# Optimal feature set and minimal training size for pronunciation adaptation in TTS.

Marie Tahon, Raheel Qader, Gwénolé Lecorvé and Damien Lolive

IRISA,
EXPRESSION team, Lannion, France

October 12, 2016

# OUTLINE


1. Introduction
2. Material and methodology
3. Optimal feature set
4. Minimal training data size
5. Example
6. Conclusions and perspectives

# OUTLINE


1. Introduction
2. Material and methodology
3. Optimal feature set
4. Minimal training data size
5. Example
6. Conclusions and perspectives

# Context: the ANR project SynPaFlex

The project SynPaFlex aims at:

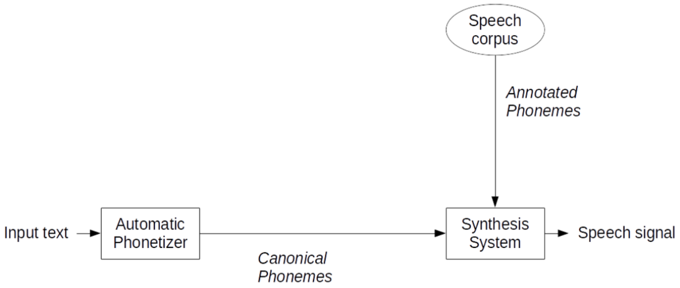
- improving flexibility of TTS systems (especially for audiobooks),
- generating high quality expressive speech.

We want to adapt pronunciation and prosody according to the semantic context.

$\rightarrow$ focus on pronunciation adaptation.

One of the main challenges when dealing with expressive speech is the lack of data (small-scaled corpora, no data at all, etc...)

# Introduction




⇒ How to reduce inconsistencies between phonemes as labeled in the speech corpus and phonemes generated by the phonetizer?

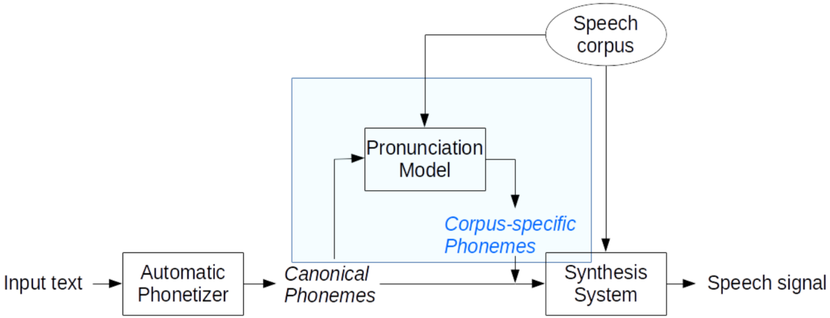
The speech corpus must be a big database carefully segmented and labeled. It is too expensive to consider pronunciation variants, or a new speech database.

⇒ Is it possible to use small expressive pronunciation databases?

# Introduction

Adaptation of the phonemes generated by a phonetizer to a specific pronunciation style ⇒ train a corpus-specific P2P model.

- As a case study, the considered pronunciation is the one uttered in the speech corpus itself
- To deploy this method to various cases, investigations are conducted on (i) the choice of optimal features, (ii) the minimal size of the pronunciation corpus to train reasonable adaptation models



# OUTLINE


1. Introduction
2. Material and methodology
3. Optimal feature set
4. Minimal training data size
5. Example
6. Conclusions and perspectives

# Speech Corpus

Overall description of the corpus:

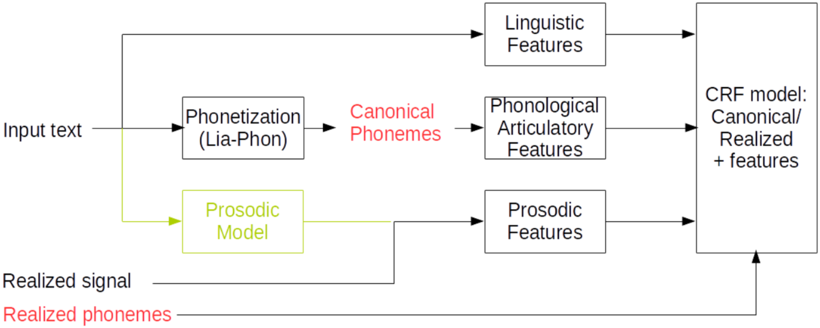
- Neutral female voice (16 kHz)
- 7208 utterances, 196,190 phonemes
- This corpus covers all French diphonemes and comprises most used words in the telecommunication field.
- Managed under the Roots toolkit [Chevelu,2014]
- Randomly split into a training set (70%) and a validation set (30%).
  - Training set: select and combine features in cross-validation conditions (7 folds)
  - Validation set: evaluate the resulting models in terms of PER and through perceptual tests.

Distribution of the corpus according to training and validation set, training set is divided in 7 folds in cross-validation conditions.

# Feature extraction

A 52 feature set [Qader,2015; Tahon,2016]:

- Canonical phonemes
  - Generated with Lia-Phon [Béchet, 2001]
- Linguistic
- Phonological
- Articulatory
- Prosodic (oracle)



# Feature extraction

A 52 feature set [Qader,2015; Tahon,2016]:

- Canonical phonemes
- Linguistic
  - Word frequencies are extracted with Google ngrams
  - Lemma and POS are extracted with Synapse
- Phonological
- Articulatory
- Prosodic (oracle)

| Linguistic features (18) |
|---|
| Word ♦ Stem ♦ Lemma ♦ POS ♦ Stop word ♦ Word, stem, lemma freq. in French (common, normal, rare) ♦ Word, stem, lemma freq. in corpus ♦ Word freq. knowing previous word in French, in corpus ♦ Word freq. knowing next word in French in corpus ♦ Number of word occurence in corpus (numerical) ♦ Word position, reverse position in utterance (numerical) |

# Feature extraction



A 52 feature set [Qader,2015; Tahon,2016]:

- Canonical phonemes
- Linguistic
- Phonological
  - Extracted using phonemes, syllable, pauses and word positions
  - Syllable structure using IPA information of its phonemes.
- Articulatory
- Prosodic (oracle)

| Phonological features (17) |
|---|
| Canonical syllables ♦ Phoneme in syllable position ♦ Phoneme in word position (begin, middle, end) ♦ Syllable in word position ♦ Phoneme position and reverse position in syllable (numerical) ♦ Phoneme position and reverse position in word (numercial) ♦ Syllable position and reverse position in word (numercial) ♦ Word length in phoneme (numerical) ♦ Word length in syllable (numerical) ♦ Syllable short and long structure (CVC, CCVCC) ♦ Syllable type (open, closed) ♦ Phoneme in syllable part (onset, nucleus, coda) ♦ Pause per Syllable (low, normal, high) |

# Feature extraction

A 52 feature set [Qader,2015; Tahon,2016]:

- Canonical phonemes
- Linguistic
- Phonological
- Articulatory
    - IPA phoneme information
- Prosodic (oracle)

| Articulatory features (9) |
|---|
| Phoneme type (vowel, consonant) ⧫ Phoneme aperture, shape, place and manner (open, close, front, central, undef, etc.) ⧫ Phoneme is affricate, rounded, doubled or voiced ? (boolean) |

# Feature extraction

A 52 feature set [Qader,2015; Tahon,2016]:

- Canonical phonemes
- Linguistic
- Phonological
- Articulatory
- Prosodic (oracle)
  - Extraction of energy (MFCC0), F0 and duration
  - F0 shape is based on a glissando value perceptually defined [d'Alessandro,1998]

| Prosodic features (7) |
|---|
| Syllable Energy (low, normal, high) ♦ Syllable and phoneme tone (from 1 to 5) ♦ $F_0$ phoneme contour (decreasing, flat, increasing) ♦ Speech rate (low, normal, high) ♦ Distance to next and previous pause (from 1 to 3) |

# OUTLINE

1. Introduction
2. Material and methodology
3. Optimal feature set
4. Minimal training data size
5. Example
6. Conclusions and perspectives

# Feature selection protocol

Three steps:

1. For each of the four feature groups: cross-validation (7 folds) forward feature selection without phoneme window
2. Feature combination of groups of selected features
3. Effect of phoneme window.

# Selected features

Protocol: Models are trained and evaluated in cross-validation on the training set (7 folds), without any phoneme window.
Forward selection process based on PER criteria + voting process.
Results:

- No articulatory features selected
- Most of prosodic features were selected
- Word frequencies were not selected: only word and stem remain
- Phoneme position in the utterance features were selected, but characteristics of syllables were not (nucleus, onset, VCV, CV, etc.)

| Group of feature | # selec/all | Selected features |
|---|---|---|
| Linguistic (L) | 2 / 18 | Word ♦ Stem |
| Phonological (Ph) | 7 / 17 | Canonical syllables ♦ Syllable in word position ♦ Phoneme reverse position in syllable (numerical) ♦ Phoneme position and reverse position (numerical) ♦ Word length in phoneme (numerical) ♦ Pause per Syllable (low, normal, high) |
| Articulatory (A) | 0 / 9 | - |
| Prosodic (Pr) | 6 / 7 | Syllable Energy (low, normal, high) ♦ Syllable and phoneme tone (from 1 to 5) ♦ $F_0$ phoneme contour (decreasing, flat, increasing) ♦ Speech rate (low, normal, high) ♦ Distance to previous pause (from 1 to 3) |

# Feature groups combination

Protocol: Models are trained on the training set (7 folds) and evaluated on the validation set.

Results:

- With only two apparently redundant features (word and its stem) a drop of 6.8 pp is obtained from the baseline.
- With very few features (8/52), the combination of linguistic and prosodic groups leads to a significant drop of 7.7 pp. from baseline
- The combination of the three groups (with a third of the initial set of feature) leads to the best PER with an improvement of 7.9 pp. from baseline
- We found a small subset of 15 features which leads to a significant improvement in terms of PER

| | | |
|---|---|---|
| Baseline (no adaptation) | | 11.2 [0.0] |
| Canonical phoneme only (C) | | 6.6 [-4.6] |
| C + L | 2 | 4.4 [-6.8] |
| C + Ph | 7 | 4.5 [-6.7] |
| C + Pr | 6 | 4.8 [-6.4] |
| C + L + Ph | 9 | 4.0 [-7.2] |
| C + L + Pr | 8 | 3.5 [-7.7] |
| C + Ph + Pr | 13 | 3.7 [-7.5] |
| C + L + Ph + Pr | 15 | 3.3 [-7.9] |

# Perceptive tests: example

Nous sommes responsables de tout le monde [*We are responsible for everyone*]

| Model | Phoneme sequence | HTS | Unit Selec. |
|---|---|---|---|
| Baseline | n u s ɔ m ə ʁ ɛ s p ɔ̃ s a b l ə d ø t u l ø m ɔ̃ d ə | | |
| Adapted C | n u s ɔ m - ʁ ɛ s p ɔ̃ s a b l - d ø t u l ø m ɔ̃ d - | | |
| Adapted CLPh | n u s ɔ m - ʁ ɛ s p ɔ̃ s a b l - d ø t u l ø m ɔ̃ d - | | |
| Adapted CLPhPr | n u s ɔ m - ʁ ɛ s p ɔ̃ s a b l ə d ø t u l - m ɔ̃ d - | | |
| Realized | n u s ɔ m - ʁ ɛ s p ɔ̃ s a b l ə d ø t u l - m ɔ̃ d - | | |

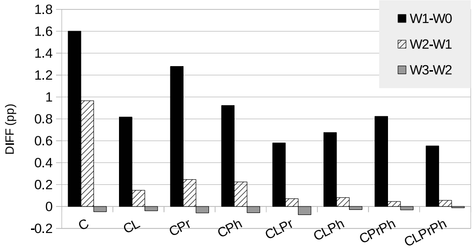
La guerre devient un peu moins improbable [*War becomes a bit less improbable*]

| Model | Phoneme sequence | HTS | Unit Selec. |
|---|---|---|---|
| Baseline | l a g ɛ ʁ ə d ø v j ɛ̃ - ɛ̃ p ø m w ɛ̃ - ɛ̃ p ʁ ɔ b a b l ə | | |
| Adapted C | l a g ɛ ʁ - d ø v j ɛ̃ - ɛ̃ p ø m w ɛ̃ - ɛ̃ p ʁ ɔ b a b l - | | |
| Adapted CLPh | l a g ɛ ʁ - d ø v j ɛ̃ - œ̃ p ø m w ɛ̃ - ɛ̃ p ʁ ɔ b a b l - | | |
| Adapted CLPhPr | l a g ɛ ʁ - d ø v j ɛ̃ - œ̃ p ø m w ɛ̃ - ɛ̃ p ʁ ɔ b a b l - | | |
| Realized | l a g ɛ ʁ - d ø v j ɛ̃ t œ̃ p ø m w ɛ̃ z ɛ̃ p ʁ ɔ b a b l - | | |

# Effect of phoneme window

Protocol: Models are trained on the training set (7 folds) and evaluated on the validation set.
Four symmetrical phoneme windows are tested; window are applied to current phoneme but also is associated features.
Results:

- The addition of one or two surrounding phonemes improves the PER (all the more so as feature set is small)
- A seven phoneme window, W3, degrades the results (overfitting)
- Windowing has a higher effect with prosodic features than linguistic or phonological features.
- W2 + 15 features brings the best improvement from baseline (-8.5 pp)



# OUTLINE


1. Introduction
2. Material and methodology
3. Optimal feature set
4. Minimal training data size
5. Example
6. Conclusions and perspectives

# Effect of the quantity of training material

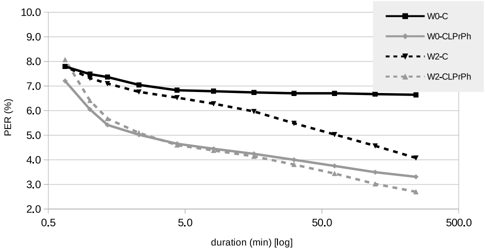
Protocol: Reduction of the training data by splitting the training set.
Cross-validation with 7 folds to 100 folds.

- Max size: 243.3 min of training data, 7 folds, 4321 utterances each
- Min size: 40 s of training size, 100 folds, 12 utterances each
- Validation: 120.2 min, 2161 utterances.

# Effect of the quantity of training material

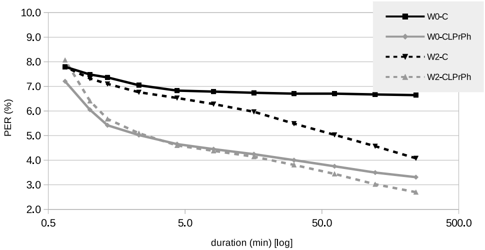
Results:

- Small durations reach a PER improvement of 4.0 pp (W0-CLPrPh) $\rightarrow$ small training sets allows fixing many errors. But STD is high, the choice of the training set is crucial.
- If duration $>$ 4.4 min, PER is almost linear with duration (in agreement with ASR result [Moore,2003])




| Training duration | Lin. Reg. | W0-C | W0-CLPrPh | W2-C | W2-CLPrPh |
|---|---|---|---|---|---|
| $>$ 0.7 min | Slope | -0.17 | -0.54 | -0.58 | -0.73 |
| | Corr. coef. | 0.74 | 0.85 | 0.99 | 0.86 |
| $>$ 4.4 min | Slope | -0.04 | -0.34 | -0.62 | -0.48 |
| | Corr. coef. | 0.96 | 1.00 | 0.99 | 0.99 |

# Effect of the quantity of training material




Conclusion:

| Durations | $<1$ min | 1-4 min | $>5$ min |
|---|---|---|---|
| Window effect | no | no | strong |
| Feature effect | no | strong | small |
| Linearity | no | no | yes |
| Improvement from baseline (in PER)$^{\star}$ | 4.0 pp | 6.6 pp | 8.5 pp |
| Improvement and duration | - | $\times 6.6 \rightarrow -2.6$ pp | $\times 10 \rightarrow -0.5$ pp |
| Best configuration | W0-CLPrPh | CLPrPh | W2 |

An ideal PER $= 0$, would be reached for $3 \cdot 10^8$ hours of speech !!!!

$^{\star}$: for best configuration

# OUTLINE


1. Introduction
2. Material and methodology
3. Optimal feature set
4. Minimal training data size
5. Example
6. Conclusions and perspectives

# Pronunciation adaptation: example

Example of pronunciation adaptations with different windows, features and training size. The input text is *Dans la montagne, les couleurs sont exceptionnelles.* “In the mountains, colors are remarkable”

| Win. | Features | dur(min) | Phoneme sequence |
|---|---|---|---|
| Realized | | | d ã l a m ɔ̃ t a n j - l e k u l œ ʁ s ɔ̃ t ɛ k s ɛ p s j o n ɛ l - |
| Canonical | | | d ã l a m ɔ̃ t a ɲ - ə l e k u l œ ʁ s ɔ̃ - ɛ k s ɛ p s j ɔ n ɛ l ə |
| W2 | CLPrPh | 243.3 | d ã l a m ɔ̃ t a n j - l e k u l œ ʁ s ɔ̃ z ɛ k s e p s j ɔ n ɛ l - |
| W2 | C | 243.3 | d ã l a m ɔ̃ t a n j - l e k u l œ ʁ s ɔ̃ t e k s ɛ p s j o n ɛ l - |
| W0 | C | 243.3 | d ã l a m ɔ̃ t a n j - l e k u l œ ʁ s ɔ̃ - ɛ k s ɛ p s j ɔ n ɛ l - |
| W2 | CLPrPh | 4.4 | d ã l a m ɔ̃ t a n j ə l e k u l œ ʁ s ɔ̃ t ɛ k s ɛ p s j o n ɛ l - |
| W2 | C | 4.4 | d ã l a m ɔ̃ t a n j - l e k u l œ ʁ s ɔ̃ t ɛ k s ɛ p s j o n ɛ l - |
| W0 | C | 4.4 | d ã l a m ɔ̃ t a n j - l e k u l œ ʁ s ɔ̃ - ɛ k s ɛ p s j o n ɛ l - |
| W2 | CLPrPh | 0.7 | d ã l a m ɔ̃ t a g - e l e k u l œ ʁ s ɔ̃ - ɛ k s ɛ p s j o n ɛ l - |
| W2 | C | 0.7 | d ã l a m ɔ̃ t a ʁ - - l e k u l œ ʁ s ɔ̃ t ɛ k s ɛ p s j o n ɛ l - |
| W0 | C | 0.7 | d ã l a m ɔ̃ t a ʁ - - l e k u l œ ʁ s ɔ̃ - ɛ k s ɛ p s j ɔ n ɛ l - |

- LIAISONS: W0 is not able to model French liaison: /s ɔ̃ t ɛ/, but W2 do; not always the correct one: /z/ instead of /t/
- ALPHABET: with 40 s of training data, models are not able to lable correctly the symbol /ɲ/: labels /n j/ are not found but /ʁ/, or /g/
- SCHWA: in the realized sequence, schwa is not pronunced, all models but WO-CLPhPr delete the canonical symbol /ə/.
- PRONUNCIATION: the substitution /ɔ/ → /o/ is better modeled with W2

# OUTLINE

1 Introduction

2 Material and methodology

3 Optimal feature set

4 Minimal training data size

5 Example

6 Conclusions and perspectives

# Conclusion

Objective of the presented work:

- Adaptation of phonemes generated by a phonetizer to the phonemes as labeled in the speech corpus in order to reduce inconsistencies.
- Investigation of an optimal feature set and a minimal training size.

Proposed solution:

- Train a CRF pronunciation model with linguistic, articulatory, phonological and prosodic features
- Reduce feature set dimension in cross-validation conditions.
- Reduce the quantity of training data for modeling pronunciations.

Main results:

- Reduction of the initial feature set from 52 to 15 features
- Corpus-specific adaptation method brings an improvement of 8.5 pp. in terms of PER (with W2-CLPrPh configuration)
- Over 5 min of training material, the addition of new data has a high cost but a weak improvement in accuracy ($\times 10 \rightarrow -0.5$ pp only). An ideal PER=0, would be reached for $3 \cdot 10^8$ hours of training data.
  $\Rightarrow$ For exploratory researches on pronunciation, 5 minutes seem enough, for end-users applications: the more data, the better.

# Perspectives

The presented pronunciation adaptation method (i) improves TTS quality, (ii) brings interesting perspectives in the use of small-scaled corpora for expressive TTS.

Further works:

- Phoneme adaptation to expressive speech (speaking style, emotions, direct/narrative style, regional accent, etc.)
- Introduction of n-best predicted phonemes into lattices for synthesis applications.

Thank you for your attention.
Any questions ?