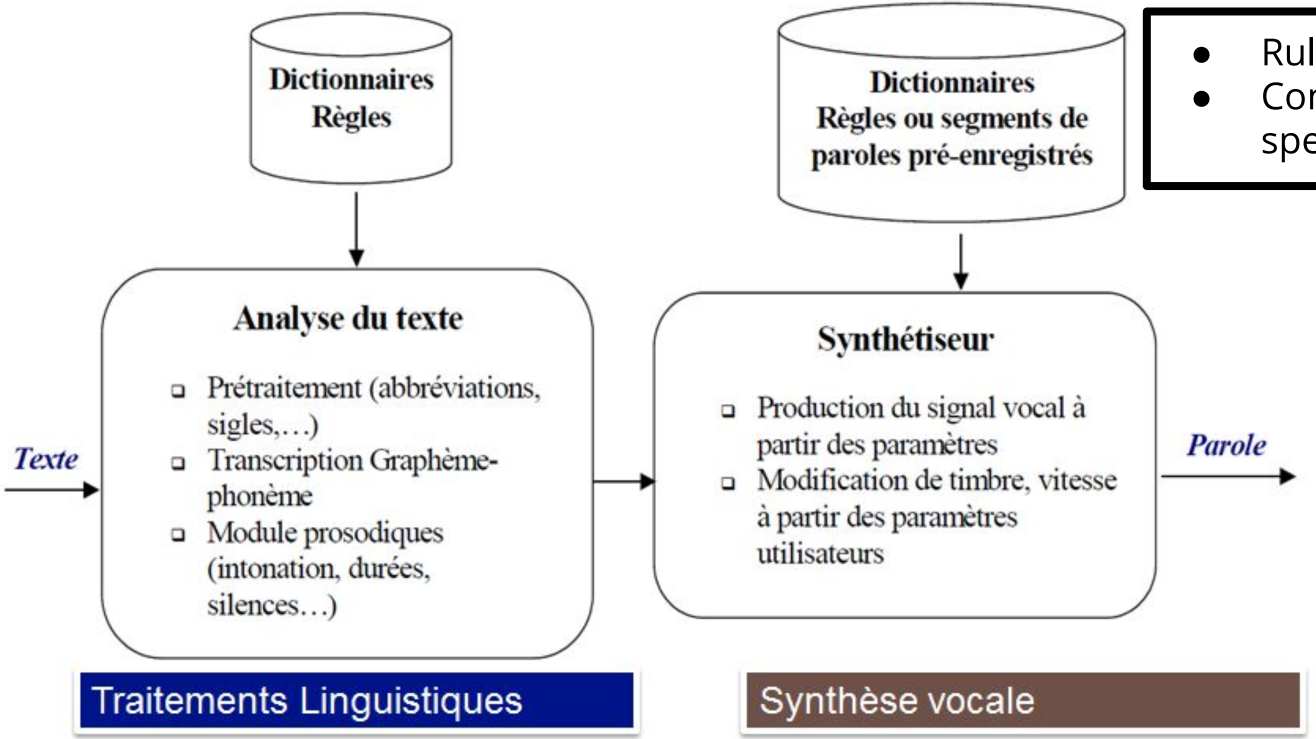
The Journal of the Acoustical Society of America

Different approaches

1. Rule-Based Approaches [Klatt, 1980]
2. Synthesis by concatenation [Hunt & Black, 1996].
3. Parametric synthesis: generative synthesis from models (HMM) [Zen et al., 2009].
4. Deep learning approaches (WaveNet) [Van Den Oord et al., 2016].

Examples :
https://deepmind.com/blog/article/wavenet-generative-model-raw-audio

Classical architecture of a TTS



Dictionnaires
Règles

Dictionnaires
Règles ou segments de paroles pré-enregistrés

Texte

Analyse du texte

- Prétraitement (abbréviations, sigles,…)
- Transcription Graphème-phonème
- Module prosodiques (intonation, durées, silences…)

Synthétiseur

- Production du signal vocal à partir des paramètres
- Modification de timbre, vitesse à partir des paramètres utilisateurs

Parole

Traitements Linguistiques

Synthèse vocale

- Rule-based TTS
- Concatenative TTS : short speech fragments

Analogy with the mode of speech production

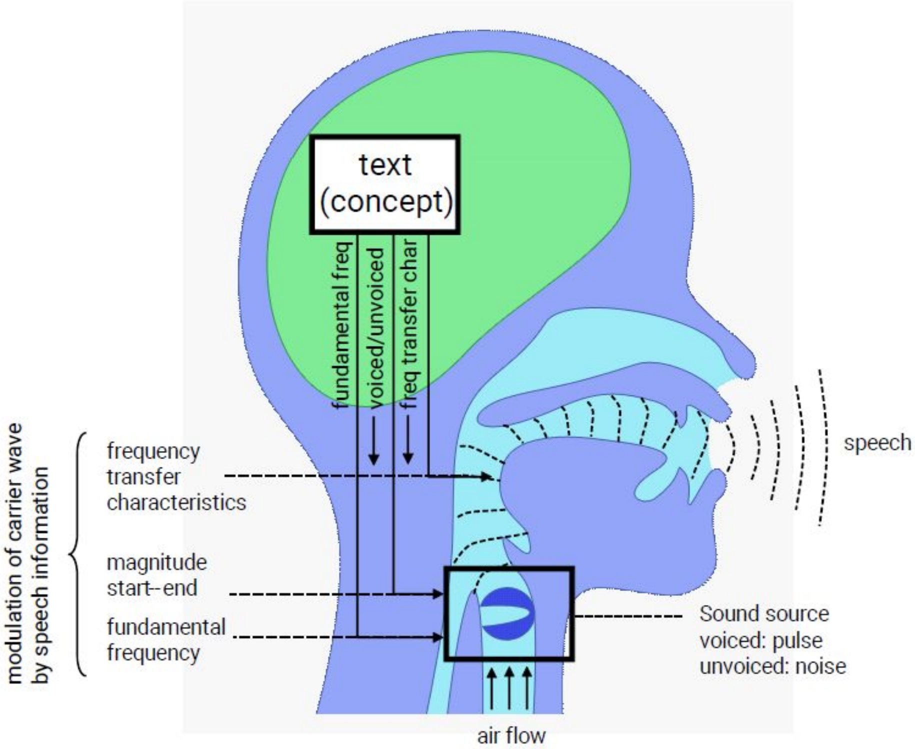
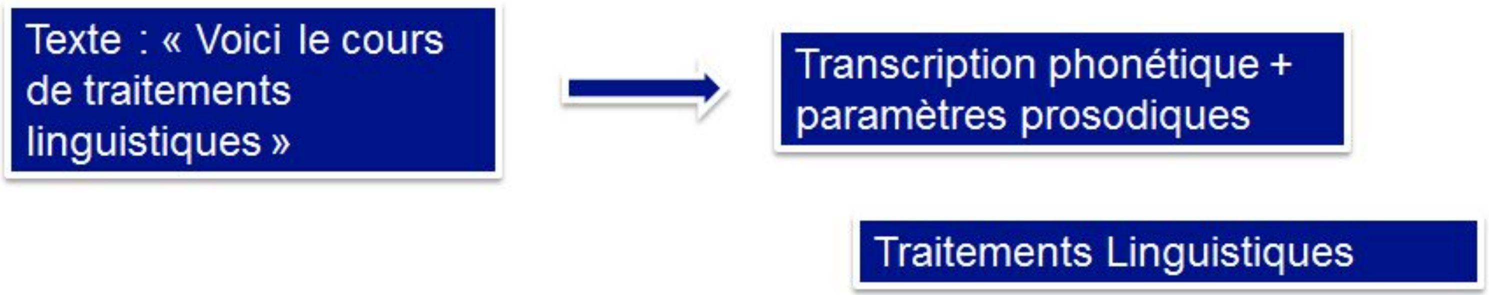


Figure tirée de
[1]

Architecture of a two-block TTS system

Text analysis / natural language processing

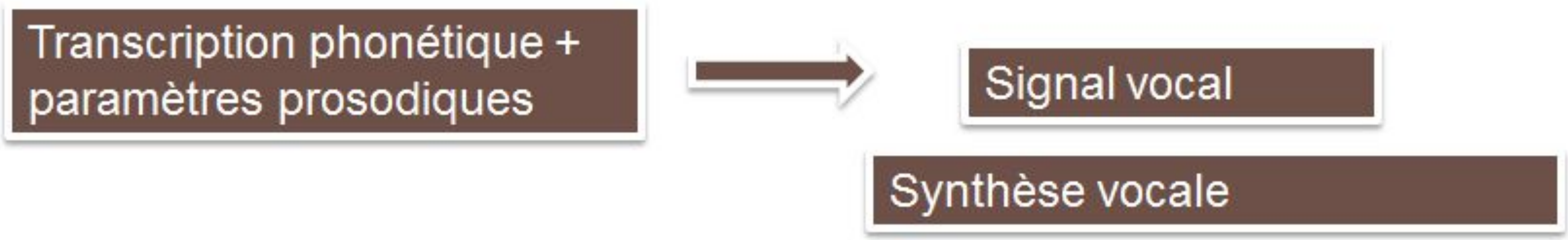


Texte : « Voici le cours de traitements linguistiques »

Transcription phonétique + paramètres prosodiques

Traitements Linguistiques

Synthesis : production of vocal signal



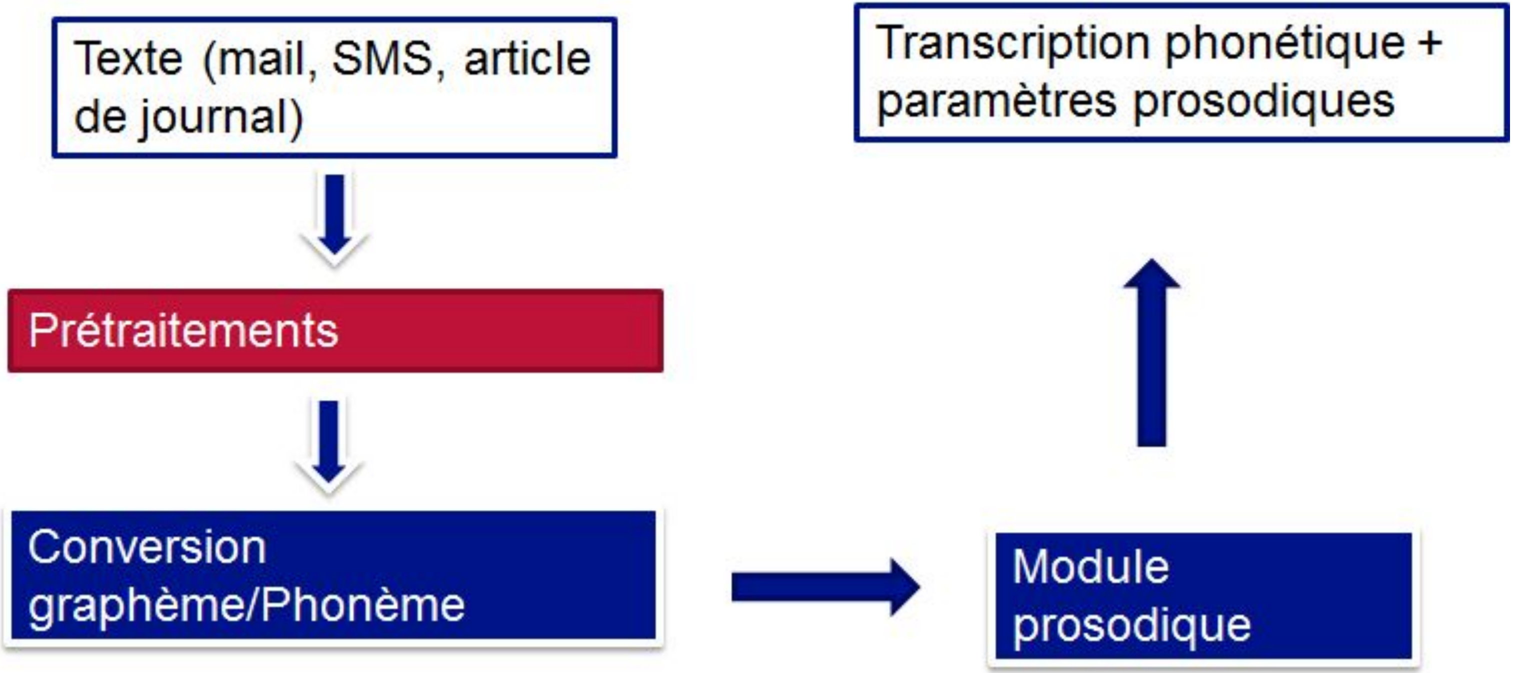
Transcription phonétique + paramètres prosodiques

Signal vocal

Synthèse vocale

Text Pre-processing

The bricks of the language processing block



Texte (mail, SMS, article de journal)

Prétraitements

Conversion graphème/Phonème

Module prosodique

Transcription phonétique + paramètres prosodiques

Preprocessing and Linguistic processing

Shared steps by all the different approaches :

- Rule-Based Approaches [Klatt, 1980]
- Synthesis by concatenation [Hunt & Black, 1996].
- Parametric synthesis: generative synthesis from models (HMM) [Zen et al., 2009].
- Deep learning approaches (WaveNet) [Van Den Oord et al., 2016] : sequence of linguistic and phonetic features (which contain information about the current phoneme, syllable, word, etc.) fed into WaveNet.

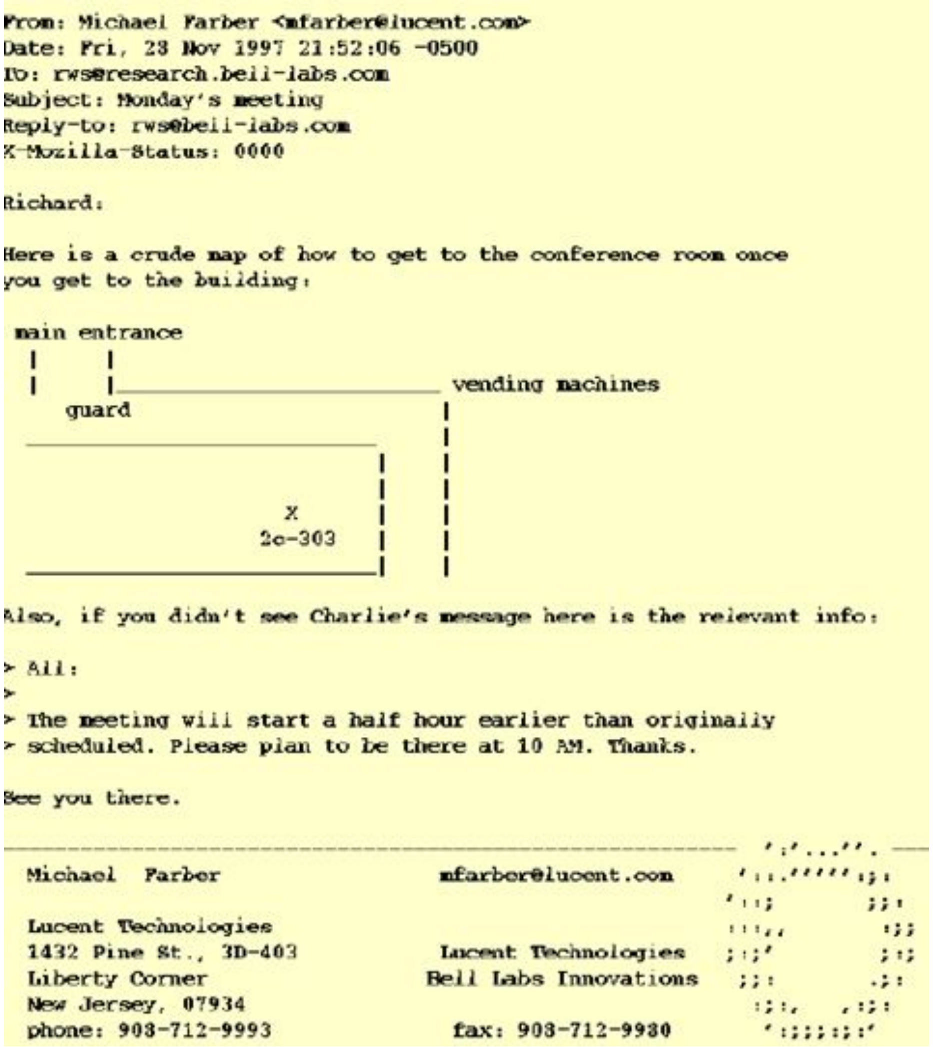
Preprocessing -example

Voice synthesis of mail for the blind or by a virtual assistant

How can we reproduce the process used by humans to read an email?

What are the different stages?

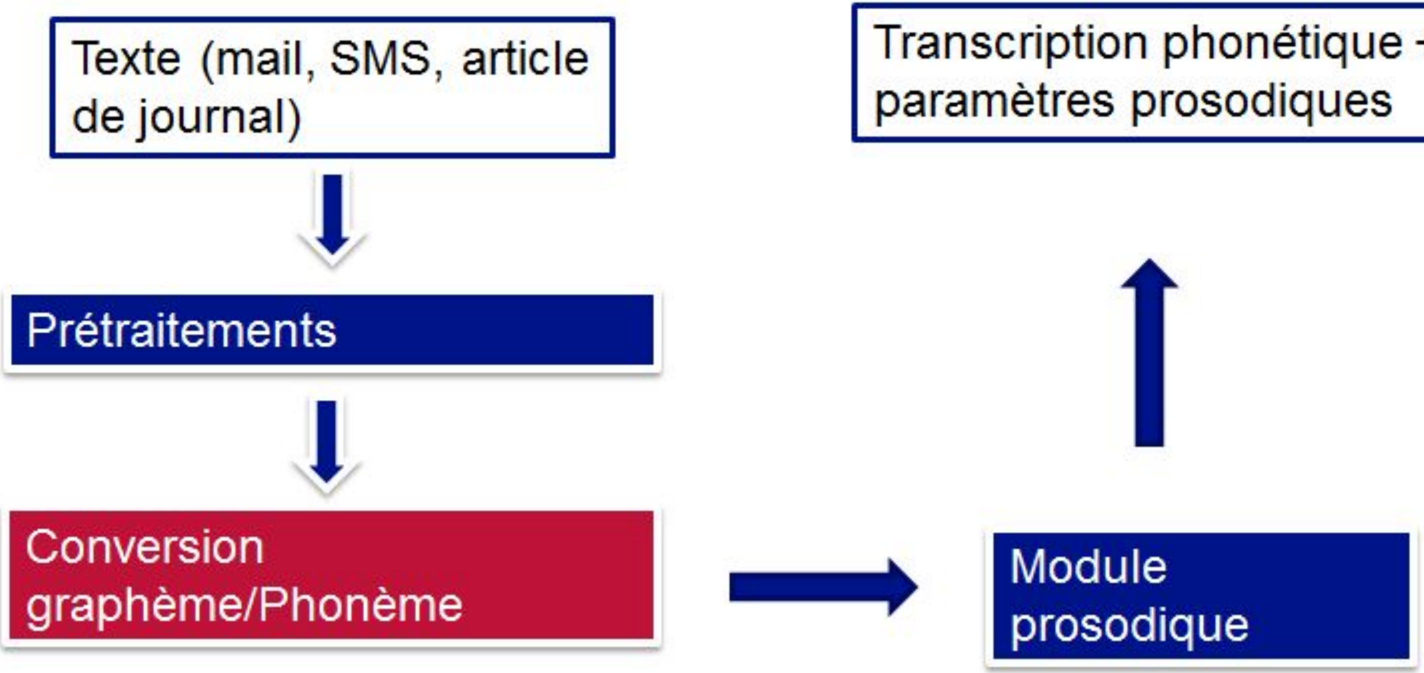
*Mail structure analysis
*Identification of problematic character sequence
*Language identification



```
From: Michael Farber <mfarber@lucent.com>
Date: Fri, 28 Nov 1997 21:52:06 -0500
To: rws@research.bell-labs.com
Subject: Monday's meeting
Reply-to: rws@bell-labs.com
X-Mozilla-Status: 0000

Richard:

Here is a crude map of how to get to the conference room once
you get to the building:

 main entrance
  |     |
  |     |_______________________  vending machines
     guard                        |
  _______________________         |
                          |       |
                          |       |
                     X    |       |
                  2c-303  |       |
  ________________________|       |

Also, if you didn't see Charlie's message here is the relevant info:

> All:
>
> The meeting will start a half hour earlier than originally
> scheduled. Please plan to there at 10 AM. Thanks.

See you there.

-------------------------------------------------------  ';'...''. --
  Michael   Farber                  mfarber@lucent.com   ':;.''''';;:
                                                        ':;;        ;;:
  Lucent Technologies                                   ::;,,        :;;
  1432 Pine St., 3D-403             Lucent Technologies  ;:;'         ;:;
  Liberty Corner                  Bell Labs Innovations   ;;:        .;:
  New Jersey, 07934                                        :;:,    ,:;:
  phone: 908-712-9993                fax: 908-712-9930      ':;;;:;:'
```

Linguistic processing
- grapheme-phoneme conversion-

The bricks of the language processing block



Texte (mail, SMS, article de journal)

Prétraitements

Conversion graphème/Phonème

Module prosodique

Transcription phonétique + paramètres prosodiques

Principle of grapheme-phoneme conversion

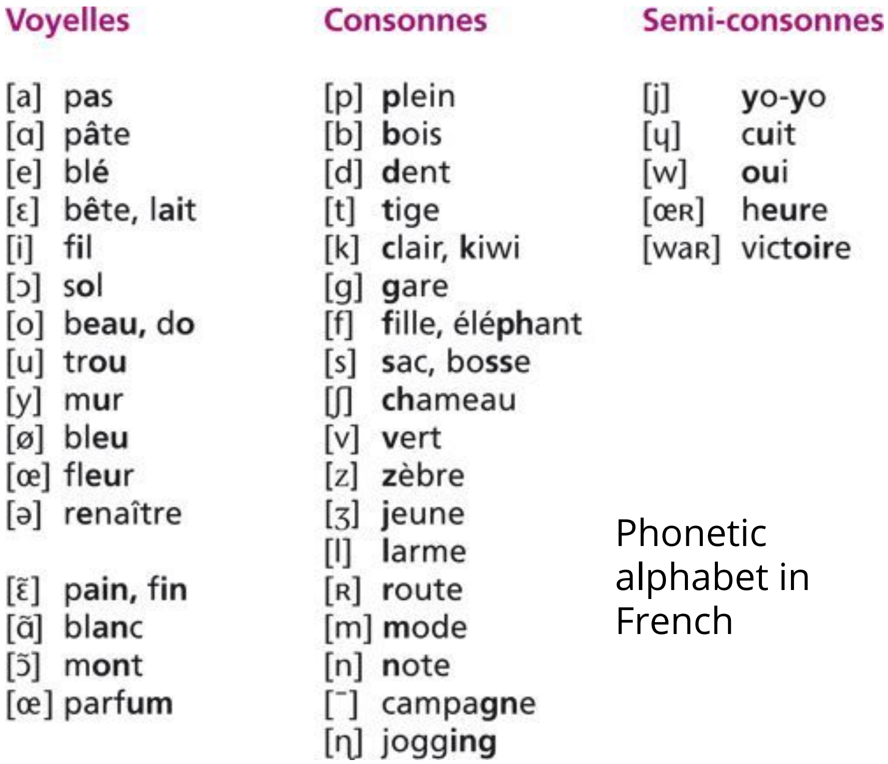
orthographic text (grapheme)

↓

phonetic text

(sequence of phonemes)

ex : /vwaʁ/)



Voyelles		Consonnes		Semi-consonnes	
[a]	pas	[p]	plein	[j]	yo-yo
[ɑ]	pâte	[b]	bois	[ɥ]	cuit
[e]	blé	[d]	dent	[w]	oui
[ɛ]	bête, lait	[t]	tige	[œʀ]	heure
[i]	fil	[k]	clair, kiwi	[waʀ]	victoire
[ɔ]	sol	[g]	gare		
[o]	beau, do	[f]	fille, éléphant		
[u]	trou	[s]	sac, bosse		
[y]	mur	[ʃ]	chameau		
[ø]	bleu	[v]	vert		
[œ]	fleur	[z]	zèbre		
[ə]	renaître	[ʒ]	jeune		
		[l]	larme		
[ɛ̃]	pain, fin	[ʀ]	route		
[ɑ̃]	blanc	[m]	mode		
[ɔ̃]	mont	[n]	note		
[œ]	parfum	[ˉ]	campagne		
		[ŋ]	jogging		

Phonetic alphabet in French

Principle of grapheme-phoneme conversion

Transcription +/- difficult depending on the language and complex for French

No bijection between letter space and phoneme space

Ex: A sequence of letters can have several pronunciations:

‘ch’ can be pronounced:

- /k/ in *chlore*

- /ʃ/ in *château*

Transcription of isolated words: how does it work?

Use of lexicons and morphological analysis to define pronunciation rules

s -> z / V __ V (in French, ex : oiseau)

Rule that does not work for *a+social* => Exception rule

s -> s / #a,#anti,#pro... __ V

Use of machine learning approaches on corpora (pronunciation dictionary)

Transcribe word in context

Syntactic analysis
Reduce the number of possible lexical categories for each word according to its neighbours.

Ex: one word -> several lexical categories (50% of the occurrences)

(2) La/DET/N/PRO

(3) couvent/N/V

A sentence of 20 words has 210 possible analyses

Syntactic analysis for transcription

Aim : disambiguate homographes-heterophones (words with the same spelling but with a different pronunciations)

est-V vs est-N,
couvent-V vs. couvent-N,
bus-V vs. bus-N,
violent-V vs. violent-A,
portions-V vs. portions-N,
fils-N vs. fils-N….

Transcribe word in context

Phonological adjustment - e-mute
/ə/ is an unstable phoneme, which may or may not be pronounced depending on its context.
Influence of accents, style, rate and context of speech, rhythmic factors...

Example of rules :

1//ə/ elides at the end of the word, and sometimes at the initial (renard, but not pelote, vedette)

2/ 3 consonant rule: an "e" is pronounced if its disappearance causes the pronunciation of 3 successive consonants (ex : table rouge)

Transcribe word in context

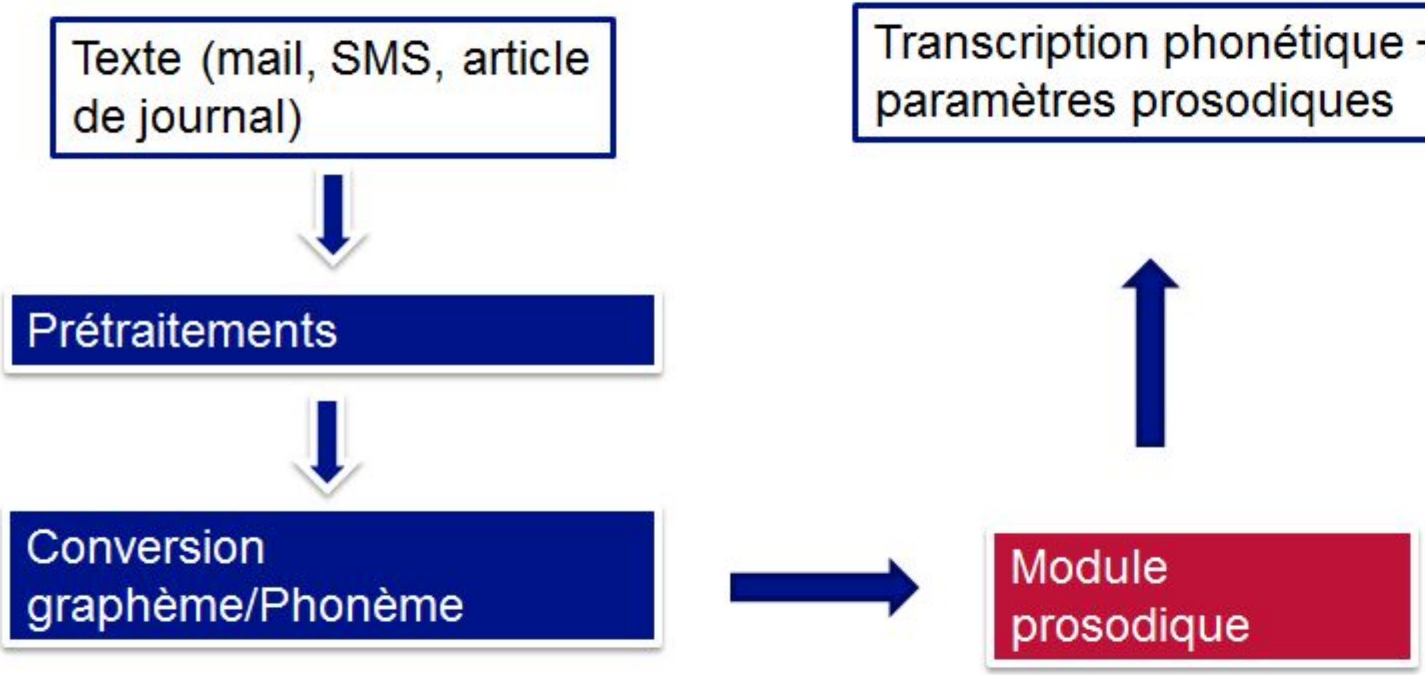
Phonological adjustment : the liaison (EX : “les enfants”) **-** Contextual realization of a latent consonant segment at the end of a word.

Build rules of pronunciation which depend on

- syntactic factors
  - ex : obligatory liaisons : DET+NOM (Ex : les enfants) *vs.* forbidden liaisons : NOM+VRB (Ex : les enfants ont …)
- phonological factors
  - ex: prefer CV pronunciation to hiatus (two adjacent vowels) by inserting a phoneme « les[z] enfants »)
- Inter-speaker and intra-speaker variability

Linguistic processing
- prosodic module-

The bricks of the language processing block



Texte (mail, SMS, article de journal)

Prétraitements

Conversion graphème/Phonème

Module prosodique

Transcription phonétique + paramètres prosodiques

Prosodic module : objectives

Calculate prosodic parameters (also called intonation symbols) automatically from the text

Supra-segmental speech characteristics :

Melody, Rhythm, Accent and Emphasis

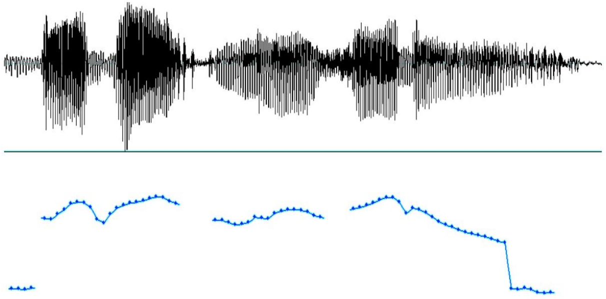
Three main parameters:

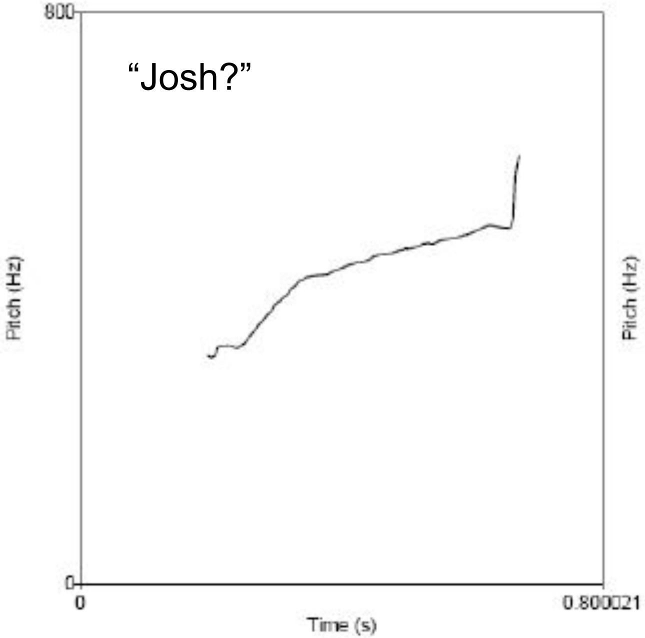
- Intonation: variations in the fundamental frequency...
- Duration: Segment and pause durations
- Intensity: a function of energy

Extraction of 4 types of symbolic information from the text :

1/Modality of the statement (declarative, imperative, interrogative) -> overall shape of the intonative curve

Ex : Declination: the melodic line generally decreases from the beginning to the end of a declarative sentence.





Extraction of 4 types of symbolic information from the text :

2/Identification of Prosodic Groups -> position of accents and pauses within the sentence, (re)initialization of the declination curve

Demarcative function of the prosody : from the sentence structure (dot, coma, chunk), we can decide where to put the pauses

Example of tricky cases : « Jacqueline (entend (le bruit de la fenêtre)) » vs. « Jacqueline (entend (le bruit) (de la fenêtre)) »

2/Identification of Prosodic Groups -› position of accents and pauses within the sentence, (re)initialization of the declination curve

Method:

- rules/heuristics
- morpho-syntactic analysis for sentence segmentation
- machine learning approaches on corpora labelled in pauses

Extraction of 4 types of symbolic information from the text :

1/Modality of the statement (declarative, imperative, interrogative)

2/Identification of Prosodic Groups

3/Identification of accented syllables within prosodic groups -> duration, energy, micro-melody, rhythm.

e.g. in English, each lexical word has a main emphasis on one syllable and the other syllables are unaccented

3/Identification of accented syllables within prosodic groups -> duration, energy, micro-melody, rhythm.

Lexical accent

In French: no lexical accent

In English, each lexical word has a main accent on one syllable and the other syllables are unaccented.

In Spanish, accent distinguishes words with different meanings : « ’termino » (le terme) « ter’mino » (je termine) « termi'no » (il a terminé)

-> distinctive function of prosody

3/Identification of accented syllables within prosodic groups -> duration, energy, micro-melody, rhythm.

Emphatic accent to emphasize a particular point in the sentence (Expressive function: emphasis)

Opposition: « on ne dit pas **la** garçon , on dit **le** garçon »

Emphasis : c'est **su**per beau

Differentiating : des échanges **hu**mains, **co**mmerciaux…
Positions itself on the first syllable of the linguistic unit,

> Example of tricky case : « mettez vos livres sous votre table» (by opp to "sur" or to "sa")

3/Identification of accented syllables within prosodic groups -› duration, energy, micro-melody, rhythm.

The emphatic accent is expressed either by:

an increased strength and duration of the consonant (“la garçon”).

a glottis stroke

A higher melodic rise

Prosodic module : models

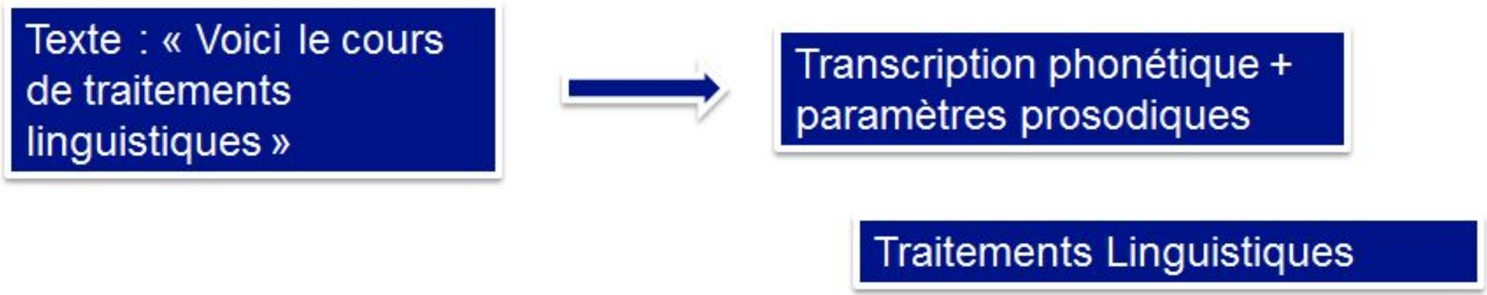
A model makes it possible to summarize all the intonation phenomena by a few parameters or intonation symbols.

E.g.: TOBI Tonal Model, Stylization (Dutch school), Fujisaki Model: reports the acoustic characteristics of intonation

Generation of speech signal/ Vocal synthesis

Architecture of a two-block TTS system

Text analysis / natural language processing

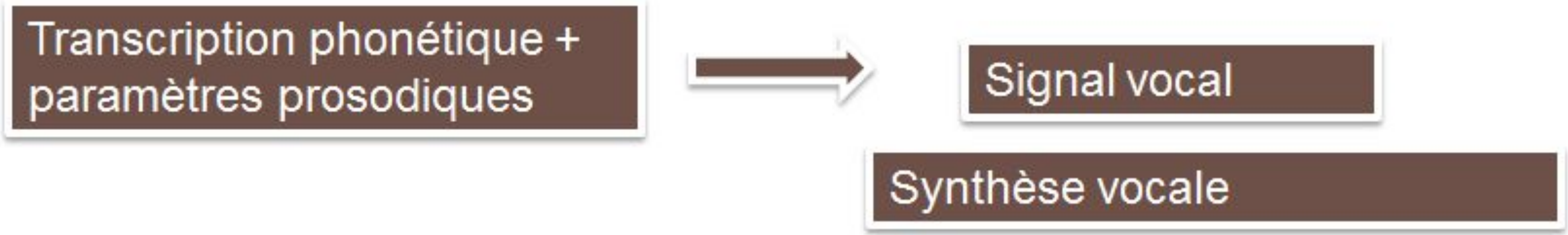


Texte : « Voici le cours de traitements linguistiques »

Transcription phonétique + paramètres prosodiques

Traitements Linguistiques

Synthesis : production of vocal signal



Transcription phonétique + paramètres prosodiques

Signal vocal

Synthèse vocale

Different approaches

1. Rule-Based Approaches [Klatt, 1980]
2. Synthesis by concatenation [Hunt & Black, 1996].
3. Parametric synthesis: generative synthesis from models (HMM) [Zen et al., 2009].
4. Deep learning approaches (WaveNet) [Van Den Oord et al., 2016].

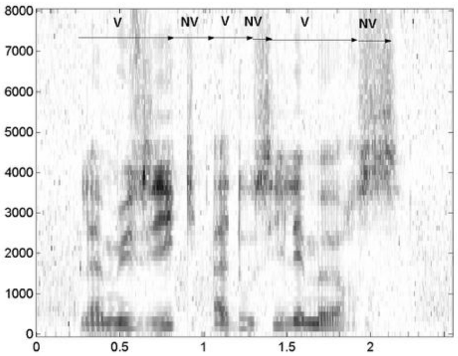
Examples :

https://deepmind.com/blog/article/wavenet-generative-model-raw-audio

1/ Rule-based approaches [Klatt, 1980]

Objective: to build the rules to create a sound signal from a sequence of given phonemes and the prosody (calculated from the linguistic analyses).

Principle: Reverse the process of reading the spectrogram, by doing formant synthesis



Spectrogram of the sentence : « la musique adoucit les mœurs »

1/ Rule-based approaches [Klatt, 1980]

Advantages:

- Little data to store

- Integrates knowledge about speech

Disadvantages:

- Long and tedious rule setting

- Rules depend largely on the language and to a lesser extent on the speaker

2/ Concatenative synthesis [Hunt & Black, 1996]

Principle :

Assembling speech segments (stored in a database) corresponding to the phoneme sequence

Purely acoustic smoothing of discontinuities that may appear at the points of concatenation

Note : Requires only limited knowledge of the speech signal.

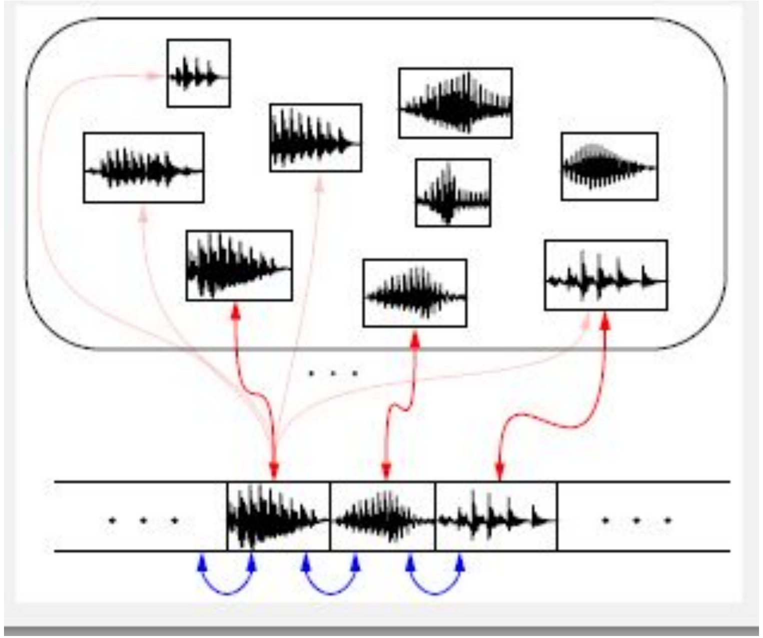


Schéma tiré de [1]

2/ Concatenative synthesis [Hunt & Black, 1996]

Database of speech segments :

From mono-speaker speech recordings with a wide linguistic diversity

Split the acoustic signal in a relevant acoustic unit (the diphone)

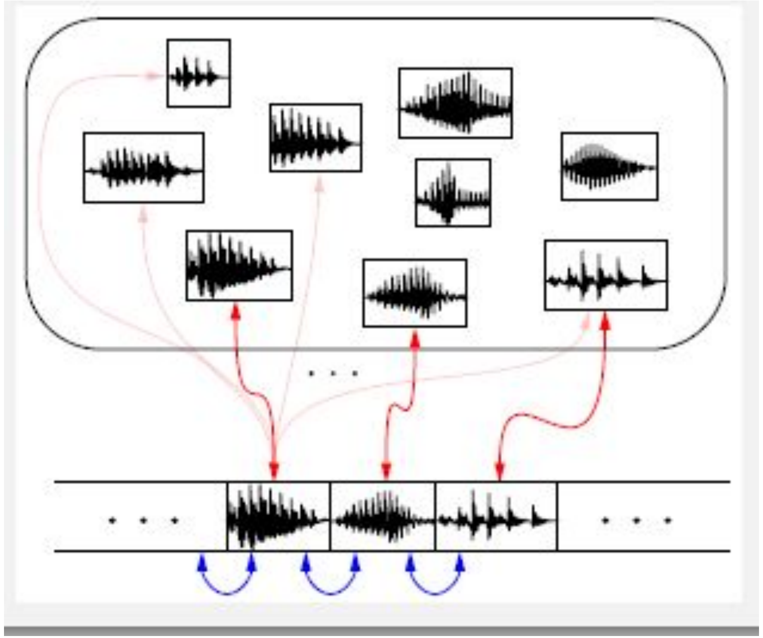


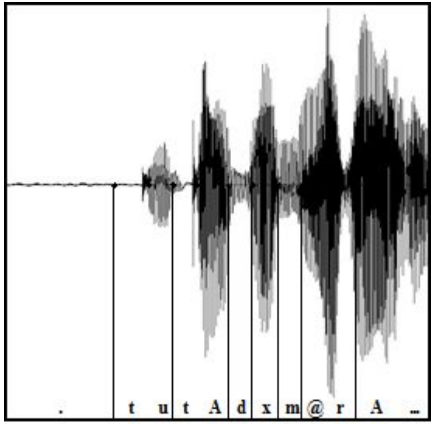
Schéma tiré de [1]

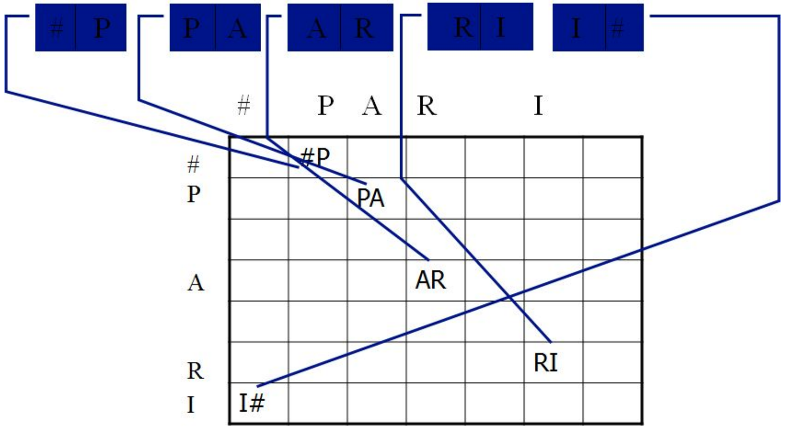
2/ Concatenative synthesis

Diphones: (blue) acoustic unit that begins in the middle of the stable range of one phoneme and ends in the middle of the stable range of the next phoneme.

Splitting into diphones requires to have an alignment between the signal and the phonemes

Allows for the inclusion of coarticulation phenomena.





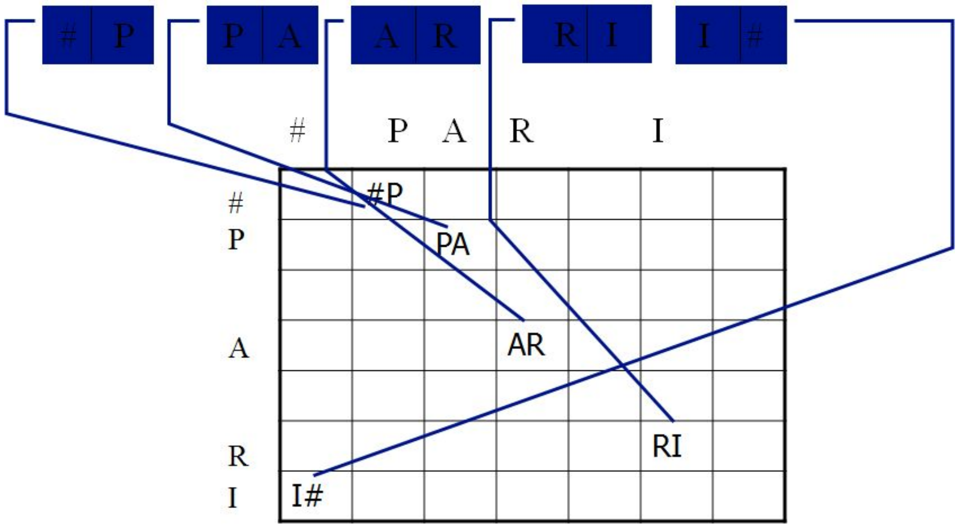
2/ Concatenative synthesis

Considering the input phoneme sequence

Select the synthesis units that will minimize future concatenation problems.

Two types of selection :

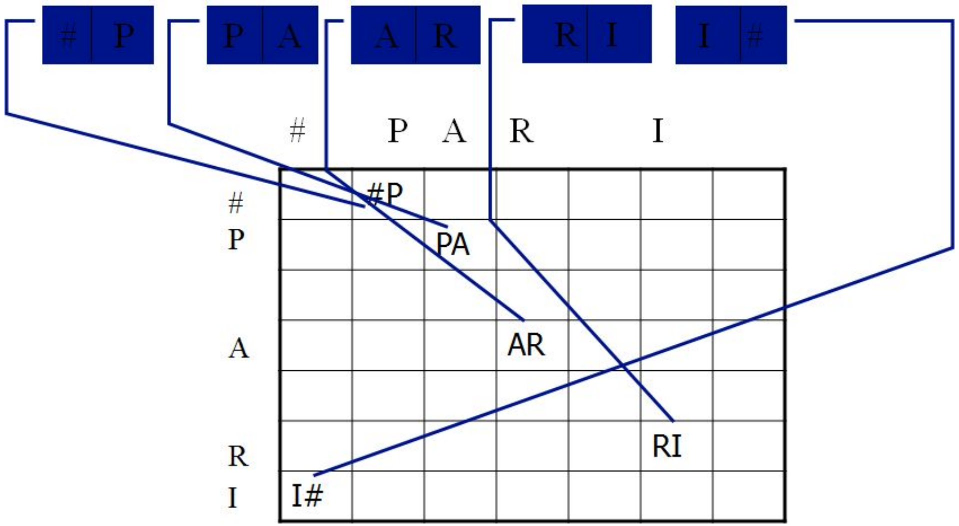
static or dynamic



2/ Concatenative synthesis

Static selection

- Only one choice per unit
- Pair of phonemes of the input phonetic chain <-> a diphone



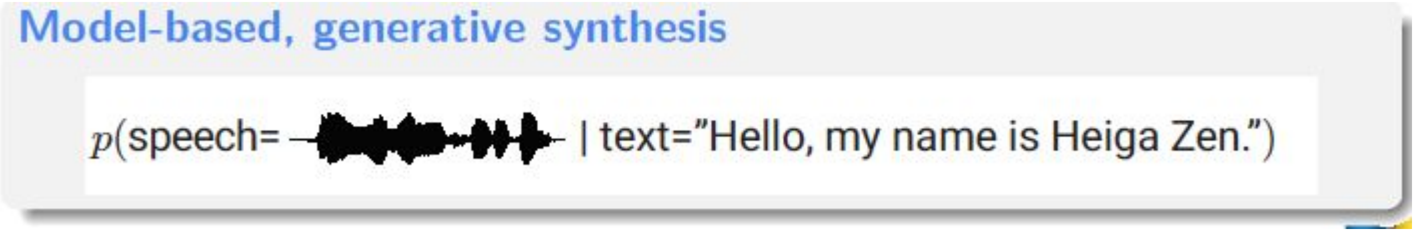
2/ Concatenative synthesis:

Dynamic selection

- Several instances of the same diphone are available with different prosodies and positioned in different phonetic contexts.
- Choice made at the time of the synthesis according to a global selection cost i.e. choice of the instance having:
    - the phonetic context as close as possible to the phonetic string to be synthesized (representation cost)
    - the prosody as close as possible to the prosody to be produced (representation cost)
    - Starts and ends with the fewest spectral discontinuities (cost of concatenation)
- Method:
    - Use of dynamic programming (Viterbi) in the lattice of usable segments.

3/Parametric synthesis: generative synthesis from models

Problem formulation of generative synthesis



Model-based, generative synthesis

p(speech= | text="Hello, my name is Heiga Zen.")

Instead of text : use phonetic and prosodic transcription

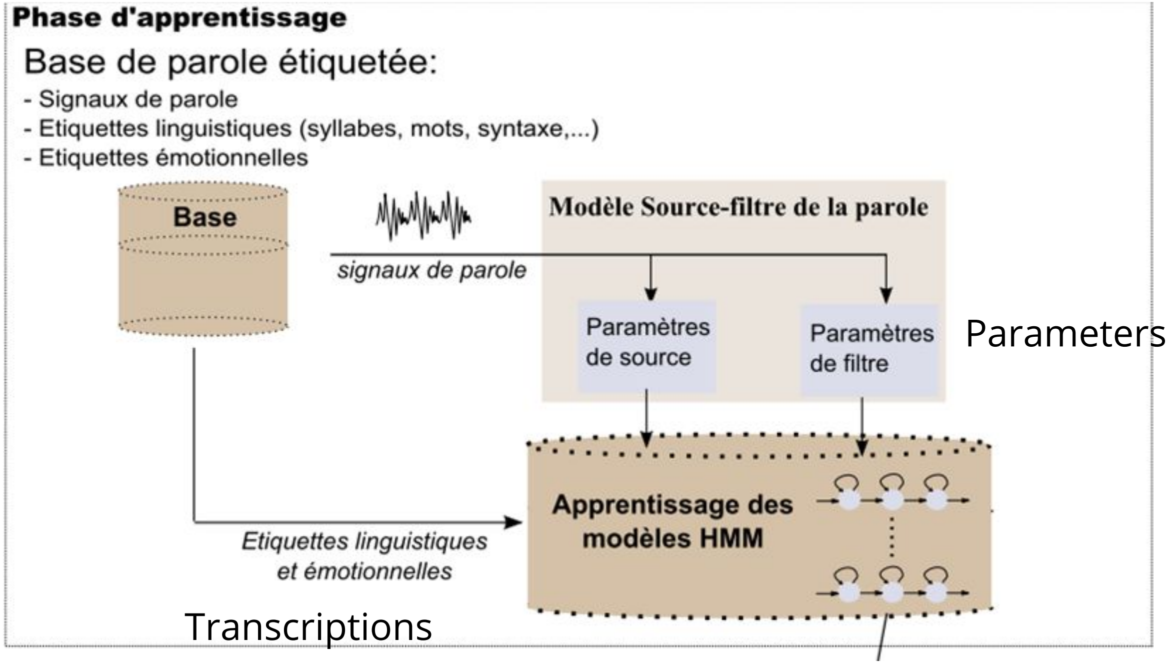
Instead of speech waveform : use parameters of speech waveform (e.g. source and filter parameters)

Model built from a training corpus :

Learn mapping between transcription <-> parameters of speech waveform

3/Parametric synthesis: generative synthesis from models

- HMM (Hidden Markov Models) - training phase



Phase d'apprentissage

Base de parole étiquetée:

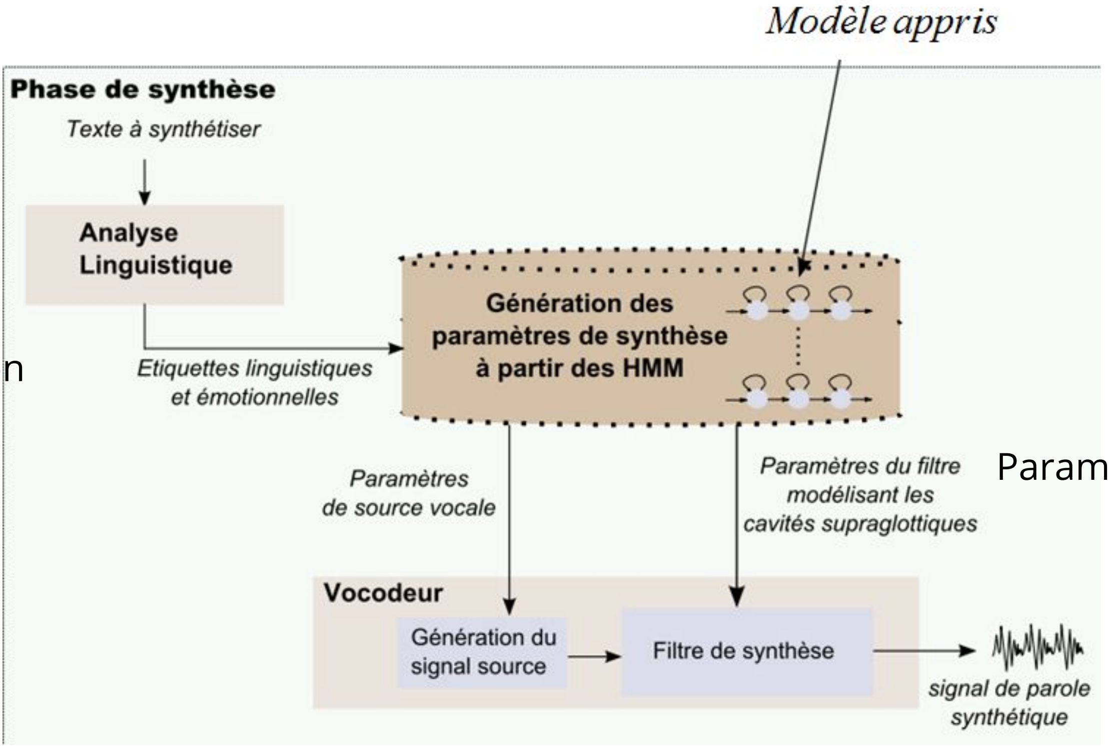
- Signaux de parole
- Etiquettes linguistiques (syllabes, mots, syntaxe,...)
- Etiquettes émotionnelles

Base
signaux de parole
Modèle Source-filtre de la parole
Paramètres de source
Paramètres de filtre
Apprentissage des modèles HMM
Etiquettes linguistiques et émotionnelles

Parameters of the speech form

Transcriptions

HMM synthesis


Modèle appris
Phase de synthèse
Texte à synthétiser
Analyse Linguistique
Génération des paramètres de synthèse à partir des HMM
Etiquettes linguistiques et émotionnelles
Paramètres de source vocale
Paramètres du filtre modélisant les cavités supraglottiques
Vocodeur
Génération du signal source
Filtre de synthèse
signal de parole synthétique

Transcription

Parameters of the speech form

Vocoder for speech signal synthesis

WaveNet speech synthesis

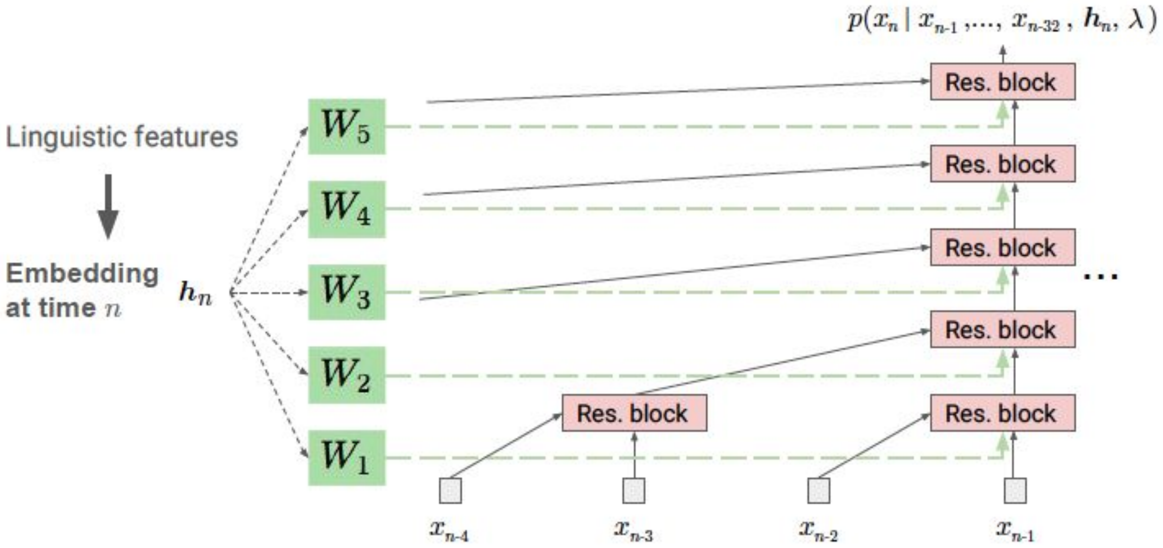
1/transforming the text into a sequence of linguistic and phonetic features (which contain information about the current phoneme, syllable, word, etc.)

Linguistic features : phone, syllable, word, phrase, and utterance-level features (e.g. phone identities, syllable stress, the number of syllables in a word, and position of the current syllable in a phrase, see http://www.cs.columbia.edu/~ecooper/tts/lab_format.pdf)

Phonetic features: frame position and phone duration and F0 features (obtained using LSTM-RNN-based phone duration and autoregressive CNN-based logF0 prediction models)

WaveNet speech synthesis

2/feeding this sequence of linguistic and phonetic features into WaveNet (a fully convolutional neural network)



WaveNet speech synthesis

3/ WaveNet generates raw audio waveforms using a model

- Fully probabilistic: compute the probability for each audio sample
- Autoregressive : the probability is conditioned on all previous audio samples and the sequence of linguistic and phonetic features of the text to synthetize

References

[1] Generative Model-Based Text-to-Speech Synthesis Andrew Senior (DeepMind London)

T. Dutoit [27]

cours de F. Beaugendre [10]

D. Klatt. Real-time speech synthesis by rule. Journal of ASA, 68(S1):S18{S18, 1980.

A. Hunt and A. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In Proc. ICASSP, pages 373{376, 1996.

References

J. Benesty, M. Sondhi, Y. Huang, « Handbook of Speech Processing », Springer, 2008 (1176 pages !!)

C. d’Alessandro et G. Richard, “Synthèse de la parole à partir du texte”, Collection Techniques de l’ingénieur, Paris, 2013 (à paraître)

O. Boeffard et C. d’Alessandro, « Synthèse de la parole » dans Analyse, Synthèse et Codage de la parole, Hermès, Lavoisier, 2002.

R. Boite, H. Bourlard, T. Dutoit, J. Hancq, and H. Leich. Traitement de la parole. Presses polytechniques et universitaires romandes,Lausanne, 2000.

H. Zen, K. Tokuda, A. Black « Statistical Parametric Speech Synthesis » , Speech Com. Volume 51, Issue 11, November 2009, Pages 1039–1064

Van Den Oord et al. "WaveNet : A generative model for raw audio.", 2016