Evaluation: Impact of Corpus Phonetic Alignment on the HMM-Based Speech Synthesis Quality

Marc EVRARD

Ph.D. in Computer Science

LIMSI, CNRS, Université Paris-Saclay

3rd International Conference on
Statistical Language and Speech Processing, **SLSP 2015**

Budapest, Hungary

Introduction

- HMM-based speech synthesis (HSS) models are trained on speech corpora
- Utterances read by a speaker, and annotated with phonetic labels
- Process of annotating a corpus starts with grapheme-phoneme (GP) conversion (complex probl.)
- State-of-the-art systems are still imperfect for most languages [Jouvet, D. *et al.* 2012]

Introduction
GP conversion

- GP conversion is a deterministic process, while the speaker phoneme realization is not
- Particularly true for the schwa, which is not realized systematically the same way by different speakers, in different situations
- In French, there are also liaisons between words, whose realizations particularly vary between speakers [Woehrling, C. & Boula de Mareuil, P. 2006]

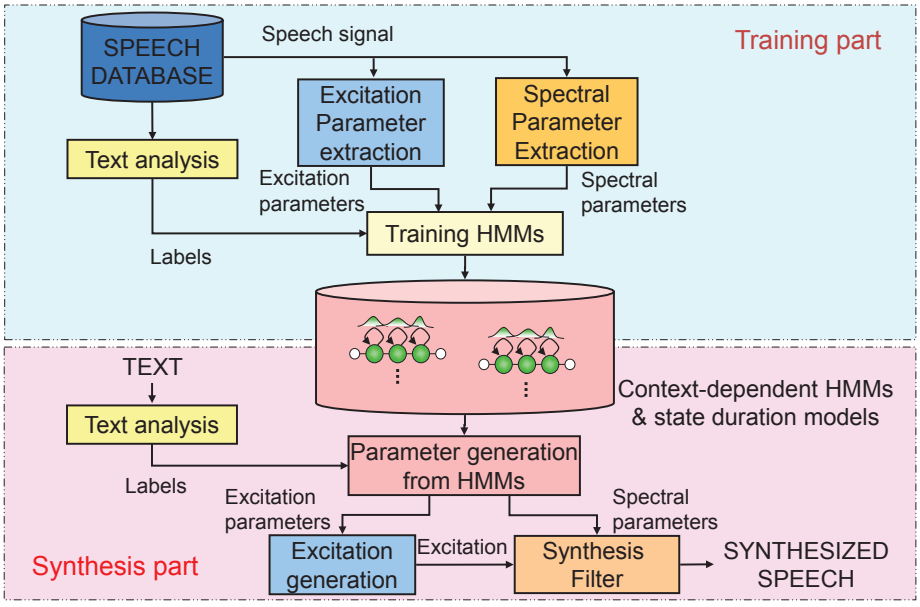
Introduction
Types of error tested

- After GP conversion, the alignment process segments the speech utterances
- 2 types of error can arise from corpus annotating processes:
  - Phonetic label errors:
    - pronounced by the speaker but not generated in the labels
    - not realized by the speaker but generated in the labels
  - Alignment errors
- In HSS, correspondence between label units and speech units is not direct (unlike in unit selection synthesis, USS)
- To what extent these systems are sensitive to the corpus annotation error?

Synthesis platform

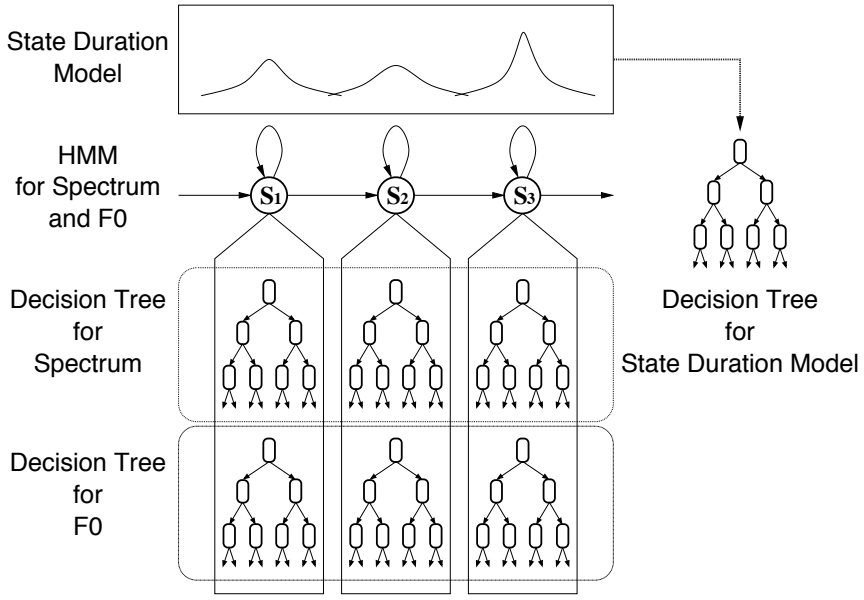
- LIMSI Parametric Speech Synthesis System (LIPS^3) is a TTS platform built around the HMM-based speech synthesis system (HTS) [Tokuda, K. 2013]
- Vocoder: sptk-3.7 (analysis) and hts-engine-1.08 (synth.) Basic excitation-filter model: impulse excitation and configured to use pure Mel-Frequency Cepstrum Coefficients (MFCC) for spectral envelop
- Language processing modules (for French), were developed in situ [Evrard, M. *et al.* 2015]

Synthesis platform
General process of HSS



[Tokuda, K. *et al.* 2013]

Synthesis platform
HTS



[Tokuda, K. *et al.* 2002]

Synthesis platform
GP

- GP conversion developed on a core set of rules previously created at LIMSI and evaluated in [Yvon, F. *et al.* 1998]
- Exclusively rule based, consisted of 7 stages:
  - sentence chunking
  - normalization
  - basic part of speech (POS) tagging
  - standard phonetization
  - peculiar rules application
  - liaisons management
  - syllabation

Synthesis platform
Corpus

- The text corpus was designed and recorded by an industrial partner (Vocally) in a collaborative project
- It consists of:

1 402	sentences
10 313	words
15 552	syllables
36 362	phonemes

- Speaker: professional actress, L1 of Parisian French
- The corpus was aligned using the Ergodic hidden Markov models (EHMM) tool [Prahallad, K. *et al.* 2006] from the Festvox tool suite

Limsi
Marc Evrard – Impact of Corpus Phonetic Alignment
SLSP 2015 – November 24, 2015
UNIVERSITÉ PARIS SUD
université PARIS-SACLAY

Synthesis platform
Linguistic contextual features

prev_prev_ph	Previous-previous phoneme
prev_ph	Previous phoneme
ph	Current phoneme
next_ph	Next phoneme
next_next_ph	Next-next phoneme
phone_from_syl_start	Position of the current phoneme in the syllable
phone_from_syl_end	(ditto counted from the syllable end)
syl_numphones	Number of phonemes in the syllable
syl_from_word_start	Position of the current syllable in the word
syl_from_word_end	(ditto counted from the word end)
syl_from_phrase_start	Position of the current syllable in the phrase
syl_from_phrase_end	(ditto counted from the phrase end)
syl_vowel	Vowel in the current syllable
word_numsyls	Number of syllable in the word
word_accent	Prominence of the current word
phrase_end	Final punctuation of the phrase
utt_numsyls	Number of syllables in the utterance
utt_numwords	Number of words in the utterance
utt_numphrases	Number of phrases in the utterance

Experiment

- Sensitivity of the annotation errors were tested:
  Different text-to-speech (TTS) systems were built,
  using the same speech corpus,
  with various altered annotations
- 2 types of variations in these systems:
  - Number of schwa and liaison realizations
  - Label alignment
- Set of sentences synthesized using the different systems
- Subjective evaluation to assess the quality differences

Experiment
Phonetic changes

- Phonetic realization changes:
  Schwa and Liaisons

Schwa	Suppr.	Add.
Content words	1917	227
Functional words	513	42
Total	2430	269
Ratio	6.68%	0.74%

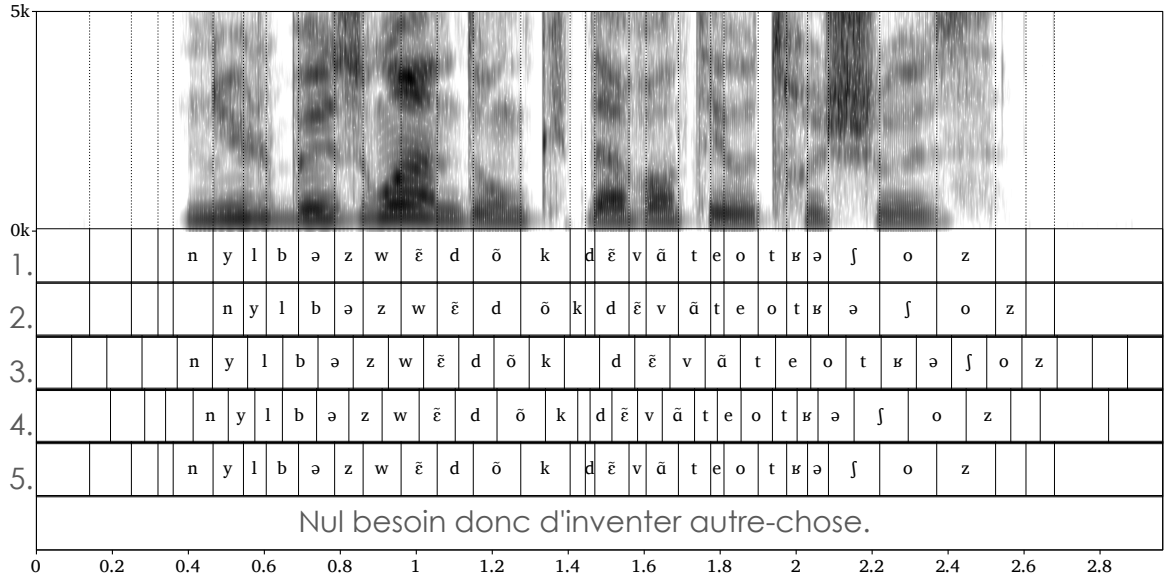
Liaisons	Suppr.	Add.
/z/	303	117
/t/	227	44
/n/	131	1
/p/	10	0
Total	671	162
Ratio	1.85%	0.45%

- GP rules modified to artificially increase/decrease occurrences

Experiment
Phonetic changes

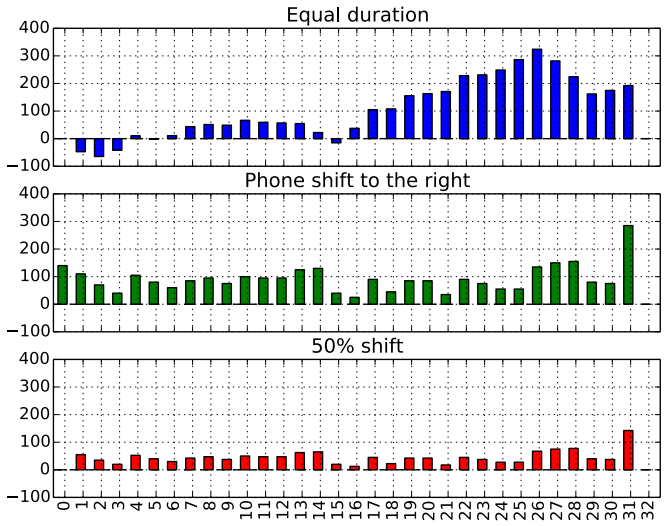
- Schwa change example ("Vous êtes le peuple souverain."):
  - Reference:
    /vu zɛt lə pœplə suvʁ~ɛ/
  - Schwas added:
    /vu zɛtə lə pœplə suvʁ~ɛ/
  - Schwas removed:
    /vu zɛt lə pœpl suvʁ~ɛ/
- Liaison change example
  ("Puis il remit avec orgueil son mouchoir dans sa poche."):
  - Reference:
    /pɥi zil ʁəmi avɛk ɔʁgœj s~ɔ muʃwaʁ d~ɑ sa pɔʃ/
  - Schwas added:
    /pɥi zil ʁəmit avɛk ɔʁgœj s~ɔ muʃwaʁ d~ɑ sa pɔʃ/
  - Schwas removed:
    /pɥi il ʁəmi avɛk ɔʁgœj s~ɔ muʃwaʁ d~ɑ sa pɔʃ/

Experiment
Boundary shifts



Experiment
Boundary shifts

- Time difference of the boundary position (ms) relative to the manually labeled corpus for each segment (here 32 segments)
- Highest value is reached near the end of the sentence for the isochronous segmentation (B1)



Results

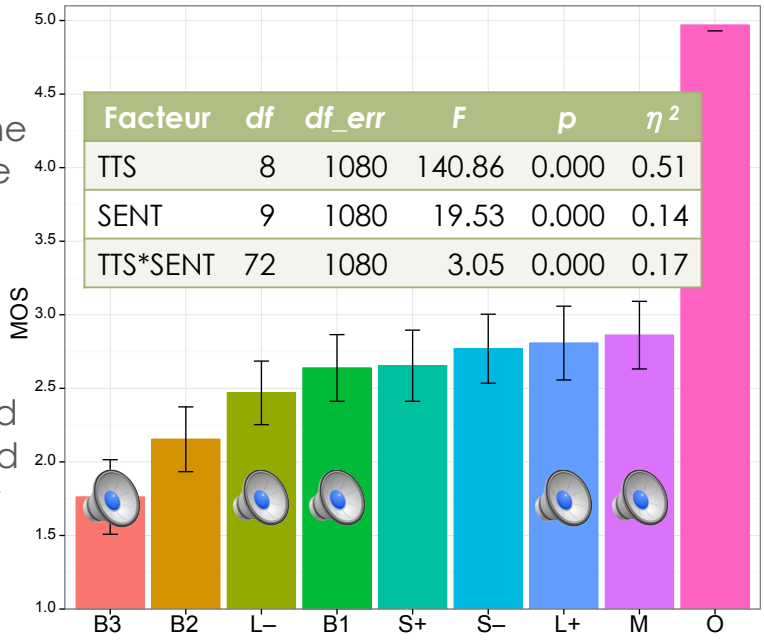
- 8 **systems** tested along with natural reference:
- 10 **sentences** synthesized with each system
- 13 **subjects** rated the overall quality of each sentence on a **MOS** (mean opinion score)

Code	TTS system
O	Natural
M	Manually corrected
L–	labels Less liaisons
L+	More liaisons
S–	Less schwas
S+	More schwas
B1	Isochronous segmentation
B2	Phone shifted right
B3	50% shift

Results
MOS

- "TTS" (systems) explains most of the observed variance
- 5 best systems: comparable quality
- 2 best (Natural and manual) perceived significantly better that the 3 worst

Facteur	df	df_err	F	p	η^2
TTS	8	1080	140.86	0.000	0.51
SENT	9	1080	19.53	0.000	0.14
TTS*SENT	72	1080	3.05	0.000	0.17



Limsi
Marc Evrard – Impact of Corpus Phonetic Alignment
SLSP 2015 – November 24, 2015
UNIVERSITÉ PARIS SUD
université PARIS-SACLAY

Results
MOS

	TTS system	Mean	Group
O	Natural	4.9692	A
M	Manually corrected	2.8615	B
L–	labels Less liaisons	2.8076	B
L+	More liaisons	2.7692	BC
S–	Less schwas	2.6538	BC
S+	More schwas	2.6385	BC
B1	Isochronous segmentation	2.4692	CD
B2	Phone shifted right	2.1538	D
B3	50% shift	1.7615	E

Results
Analysis: Phoneme occurrence

- Number of changes (phonemes) must be an important factor to explain the quality loss
- Not the only one: adding 6% of unperformed schwas in the labels does not lead to a significant quality loss
- But "forgetting" actually performed liaisons has a stronger effect on the quality output

Results
Analysis: Boundary shifts

- Resilience of the learning process to boundary shifts (to some degree)
- "50% shift" causes a stronger degradation, than the "phone shift to the right"
- "50% shift" leads to an alignment that maximizes mixes among phoneme labels (units located on phoneme transitions)
- "isochronous" (equal duration) segmentation results in a system whose quality is comparable with the reference

Conclusion

- HSS seems fairly robust to training corpus labeling errors
- According to these results, phonetic alignment precision should not be seen as a priority for HSS training corpora
- Observation of significant quality degradations linked to phoneme deletion supports the hypothesis of a greater sensitivity of the learning process to missing labeling
- Should push GP designers to favor realization of phonemes for ambiguous cases

Conclusion
Perspectives

- A next step for the analysis of phonetic variation sensitivity: use a fixed text corpus and phonetization, along with different phonetic realizations by the speakers
- Typical condition of expressive speech synthesis using common text for the different expressive corpora
- Provide a deeper analysis and an objective measurement of the resulting HMM model quality

Questions?

Thanks for your attention ☺

http://marcevrard.github.io/

marc.evrard@limsi.fr

References

[Jouvet, D. *et al*. 2012]
Evaluating grapheme-to-phoneme converters in automatic speech recognition context. In: ICASSP 2012.

[Woehrling, C. & Boula de Mareuil, P. 2006]
Identification d'accents regionaux en francais: perception et analyse. Revue Parole 37, 55.

[Tokuda, K. 2013]
Speech synthesis based on hidden Markov models. Proceedings of the IEEE 101(5).

[Evrard, M. *et al.* 2015]
Comparison of chironomic stylization versus statistical modeling of prosody for expressive speech synthesis. In : INTERSPEECH 2015.

References

[Yvon, F. *et al.* 1998]
Objective evaluation of grapheme to phoneme conversion for text-to-speech synthesis in French. Computer Speech & Language 12(4), 393–410 (1998).

[Prahallad, K. *et al.* 2006]
Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis. ICASSP 2006.

[Tokuda, K. et al. 2002]
An HMM-based speech synthesis system applied to English. Proceedings of 2002 IEEE Workshop on Speech Synthesis.