Can we Generate Emotional Pronunciations for Expressive Speech Synthesis?

Marie Tahon, Gwénolé Lecorvé, Damien Lolive

Can we Generate Emotional Pronunciations for Expressive Speech Synthesis?

Marie Tahon, Gwénolé Lecorvé and Damien Lolive

Abstract—In the field of expressive speech synthesis, a lot of work has been conducted on suprasegmental prosodic features while few has been done on pronunciation variants. However, prosody is highly related to the sequence of phonemes to be expressed. This article raises two issues in the generation of emotional pronunciations for TTS systems. The first issue consists in designing an automatic pronunciation generation method from text, while the second issue addresses the very existence of emotional pronunciations through experiments conducted on emotional speech. To do so, an innovative pronunciation adaptation method which automatically adapts canonical phonemes first to those labeled in the corpus used to create a synthetic voice, then to those labeled in an expressive corpus, is presented. This method consists in training conditional random fields pronunciation models with prosodic, linguistic, phonological and articulatory features. The analysis of emotional pronunciations reveals strong dependencies between prosody and phoneme assimilation or elisions. According to perceptual tests, the double adaptation allows to synthesize expressive speech samples of good quality, but emotion-specific pronunciations are too subtle to be perceived by testers.

Index Terms—Expressive speech synthesis, emotion, pronunciation adaptation, conditional random fields.

1 Introduction

Speech synthesis usually consists in the conversion of a written text to a speech sound, process named as Text-To-Speech (TTS). Nowadays TTS tries to move from a neutral and machine-like style to expressive speech with different speaking styles, under various emotional states. This shift would greatly contribute to the fields of human-machine interactions, education, entertainment, etc. [1]. As a result, there is a crucial need not only for just intelligible speech carrying linguistic information, but also for the expression of affect, i.e. expressivity.

This paper investigates the generation of expressive speech for TTS systems focusing on the prediction of emotional pronunciations. In fact, expressive speech synthesis (ESS) presents two requirements: one is the detection of affect in the text itself [2], the other is the generation of a speech signal consistent with the suited affective message. Only the second issue is developed in the followings.

One of the main challenges in expressive TTS is to find the harmony between affective states in the input and the realization of prosodic characteristics to express them in the output speech [3]. Undoubtedly, prosody is the most important cue in the perception of affect in speech as many works in emotion detection have shown. However, the syntactic structure of a sentence defines a sequence of phonemes which design suprasegmetal prosodic cues (silences positions and durations, phrase breaks, etc.). Therefore, the study of emotional pronunciation is complementary to the study of emotional prosody. Unfortunately, research in this field usually focuses on suprasegmental prosodic variations only. And, as far as the authors know, no investigations have been done on the shared influences of pronunciation and emotional speech, be that on synthetic or natural speech.

The present article raises two issues in the generation of emotional pronunciations for TTS systems. The first issue consists in the automatic generation of emotional pronunciations from text. The fact that phoneme sequence greatly influences the prosodic and spectral parameters of the expressive speech signal is a challenging issue that have not been tackled yet in ESS. In the present article, grapheme sequences are automatically converted into *canonical* phoneme sequences using a rule-based phonetizer. Starting from a pronunciation adaptation method originally developed to improve the perceived quality in TTS [4], expressive pronunciation models are trained to adapt *canonical* neutral pronunciations to *target* emotional pronunciations as transcribed in an emotional pronunciation corpus. Different aspects of the automatic adaptation framework are studied and evaluated at the phonetic level. Because our aim is to improve ESS, the impact of pronunciation adaptation on the quality and expressivity of synthetic speech is evaluated through perceptual tests. With this aim in mind, neutral and expressive sentences are phonetized using our pronunciation adaptation framework, synthesized by querying a voice-specific speech database with a unit selection TTS system, and finally evaluated perceptually.

The second issue this article addresses, is the characterization of an emotional pronunciation. Through analyses, objective and perceptual evaluations, we study pronunciation variations in the expression of different emotional states, first on natural speech, then on generated pronunciations.

Section 2 presents a short review of emotional databases, approaches for the generation of expressive speech and pro-

- *M. Tahon was with the Expression team at IRISA (Lannion, France) when she conducted the present work. She is now with the LST team at LIUM (Le Mans, France) and also with Le Mans University, Le Mans, France.*
  E-mail: marie.tahon@univ-lemans.fr
- *G. Lecorvé and D. Lolive are with the Expression team at IRISA (Lannion, France) and also with the University of Rennes 1, ENSSAT (Lannion, France).*
  E-mail: {gwenole.lecorve, damien.lolive}@irisa.fr

nunciation variants studies. Section 3 describes the emotional pronunciation adaptation framework and the databases used for pronunciation adaptation and synthesis purposes. Section 4 details the training protocol for the generation of emotional pronunciations and their evaluation. Section 5 presents an analysis of target and generated emotional pronunciations. Finally, the results of perceptual tests are presented in Section 6.

2 Related work

First, expressive content of existing emotional databases is described highlighting the possible variations for data-driven TTS systems. Then, current state of the art data-driven systems for ESS are presented. Assuming that pronunciation together with prosody has a significant impact on expressive speech, studies on pronunciation variants in emotional speech are detailed.

2.1 Emotional databases

Emotion is a complex phenomena and finding a consensual definition is a hard task. According to Scherer [5], the number of human emotions occuring in the context of social interactions is infinite, subtle and often mixed. Most expressive databases are limited to the expression of a few acted emotional states, usually Ekman's *big six* [6]. According to Campbell [3], "part of the reason for the dominance of discrete emotions is the ease of collecting training data". In order to take into consideration other aspects of expressive voice such as social cues, intention or interactive cues, the complex nature of affect in speech can be described with continuous dimensions, notably activation and valence [7], but also control, dominance or intention [8].

There are mainly three types of emotional databases in the literature [9]. One type is emotional data simulated by professional actors (for example the well-known EMO-DB emotional database [10]). The second type is spontaneous data collected with real-life scenarios (for example the conversational database CHATR [11]). The third type is induced data obtained with scripted scenarios (for example human-robot interaction databases [**?**]). The collection of induced or spontaneous data is more challenging than acted data, thus such corpora turn to be not freely available. In the context of speech synthesis, acted data is recommended for cartoon animations or commercial applications while spontaneous data is better to deploy dialogue systems where the machine has to interact naturally with the user [12].

In the field of ESS, there is a need for carefully segmented and transcribed speech of good audio quality. The simplest solution is to use one of the few available databases which contain few acted emotional states or speaking styles [13], [14], thus leading to prototypical expressions of emotion. In order to model and synthesize more subtle and variable affective speech, data-driven ESS shows a growing interest for audio books as demonstrated by the recent Blizzard Challenges 2016 and 2017 [15], [16]. Audio books are very interesting for TTS as they contain both a text of interest, with different characters, speaking styles and emotions, and the corresponding audio signal [17]. In this paper, a large neutral speech database and a small emotional speech database (see section 3.3) are used to train, synthesize and evaluate models. However, to enlarge the variants in synthesized speech, the proposed protocol will be applied on an audio book in further work.

2.2 Generation of expressive speech

As aforementioned, two data-driven approaches coexist for TTS system: unit selection and statistical parametric systems (mainly HMM or DNN based), and both require emotional speech data of good audio quality. According to Schröder [18], parametric representation enables more flexibility than unit selection approach, because interpolation between styles contained in the database is possible. However, in both approaches, a solution to introduce flexibility in TTS consists in training acoustic models on speech produced with different speaking styles or in adapting models to specific voices or prosodic styles [19], [20]. This solution requires as many speech databases as required speaking styles and raises the issue of the consistency between semantics and expressivity. A speaker and expressivity factorization could help to solve this problem [21]. Otherwise, expressivity can also be controlled in symbolic terms (diphone identity, position, etc.) [22] and in prosodic terms (fundamental frequency, energy, duration) [23], [24]. Those elements are usually used in the speech synthesizer directly in the cost function of unit selection systems or in the construction of the acoustic model of parametric systems [25].

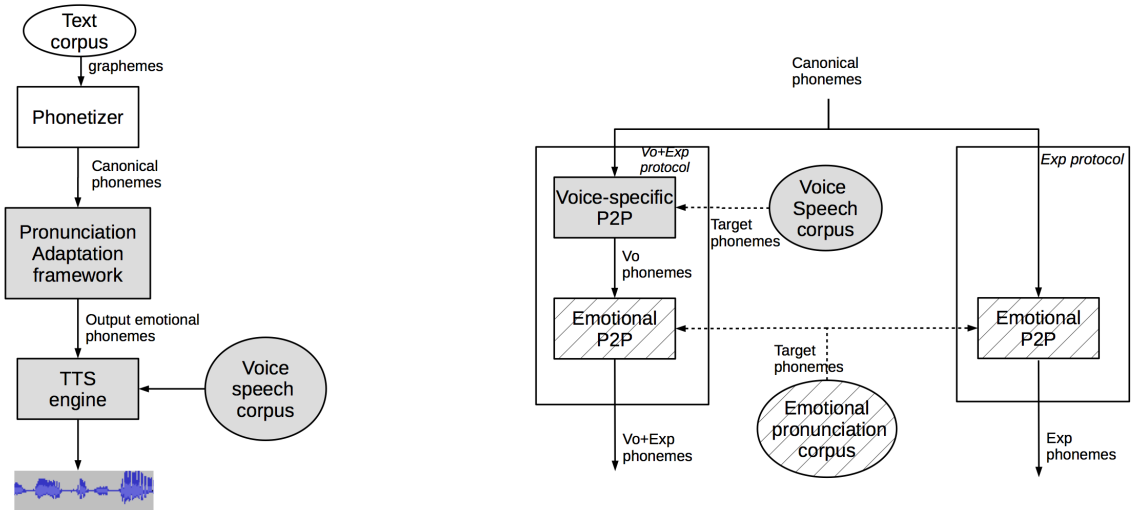
While most of the mentioned works on ESS focus on back-end acoustic models and prosody variations, these works generally use rule-based grapheme-to-phoneme tools without adapting pronunciation at the symbolic level to the target expressivity. The fact that phoneme sequence greatly influences the prosodic and spectral parameters of the expressive speech signal is challenging and has not been investigated so far. This paper explores front-end pronunciation variations in emotional speech for synthesis without controlling any prosodic or acoustic parameter. The next section gives an overview of studies related to pronunciation variants modelling.

2.3 Studies on pronunciation variants modelling

A possible way to introduce pronunciation variants into TTS is to manually add alternative pronunciations directly into the dictionary [26]. Rule-based pronunciation lattices have also been used to reduce inconsistencies between the phoneme sequence generated by the front-end text processing system (i.e. the phonetizer) and the phoneme sequence as transcribed in the training speech corpus [27]. However, building and maintaining such hand-crafted pronunciation lexicons designed by linguistic experts and transpose them to diverse speaking styles and expressivity is expensive and time consuming. Machine learning techniques have shown great advantages in this field, especially in Automatic Speech Recognition (ASR).

Some recent works in ASR have been done to characterize confusion matrices between phonemes with neural networks and conditional random fields (CRF) models [28]. Probabilistic acoustic models were implemented with expectation-maximization algorithm (EM) and weighted finite state transducers (WFST) and Viterbi approximation [29], thus improving the word recognition rate. Probabilistic pronunciation lattices were also predicted with CRF and WFST [30], [31].

Both hand-crafted and model-based pronunciation models require some features. Articulatory features describe physiological properties of the speech production process. They



(a) Integration of pronunciation adaptation framework in the TTS chain.

(b) Pronunciation adaptation protocols: Vo+Exp (left) and Exp (right)

Figure 1: General and detailed overviews.

were shown to be relevant for pronunciation modelling with decision trees [32] or Bayesian networks [33]. Linguistic, phonological and articulatory features can be directly derived from textual data, such as distinction between content and function words, word predictability or syllable locations [34], [35], [36]. Syllable-based features, among them schwas and liaisons, have also been investigated for pronunciation variants in French [37], [38]. A few years ago, acoustic features, mainly cepstral features, have been introduced to predict pronunciation variations [39].

Very few studies on the impact of pronunciation in the perception of expressivity were realized. A perceptual study [40] has shown that samples synthesized with the *target* pronunciation were preferred to those synthesized with the *canonical* pronunciation. Also, the adaptation of the *canonical* pronunciation to the voice corpus has shown a clear preference in terms of quality [4]. However, it seems that the generation of spontaneous speech requires some compromises between intelligibility and quality [41].

3 Method and material

This section presents the protocols and corpora used in the experiments. Different phoneme-to-phoneme (P2P) models are trained with a large set of features and two corpora designed for synthesis. The first corpus contains neutral speech while the second one contains different expressions of the *big six* emotions. Models are optimized with the phoneme error rate (PER) between predicted and target phonemes.

3.1 General overview

The presented adaptation framework (Fig. 1a) integrates in the TTS chain after the generation of *canonical* phonemes by a rule-based phonetizer and before synthesis. It adapts *canonical* phonemes to an emotion-specific pronunciation, i.e. predicts a new sequence of emotional phonemes. This new sequence is synthesized afterwards by querying a dedicated voice speech database with a unit-selection TTS system developed at IRISA [22]. Emotional phoneme sequences output by the framework under different configurations will be evaluated through phoneme error rate (see section 4.5) – and synthesized speech generated with the emotional phoneme sequences will be evaluated through perceptual tests as described in section 6.

3.2 Pronunciation adaptation framework

The goal of pronunciation adaptation is to reduce inconsistencies between a starting sequence of phonemes and a target sequence.

3.2.1 *Canonical and target phoneme sequences*

In the present work, the starting sequence is the one generated by the rule-based phonetizer from input graphemes (referred to as *canonical* sequence). The target sequence is constituted of the phonemes transcribed from an expressive or neutral speech corpus (referred to as *target* sequence). *Target* phonemes can be partially manually labeled using a given phonetic alphabet but a full and precise transcription by human annotators is not conceivable. In our case, transcriptions were extracted using text-to-speech alignement tools then manually checked. More precisely, the *canonical* phonemes derived from text, are aligned with the speech signal resulting in the *target* phonemes. Because in the present case speech is expressive and spontaneous, inconsistencies between *canonical* and *target* sequences (elisions, substitutions, insertions) are still numerous (see analysis at section 5).

3.2.2 *Exp single adaptation protocol*

As shown on Fig. 1b (right), a single P2P model represents the relation between *canonical* phonemes (observation) and *target* phonetic transcriptions of the emotional pronunciation corpus. This emotional P2P model predicts expressive adapted phonemes (referred to as *Exp* phonemes) starting from *canonical* ones. No adaptation to the voice speech corpus

is done. Exp protocol has the advantage of being fast because it uses less emotional data. The emotional P2P system should fit pretty well with emotional pronunciation, thus increasing the expressivity of output speech samples, but will probably overfit the data. Moreover, if this set-up is not adapted to the voice corpus, then inconsistencies between the corpus used for synthesis and the corpus used for pronunciation remain, lowering the TTS quality [4].

3.2.3 Vo+Exp double adaptation protocol

Fig. 1b (left) features two P2P models. The voice-specific P2P model represents the relation between *canonical* phonemes (observation) and *target* phonemes present in the voice speech corpus (which will be used to create the synthetic voice in section 6). This P2P model predicts neutral adapted phonemes (referred to as *Vo* phonemes) starting from *canonical* ones. A second emotional P2P models the relation between neutral adapted phonemes (Vo phonemes) and *target* phonemes present in the emotional pronunciation corpus. It predicts expressive adapted phonemes (referred to as *Vo+Exp* phonemes) from neutral adapted phonemes (Vo). Overcoming the disadvantages of the aforementioned protocol, the Vo+Exp protocol should reduce inconsistencies and consequently conduct to generate good quality expressive synthesized speech samples. To conclude, the two designed experiments aim at finding the best compromise between quality and expressivity for ESS.

3.3 Databases

Two speech databases designed for speech synthesis are used in the present article. *TelecomVo* corpus is a neutral corpus, while *EmotionPron* is an emotional corpus. They are split in order to keep data for evaluation. Their characteristics are given in Table 1. Corpora and their annotations are managed using the Roots toolkit [42].

3.3.1 TelecomVo corpus

TelecomVo is a French speech corpus originally dedicated to an interactive vocal server. As such, this corpus covers all diphonemes present in French and comprises most used words in the telecommunication field. It features a neutral female voice sampled at 16kHz. The corpus is composed of $7,208$ utterances, totaling 6h55' of speech. Pronunciations and non speech sounds have been strongly controlled during the recording process. For each sentence, the speech material has been automatically aligned[1] to the corresponding text. Obtained phonemes and alignement have been manually corrected. *TelecomVo* corpus has been randomly split in two subsets: 70% are left for development and training purposes (training the voice-specific P2P model and creating the synthetic voice) and the remaining 30% are kept for evaluations.

3.3.2 EmotionPron corpus

The *EmotionPron* corpus features a male speaker who recorded French sentences under the *big six* emotional states (anger, disgust, joy, fear, surprise and sadness). The sound signals were sampled at 16 kHz. About 50 sentences per emotion (360 distinct sentences in total) are recorded with a high activation degree. The same sentences are also recorded with a neutral reading style, hence enabling the comparison of prosodic features [14] between each emotion and the neutral state, on the same set of sentences. *EmotionPron* corpus gives the opportunity to analyze neutral and expressive pronunciations for different emotional states. Phonetic transcription (phonemes and segmentation) have been done automatically and manually checked [14]. The linguistic content of the sentences is informal and emotionally colored. For example: "*Mais t'es con ou quoi ?*" ("But you're a fool or what?") is informal language. The choice of such sentences greatly helps the speaker to simulate an emotion while acting. This pronunciation corpus contains little emotional data, almost 8 min for each emotional state. A previous experiment conducted on pronunciation adaptation has shown that a minimum of 5 min of training data is necessary to reach a satisfactory adaptation [43].

1. with Voxygen's alignement tool voxygen-group.com

Table 1: Characteristics of the corpora: presence of emotional speech, number of utterances, of phonemes and duration.

Corpus	Em.	# utt.	Dur.	# phon.
Pronunciation corpora (training)				
EmotionPron - 4 folds (X valid.)	yes	$\approx 6 \times 40$	0h33'	13,580
TelecomVo - 70%	no	5,046	4h51'	151,945
Voice speech corpus (unit-selection)				
TelecomVo - 70%	no	5,046	4h51'	151,945
Text corpora (evaluation)				
EmotionPron - 1 fold (X valid.)	yes	$\approx 6 \times 10$	0h08'	3,395
TelecomVo - 30%	no	2,162	2h04'	64,960

4 Pronunciation adaptation

This section details a general method which adapts a given pronunciation (here *canonical* pronunciation) to a target pronunciation (here *target* emotional pronunciation). Feature extraction and selection method as well as the training protocol of CRFs are presented. Models' evaluations in terms of PER between predicted and target phonemes are also described. PER is defined as the ratio between (a) the number of confusions between a reference sequence and a target sequence (substitutions (S), deletions (D) and insertions (I)) and (b) the total number of phonemes in the reference (N) as described in equation 1.

$$PER = \frac{S + D + I}{N} \tag{1}$$

At the end of this stage, emotional phoneme sequences are predicted for different emotional states. These sequences will be analyzed in section 5, then synthesized using a unit-selection engine and perceptually evaluated in section 6.

4.1 Features

Pronunciation models are trained with a large set of features. Precisely, four groups of features have been investigated: linguistic, phonological, articulatory and prosodic features. The corresponding set of 60 features presented in Table 2 is inspired from [41]. It has been enriched and adapted to French in [4]. Previous and next words are added in the feature set. Most features have been normalized to corpus or utterance, then discretized.

First, *canonical* phonemes are automatically determined from text with the phonetizer LiaPhon [44] – one of the most widely used utterance phonetization system for French. Word

Table 2: Groups of features used for pronunciation modelling experiments.

Linguistic features (26)
Word ♦ first and second words to current ♦ first and second words from current ♦ Stem ♦ Lemma ♦ POS ♦ first and second POS to current ♦ first and second POS from current ♦ 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 occurrence in corpus (numerical) ♦ Word position, reverse position in utterance (numerical)
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 (numerical) ♦ Syllable position and reverse position in word (numerical) ♦ 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)
Articulatory features (9)
Phoneme type (vowel, consonant) ♦ Phoneme aperture, shape (for vowels only) ♦ Phoneme place and manner (open, close, front, central, undef, etc.) ♦ Phoneme is affricate, rounded, doubled or voiced? (boolean)
Prosodic features (8)
Syllable Energy (low, normal, high) ♦ Duration ♦ 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)

frequencies in French are extracted from Google ngrams [45]. Articulatory features are derived from standard International Phonetic Alphabet (IPA) traits. In an ideal system, prosody should also be predicted from text. However, because this task is still a research issue, prosodic features have been extracted in an oracle way, i.e., directly from the recorded utterances of the speech corpus. Prosodic features are based on energy, fundamental frequency F_0 and duration. F_0 shape is based on a glissando value perceptually defined [46]. In the future, a text-to-prosody model could be included in the synthesizer, thus making prosodic features available. In any case, such a protocol allows us to know to what extent prosody affects pronunciation models.

4.2 Feature selection protocol

In the presented work, phonemic sequences are modelled with CRFs. CRFs are trained under cross-validation conditions (with 5 folds) with the Wapiti toolkit [47] and its BFGS algorithm[2]. Models are trained with *a minima* input phonemes – either *canonical* or predicted Vo phonemes – and *a maxima* all features. CRF models are evaluated with PER between predicted and *target* (either neutral or expressive) phoneme sequences.

An automatic feature selection is performed in the same way for each of the three P2P models presented in Fig. 1b. In the applied method already presented in [4], features are selected separately for each of the four groups of features (linguistic, phonological, articulatory and prosodic) reported in Table 2. For each group of features, three symmetric phoneme window sizes are tested (W_0: current phoneme only, W_1: current, previous and next phonemes, W_2: current, 2 previous and 2 next phonemes). The forward feature selection starts with input phonemes only and other features belonging to the same group are added one at a time until the best subset of feature S_x is reached for window W_x. This best subset is found when the addition of one more feature does not improve the PER.

In order to find the global subset over the 5 folds, a voting process has been set up. For each fold, a selected feature receives a vote $v = 1$, therefore the maximum of votes for the global selection process is the number of folds. Features which receive more than one vote $nv > 1$, are added in the global subset S^x for window W_x.

2. Broyden-Fletcher-Goldfarb-Shanno algorithm

4.3 Voice-specific P2P model

In their previous work [4], [43], the authors have presented the training process of a voice-specific P2P model with the corpus *TelecomVo* training subcorpus. A first set of 15 features including linguistic, phonological and prosodic features with a W_2 window, was automatically selected. With this 15-feature set, an optimal PER (between predicted and target neutral phonemes) of 2.7% was reached. In comparison, the baseline PER (between canonical and target neutral phonemes) was 11.2%.

However, a perceptual test has shown that the 6 selected prosodic features had no effect on the quality of synthesized speech samples. Furthermore, prosodic features are not generated from text yet but are estimated in an oracle way. Consequently, we decide to use only the selected linguistic and phonological 9-feature set and a 5-phoneme window (W_2) to train voice-specific P2P models on the voice speech corpus.

4.4 Emotional P2P model

Each emotional subcorpus of *EmotionPron* has been randomly split in five, each fold containing almost 10 utterances. Because very little expressive data is available, the two experiments (Exp and Vo+Exp) are conducted under a 5-fold cross-validation protocol. CRFs emotional P2P models are trained on 3 folds and optimized on 1 fold. They are evaluated on the remaining fold with the PER between predicted and target emotional phonemes. In such configuration, the optimization data (feature selection, combination and phoneme window described in this section) and the final evaluation data (described in next section 4.5) belong to the same corpus while being distinct. To strengthen our results, emotional P2P models will also be evaluated with neutral data from the *TelecomVo* evaluation subcorpus as shown in Table 1.

4.4.1 Feature selection results

In order to investigate how expressive and emotional pronunciations are, the number of selected features of each linguistic, phonological, articulatory and prosodic group is reported in Table 3 for each experiment. This table shows the relative

Table 3: Relative number of selected features, of a group for the three windows compared to the total number of selected features for the three windows, per emotion.

Feature groups	Ang	Dis	Joy	Fea	Sur	Sad	W_0	W_1	W_2	AVG
	Vo+Exp									
Linguistic	15.8	23.5	11.1	16.7	18.9	15.6	21.2	15.1	16.0	17.4
Phonological	24.6	27.9	33.3	31.3	26.4	17.8	28.8	26.4	25.5	26.9
Articulatory	24.6	17.6	11.1	14.6	17.0	20.0	12.5	19.8	20.8	17.7
Prosodic	35.1	30.9	44.4	37.5	37.7	46.7	37.5	38.7	37.7	38.0
	Exp									
Linguistic	25.8	16.7	14.8	13.6	15.7	18.5	22.5	17.4	13.9	17.9
Phonological	22.6	22.2	24.1	20.5	25.5	22.2	27.5	20.2	21.3	22.9
Articulatory	17.7	22.2	20.4	18.2	13.7	24.1	7.8	25.7	24.1	19.4
Prosodic	33.9	38.9	40.7	47.7	45.1	35.2	42.2	36.7	40.7	39.8

weight of each feature group averaged over window per emotion and averaged over emotions per window.

As aforementioned, no prosodic features have been used to train the P2P voice pronunciation model. Almost all prosodic features were selected in both experiments. This result confirms the great importance of prosody for emotional pronunciation modelling. It can be due to the fact that prosody features were extracted from target speech in an oracle way. For example when a voiced phoneme is elided in its realization, the F_0 is arbitrarily set to -1, which is a very discriminant cue.

Phonological features are also important for pronunciation, especially the canonical syllable to which the current phoneme belongs, but also the position of the phoneme in the syllable or in the word.

In both experiments, the number of selected articulatory features is quite small with W_0 window (7.8% in Exp and 12.5% in Vo+Exp) and increases with the window size. In the mean time, the number of selected linguistic features decreases inversely with the window size. This result was expected since a large phoneme window size corresponds more or less to the word level, while a small size corresponds to the phoneme level.

While the relation between selected features and window sizes is quite clear, the relation between selected features and emotions is more subtle. Further analysis on pronunciation variants and emotion are detailed in section 5.

4.4.2 Optimal feature combination and phoneme window

Once feature selection is performed for each group, the combination of these groups and phoneme windows are investigated to find the best configuration (selected features, phoneme window). All combinations are evaluated with expressive utterances under cross-validation conditions in terms of PER between predicted and *target* emotional phonemes for the two Vo+Exp and the Exp experiments. The complete results are not reported in this study.

The PER baseline between *canonical* and *target* phonemes is obtained without adaptation. PER between predicted and *target* phonemes is obtained with adaptation and compared to the baseline. Results are given for Vo+Exp experiment [resp. Exp experiment]. The adaptation of Vo [resp. *canonical*] phonemes with a minimal set of features constituted of the current phoneme only, brings a slight improvement from the baseline of 5 percentage point (pp.) [resp. 3.6 pp.] on average over emotions. The addition of an optimal feature set has a much greater effect. The best compromise between a good PER and a small number of features is reached with the configuration (W_0, S_E^0) for emotion E. In that case, the improvement is 12.1 pp. [resp. 12.7 pp.].

4.5 Evaluation with expressive and neutral utterances

Table 4: PER between reference and target pronunciations (W_0, S_E^0). PER between Vo and target is 2.7% with neutral utterances.

Reference/Target	Ang	Dis	Joy	Fea	Sur	Sad	AVG
Expressive utt. in input (target: expressive)							
Cano/target (baseline)	16.9	18.8	16.2	16.7	16.7	18.2	17.3
Vo+Exp/target	5.8	5.1	4.6	5.6	5.8	4.3	5.2
Exp/target	5.2	4.4	4.1	4.5	5.2	4.5	4.6
Neutral utt. in input (target: neutral)							
Cano/target (baseline)	11.1						11.1
Vo+Exp/target	9.1	9.3	8.8	8.8	8.7	10.0	9.1
Exp/target	10.5	10.8	10.5	10.9	10.5	11.3	10.7

In a second evaluation, predicted emotional pronunciations are evaluated with expressive and neutral utterances. Expressive text comes from *EmotionPron* and models are evaluated under cross-validation conditions. Neutral text comes from the *TelecomVo* evaluation subcorpus and models are tested with distinct training and evaluation subcorpora. The aim here is to estimate the generalization power of the emotional pronunciation adaptation models.

Emotional pronunciation models are trained with appropriate feature subset and window (S_E^0, W_0) for each emotion E. Table 4 reports the PER obtained between predicted Vo+Exp or Exp, and *target* pronunciations.

In the case of neutral utterances, the *target* phonemes are neutral, but the adapted Exp and Vo+Exp phonemes are expected to be expressive. That is the reason why the reported PER is much more important with neutral utterances than with expressive utterances. As expected under cross-validation conditions, Exp pronunciations have a better score than Vo+Exp pronunciations with expressive utterances. One can observe the contrary with neutral utterances, thus highlighting that Vo+Exp model better generalizes than Exp model. This result was expected since the voice training corpus contains almost 5,000 utterances and the expressive training corpus contains almost 40 utterances. Average PER are quite similar with experiments Vo+Exp and Exp: 5.2% vs. 4.6% with expressive text and 9.1% vs. 10.7% with neutral text. We also checked the PER between Vo+Exp and Exp predicted pronunciation. The obtained PER are 2.6% with expressive text and 4.7% with neutral text on average. It

shows that the two models indeed predict different phoneme sequences.

In conclusion, pronunciation adaptation reduces the PER between a target emotional pronunciation and a predicted emotional pronunciation. As expected, the two proposed experiments show some differences in the predicted expressive phoneme sequences. Finally, because of the size of the voice corpus, adapting pronunciation to it better generalizes to unseen data. Clear differences are noticeable between neutral and emotional pronunciations, but differences in PER between emotions are smoothed. Therefore a deeper investigation on both expressive and emotion-specific pronunciations is needed and conducted in the next section.

5 Analysis of emotional pronunciations

In this section, the main rules learnt by pronunciation models are exemplified. The influence of prosody on phoneme elisions is also shown with examples. A deeper investigation on automatic rules generated by CRFs shows the existence of emotional pronunciations which differ from one emotion to another.

Table 5: PER (%) between reference and target pronunciations for the six emotions expressed in the *EmotionPron* corpus. Average PER and relative standard deviation on emotions.

Reference/Target	Ang	Dis	Joy	Fea	Sur	Sad	AVG
Cano/neu	15.2	14.6	13.1	14.0	14.6	14.6	14.4 ± 5.0 %
Cano/emo	16.9	18.8	16.2	16.7	16.7	18.2	17.3 ± 5.8 %
Neu/emo	5.1	8.3	6.8	6.7	4.5	7.0	6.4 ±21.5 %

5.1 Changes between read and expressive target speech

5.1.1 Objective evaluation (PER)

Baseline PER reported in Table 5 are reasonably high in comparison to the baseline of 11.2% obtained on the *TelecomVo* evaluation subcorpus. Table 5 shows a clear difference in PER between neutral read (14.4% on average) and expressive (17.3% on average) speech and smoothed variation over emotions. This is probably due to the fact that the phonetizer was tuned with neutral read speech. Pronunciation variations across emotions may be due to the pronunciation and to the text itself because some word sequences are recurrent for a given emotion, but also to a specific emotional pronunciation.

Trying to go over text dependencies on emotion-specific phonemes, the PER between *target* neutral and emotional phoneme sequences is computed for each emotion (see line neu/emo in Table 5). Disgust is clearly the emotion which provides more changes in the pronunciation. Joy, fear and sadness have an intermediate number of phoneme changes and anger and surprise have the smallest number of variations with respect to the reading style.

5.1.2 Homogenisation between annotations

As mentioned previously, the phonetizer has been built with fixed rules, and some outputs may be mistaken. The pronunciation adaptation method helps at the homogenisation of alphabet's symbols between the phonetizer's outputs and the voice corpus annotations. The main difference lies in the annotation of the French schwa. For example, the phonetizer outputs the symbol /ø/ corresponding to the realization of a schwa which should be pronounced by the TTS engine (generally middle schwa) and the symbol /ə/ for each schwa which should be elided during the synthesis process (final schwa). The use of symbol /ø/ as a schwa is probably wrong as this phoneme can not be elided. In the *TelecomVo* corpus, this phoneme is annotated with either nothing when not pronounced, either with open /œ/ or closed /ø/ when pronounced. In *EmotionPron*, the symbol /ø/ is never used and schwa are represented with either /ə/ when pronounced, or nothing when elided. A last example: *canonical* and *EmotionPron* corpus annotations use the IPA symbol /ɲ/ while the *TelecomVo* corpus alphabet contains /n j/ only. Of course, some basic rules could fix most differences between corpora. However, such rules may not be relevant with different phonetizer or corpus or language, and would not be able to manage speaker's individual pronunciations.

5.1.3 Phoneme confusions between styles

In this experiment, three different pronunciations are used: *canonical*, *target* in *TelecomVo* and *target* in *EmotionPron*. Except alphabet mapping, four types of phoneme confusions have been reported. A lot of pronunciation variants, related to the pronunciation of the speaker itself, are observed for mid-vowels /ø/, /ə/, /e/, /ɛ/, /ɔ/, /o/ (for example, /e/ ↔ /ɛ/ and /o/ ↔ /ɔ/) [43], [38]. The elision of final liquids /ʁ/ and /l/ is also observed in the target pronunciation. For example the liquid /l/ is basically elided from *il y a*, thus giving the phonemes /i y a/. This phenomenon has already been reported by Brogneaux et al. [40]. Finally, voiced assimilation seems to be an important phenomenon in spontaneous speech. For example the French /ʒ ə s ɛ p a/ (*je sais pas*) is usually expressed as the unvoiced /ʃ s ɛ p a/, thus devoicing /ʒ/ to /ʃ/ and eliding /ə/. Because of vowel elisions, consonant clusters occur more often with a high speech rate. These clusters turn out to be either completely voiced or completely unvoiced [48], [49].

The main confusions between the neutral reading and emotional styles are:

- deletion of canonical /ə/,
- re-insertion of /ə/: more schwas are pronounced (mainly for anger),
- substitution of canonical /ʒ/ by /ʃ/,
- deletion of liquids: /l/, /ʁ/.

As shown in the preceding list, most confusions are deletions. On average over emotions, the speech rate is lower in the neutral speech (5.0 syllables per second) than in the expressive speech (5.7 syllables per seconds). Speech rate even reaches 6.1 for disgust and joy in the expressive speech. An increase of speech rate mechanically leads to phoneme deletions, usually schwa and liquids. As mentioned previously, the transformation of /ʒ/ into /ʃ/ is an assimilation which is also strongly linked to speech rate.

5.2 Changes across emotional generated pronunciations

5.2.1 Objective evaluation (cosine similarity)

The analysis of the pronunciation data contained in *EmotionPron*, as well as PER differences between neutral reading and emotional speech, have both established the

Table 6: Emotional pronunciation: *"J'aurais dû accepter qu'il me raccompagne."* ("I should have accepted he took me back.")

Expressive utterance (fear)																														
Canonical	ʒ	ɔ	ʁ	ɛ	d	y	a	k	s	ɛ	p	t	e	k	i	l	m	ø	ʁ	a	k	õ	p	a	ɲ	ə
Target	ʒ	o	ʁ	ɛ	d	y	a	k	s	ɛ	p	t	e	k	i	l	m	ə	ʁ	a	k	õ	p	a	ɲ	ə
Vo (Ling+Phon)	ʒ	o	ʁ	ɛ	d	y	a	k	s	ɛ	p	t	e	k	i	l	m	ø	ʁ	a	k	õ	p	a	n	j
Vo (Ling+Phon+Pros)	ʒ	o	ʁ	ɛ	d	y	a	k	s	ɛ	p	t	e	k	i	l	m	ø	ʁ	a	k	õ	p	a	n	ə
Vo+Exp (Ling+Phon)	ʒ	o	ʁ	e	d	y	a	k	s	ɛ	p	t	e	k	i	l	m	ə	ʁ	a	k	õ	p	a	n	j
Vo+Exp (Ling+Phon+Pros)	ʒ	o	ʁ	e	d	y	a	k	s	ɛ	p	t	e	k	i	l	m	ə	ʁ	a	k	õ	p	a	n	ə
Exp (Ling+Phon+Pros)	ʒ	o	ʁ	e	d	y	a	k	s	ɛ	p	t	e	k	i	l	m	ə	ʁ	a	k	õ	p	a	t	ə

Table 7: Emotional pronunciation: *"Qui peut bien m'avoir laissé ce message? Je ne vois vraiment pas."* ("Who may have left me this message? I really do not see."

Expressive utterance (surprise):																																				
Canonical	k	i	p	ø	b	j	ɛ̃	m	a	v	w	a	ʁ	l	ɛ	s	e	s	ø	m	e	s	a	ʒ	ə	ʒ	ø	n	ø	v	w	a	v	ʁ	ɛ	m	ɑ̃	p	a
Target	k	i	p	o	b	j	ɛ̃	m	a	v	w	a	ʁ	l	e	s	e	s	-	m	e	s	a	ʒ	ə	ʒ	-			v	w	a	v	ʁ	e	m	ɑ̃	p	a
Vo (Ling+Phon)	k	i	p	ø	b	j	ɛ̃	m	a	v	w	a	ʁ	l	e	s	e	s	ø	m	e	s	a	ʒ	-	ʒ	ø	n	ø	v	w	a	v	ʁ	ɛ	m	ɑ̃	p	a
Vo (Ling+Phon+Pros)	k	i	p	ø	b	j	ɛ̃	m	a	v	w	a	ʁ	l	e	s	e	s	-	m	e	s	a	ʒ	ə	ʒ	-			v	w	a	v	ʁ	ɛ	m	ɑ̃	p	a
Vo+Exp (Ling+Phon)	k	i	p	ə	b	j	ɛ̃	m	a	v	w	a	ʁ	l	e	s	e	s	-	m	e	s	a	ɛ	-	ʒ	-			v	w	a	v	ʁ	e	m	ɑ̃	p	a
Vo+Exp (Ling+Phon+Pros)	k	i	p	ə	b	j	ɛ̃	m	a	v	w	a	ʁ	l	e	s	e	s	-	m	e	s	a	ʒ	ə	ʒ	-			v	w	a	v	ʁ	ɛ	m	ɑ̃	p	a
Exp (Ling+Phon+Pros)	k	i	p	ə	b	j	ɛ̃	m	a	v	w	a	ʁ	l	ɛ	s	e	s	-	m	e	s	a	ʃ	ə	ʒ	-			v	w	a	v	ʁ	e	m	ɑ̃	p	a

existence of an emotional pronunciation. The question of an emotion-specific pronunciation cannot be answered with a similar analysis since the input texts differ for each state. To investigate this issue, emotional pronunciations for each emotional state are generated automatically using the same input text from the *TelecomVo* evaluation subcorpus and Vo+Exp models (as they have shown better results with neutral input text). A cosine similarity is computed on the confusions obtained between *canonical* and generated emotional sequences for each emotion in cross-validation.

The cosine similarity is usually used to measure the similarity between text documents. Term-specific (e.g. word) weights in the document can be represented with a TF-IDF value [50]: TF being the frequency term (eq. 2) and IDF the inverse document frequency (eq. 3).

$$\mathrm{TF}(t) = \frac{\text{Nb of term } t \text{ in a document}}{\text{Total nb of terms in the document}} \tag{2}$$

$$\mathrm{IDF}(t) = \ln\left(\frac{\text{Total nb of documents}}{\text{Nb of documents with term } t \text{ in it}}\right) \tag{3}$$

We propose to adapt this model to measure similarity between pronunciations replacing terms by phoneme confusions. For each confusion (substitutions, deletions, insertions, e.g. /e/ → /ɛ/) obtained between a canonical pronunciation p_c and an adapted pronunciation p_1, we extract the TF-IDF.

Confusions are weighted by their TF-IDF value. For example with 12 documents, a confusion which term frequency is $TF = 0.5$ and appears in a single document is weighted by 0.54 while a confusion which term frequency is $TF = 0.05$ and appears in 11 documents is weighted by 0.0018. Thus we define a vector of confusions $\overrightarrow{C}_{p_c,p_1}$ which contains all TF-IDF values. The distance between two pronunciations p_1 and p_2 adapted from p_c is given using a cosine similarity between the two associated confusion vectors (Eq. 4). Cosine tends to 1 for similar pronunciations and to 0 for perpendicular pronunciations.

$$\cos\theta = \frac{\overrightarrow{C}_{p_c,p_1} \cdot \overrightarrow{C}_{p_c,p_2}}{\left\|\overrightarrow{C}_{p_c,p_1}\right\| \cdot \left\|\overrightarrow{C}_{p_c,p_2}\right\|} \tag{4}$$

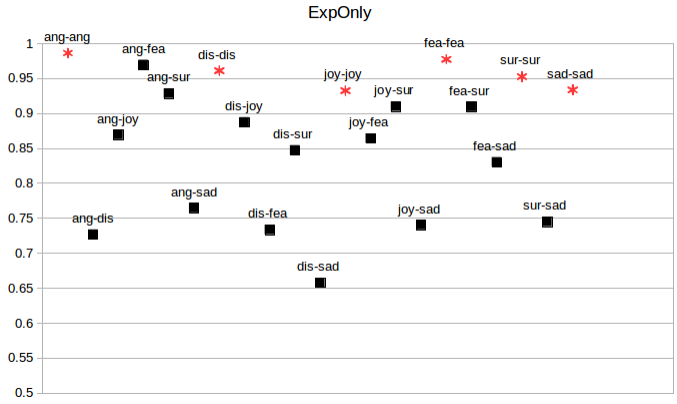


Figure 2: Cosine similarity between emotions (squares: cross-similarities and stars: auto-similarity). Confusions are evaluated from canonical phonemes between neutral utterances and predicted expressive phonemes.

Figure 2 presents cosine similarities obtained with generated emotional pronunciations using Vo+Exp protocol. The cross-validation condition allows to compute an auto-similarity, i.e. the similarity of an emotion with itself. Anger and fear seem to be the most homogeneous emotions ($\cos\theta = 0.98$) while sadness is the less ($\cos\theta = 0.93$). The second result concerns the similarity between emotions. Figure 2 shows that disgust pronunciation is very different from sadness ($\cos\theta = 0.66$) pronunciation. On the other hand, anger and fear are very similar ($\cos\theta = 0.97$). In conclusion, there are some differences in the generated emotion-specific pronunciations. Perceptual tests are needed to investigate if these differences are audible or not.

5.2.2 *Phoneme confusions and features*

A deeper analysis of the rules generated by the expressive CRFs (in both experiments) shows some very interesting results listed below. Phoneme are likely to be elided when:

- their related prosodic values are in the middle of the

scale. This result is also true in the generation of voice-specific pronunciation (Vo),
- they are related to adverbs,
- they are part of a lemma, stem or word which are common. Surprisingly word frequencies features have not been selected a lot,
- they are onset phonemes (first phonemes of a syllable) or part of open syllables located in the middle of a word.

Assimilations generally occur when prosodic features are low (energy, frequency, speech rate, decreasing F_0 shape, duration, ...), often at the end of a word and when the word is barely used in the corpus, for example /v/ → /f/ (sadness, joy, anger) or /z/ → /s/ (sadness). On the contrary, some other assimilations occur when prosodic features are high and usually at the beginning of a word: /s/ → /ʃ/. It seems like there are weak and strong assimilations according to prosody.

Two examples extracted from *EmotionPron* corpus are shown in Tables 6 and 7. Different types of adaptation are noticeable: inconsistencies in alphabets and pronunciation modifications. Liaisons and schwa elisions are not detailed in this paper. Concerning labeling strategies and alphabet choices, Table 6 shows that Vo and Vo+Exp models are able to map the *canonical* /ɲ/ to the *target* /n, j/ as labeled in the *TelecomVo* corpus. However, the Exp model is not able to do this adaptation. Concerning emotional pronunciation, in Table 6 /ø/ is present in neutral sequences (canonical and Vo) as the phonetizer outputs while /ə/ has been predicted in emotional sequences (Vo+Exp and Exp) as annotated in the corpus. The transformation of /ʒ/ in /ʃ/ (Table 7) is representative of emotional pronunciation, however Vo and Vo+Exp models do not model it while Exp does. It means that such a configuration probably never occurs in the *TelecomVo* corpus and would probably lead to poor quality speech sample.

In order to better analyse the influence of prosodic features in pronunciation modelling, pronunciations of Tables 6 & 7 are generated with and without prosodic features (feature subsets S_E^0 being adapted to each case). Table 7 presents an interesting case: the *canonical* pronunciation of /ʒ ø n ø/ is transformed in /ʒ/ in the realization of an emotional pronunciation. This is a basic case in spontaneous speech: the deletion of /ø/ gives /ʒ n ø/ and the deletion of the negative French *ne* gives the final pronunciation. For this utterance, prosodic features are all in the middle of the scale, therefore phonemes are likely to be deleted. Consequently, the deletion occurs in the Vo sequence generated with prosodic features, while all phonemes remain in the Vo sequence generated without prosodic features. The problem is that, once phonemes are deleted, it is uneasy to insert them again in the emotional pronunciation. In the case of insistence, the presence of these phonemes could be necessary, for example under anger or surprise affective states. To conclude, prosodic features are very important for elisions in emotional pronunciations. In the present work, it seems relevant that voice adaptation models are trained without prosody – thus allowing few elisions – then that emotional pronunciation models are trained with prosody thus allowing emotion-dependent elisions.

6 Perceptual tests

Because our aim is to improve ESS, the impact of pronunciation adaptation on the quality and expressivity of synthetic speech is evaluated through perceptual tests. With this aim in mind, neutral and expressive sentences are phonetized using our pronunciation adaptation framework, synthesized by querying the same dedicated voice speech database with a unit selection TTS system[3], and finally evaluated.

6.1 Protocol

Two perceptual tests were conducted with 11 participants each. Synthetic speech samples were presented randomly to the participants. The evaluation is based on AB tests with 60 utterances in which the listeners have to answer the three following questions:
- "Between A and B, which sample reaches the best quality?" (*A, B, no preference*);
- "Between A and B, which sample is the most expressive?" (*A, B, no preference*);
- "For the most expressive sample, which emotion is expressed?" (*No emotion, an emotion that I do not recognize, Anger, Disgust, Joy, Fear, Surprise, Sadness*).

In the first [resp. the second] test, utterances were randomly selected by sub-sampling *EmotionPron* corpus [resp. the *TelecomVo* evaluation subcorpus] according to the PER between canonical and target expressive [resp. neutral] pronunciation. Speech samples were synthesized using the corpus-based TTS system described in [51]. Whatever the pronunciation, the voice corpus is always the *TelecomVo* training subcorpus. Five pronunciations are evaluated: canonical pronunciation without adaptation (Cano), target pronunciation (Target), adapted pronunciation using *TelecomVo* (70%) only (Vo), using *EmotionPron* only (Exp) and both *TelecomVo* (70%) and *EmotionPron* corpora (Vo+Exp). The number of preferred samples in terms of quality, expressivity and emotion are reported in Table 8a (expressive text) and 8b (neutral text)[4].

6.2 Quality assessment

In the previous section, the authors supposed that models using both adaptation to the voice corpus and to the emotional pronunciation would lead to better quality pronunciation that if the adaptation is done to emotional pronunciation only. The perception results show that samples synthesized with Vo pronunciation applied to neutral and expressive sentences (Table 8a and 8b) are preferred in terms of quality (54% with expressive text and 65% with neutral text). It also means that emotional pronunciation adaptation degrades quality from voice-specific adaptation. Moreover Exp and Vo+Exp pronunciations are judged as similar when applied to expressive text while Vo+Exp is preferred when applied to neutral text, thus confirming the interest of a double adaptation (Vo+Exp) for quality assessment.

3. Synthetic speech samples are available at https://www-expression.irisa.fr/files/2018/03/exs_expressive_pronunciation_adaptation.zip

4. Considering a normal law, the confidence interval at 50% is ±3.7%

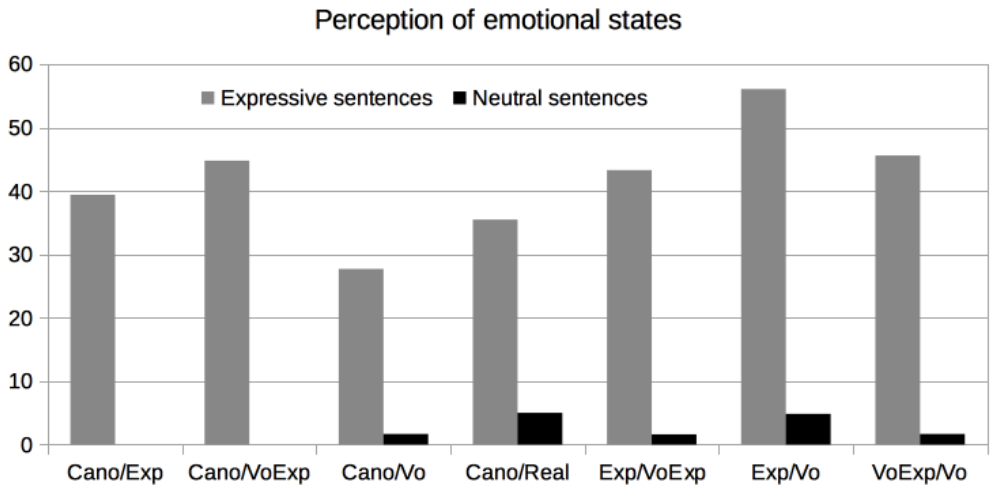


Figure 3: % of samples judged as emotional

Table 8: Preferred samples (%). Ex: 47% of participants prefer Exp pronunciation in terms of quality and 50% of the participants do not find any differences between Cano and Exp pronunciations in terms of expressivity.

(a) for expressive input text.

Samples		Quality			Expressivity		
A	B	A	B	=	A	B	=
Cano	Exp	27	**47**	26	12	38	**50**
Cano	Vo+Exp	30	**39**	31	27	**43**	30
Cano	Vo	9.2	**54**	37	28	22	**51**
Cano	Target (emot.)	**48**	23	29	18	34	**48**
Exp	Vo+Exp	21	29	**50**	19	24	**57**
Exp	Vo	21	**52**	27	36	29	35
Vo+Exp	Vo	28	**47**	25	**41**	24	35

(b) Preferred samples (%) for neutral input text.

Samples		Quality			Expressivity		
A	B	A	B	=	A	B	=
Cano	Exp	37	**40**	23	20	23	**57**
Cano	Vo+Exp	32	**48**	20	20	30	**50**
Cano	Vo	5.0	**65**	30	22	30	**48**
Cano	Target (neutral)	8.3	**67**	25	22	30	**48**
Exp	Vo+Exp	22	31	**47**	11	26	**63**
Exp	Vo	7.8	**56**	36	21	17	**62**
Vo+Exp	Vo	13	41	**46**	15	15	**70**

6.3 Expressivity assessment

Regarding the results obtained with expressive text (Table 8a), participants judged the Vo+Exp pronunciation as more expressive than other pronunciations, especially against Cano (43%) and Vo (41%) pronunciations. Concerning neutral utterances (Table 8b), where neither the voice nor the text is expressive but the pronunciation, expressivity is almost not perceived. Consequently, the proposed double adaptation (to the voice and to an emotional pronunciation) is able to improve simultaneously the quality and the expressivity of the overall synthesized signal on expressive sentences.

6.4 Emotion perception

The results of the emotion question confirm the fact that pronunciation alone does not support emotional information: testers perceive an emotion (74% on average over all AB tests) and are able to find the correct label when input text is expressive (Fig. 3) whereas they usually do not perceive any emotion (36%) and when they perceive one, they are not able to identify the affect with pronunciation only.

In conclusion of perceptual tests: (1) the double adaptation (to the voice and to an emotional pronunciation) is able to improve simultaneously quality and expressivity with expressive sentences, (2) a simple adaptation to the voice improves quality with neutral sentences, (3) when voice and sentences are neutral, no expressivity (nor emotion) is perceived even with an emotional pronunciation.

7 Conclusion

This article shows that the proposed double adaptation (to the voice and to an emotional pronunciation) is able to improve simultaneously the quality and the expressivity of the overall synthesized signal on expressive sentences. This method has the advantage of being relevant with small pronunciation databases. Regarding the semantic content, the perception of expressivity is improved when the text is also expressive.

The analysis we conducted on emotional speech enables to characterize an emotional pronunciation. It shows that prosodic features play a significant role in the deletion and assimilation of phonemes. The study also describes phoneme assimilations as weak or strong according to the values of prosodic features. Finally, objective measures show evidences of the clear existence of an emotional pronunciation with respect to target or canonical pronunciations, whereas the existence of emotion-specific pronunciation is more subtle. These results are very useful in the linguistic field but also in the design of emotional databases.

Our study raises the issue of the integration of prosody in the generation of pronunciation variants as well as the relation between linguistics, suprasegmental prosodic features and pronunciation. Finally, we have shown that emotion-specific pronunciations were probably not relevant, however it is known that emotional states greatly influence prosodic and acoustic parameters. In future work, we plan to combine a pronunciation adaptation with prosodic and acoustic DNN

or HMM-based adaptation framework. We also intend to automatically extract expressivity (and emotional states) and to predict prosodic features directly from text.

Acknowledgments

This study has been realized under the ANR (French National Research Agency) project SynPaFlex ANR-15-CE23-0015.

References

[1] Z. Handley, "Is text-to-speech synthesis ready for use in computed-assisted language learning ?" *Speech Communication*, vol. 51, no. 10, pp. 906–919, 2009.

[2] Y. Gao and W. Zhu, "Detecting affective states from text based on a multi-component emotion model," *Computer Speech and Language*, vol. 36, pp. 42–57, 2016.

[3] N. Campbell, *Expressive/Affective Speech Synthesis*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 505–518.

[4] M. Tahon, R. Qader, G. Lecorvé, and D. Lolivc, "Improving TTS with corpus-specific pronunciation adaptation," in *Interspeech*, San Fransisco, USA, 2016.

[5] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1–2, pp. 227–256, 2003.

[6] P. Ekman, *Basic Emotions*. Wiley, New-York, 1999, ch. c, pp. 301–320.

[7] S. Mariooryad and C. B. A. c, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Transactions in Affective Computing*, vol. 6, no. 2, pp. 97–108, 2015.

[8] J. Russel, *Reading emotions from and into faces: Resurrecting a dimensional-contextual perspective*. Cambridge University Press, U.K., 1997, ch. @InbookRussel1997, Title = Reading emotions from and into faces: Resurrecting a dimensional-contextual perspective, Author = J. Russel, Pages = 295–360, Publisher = Cambridge University Press, U.K., Year = 1997, Booktitle = The psychology of facial expression, Owner = tahon, Timestamp = 2017.01.20 , pp. 295–360.

[9] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, pp. 1162–1181, 2006.

[10] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Interspeech*, Lissabon, Portugal, September 2005, pp. 1517–1520.

[11] N. Campbell, "CHATR the corpus: A 20-year-old archive of concatenative speech synthesis," in *International Conference on Language Resources and Evaluation (LREC)*, Prortoroz, Slovenia, May 2016, pp. 3436–3439.

[12] D. Govind and S. Prasanna, "Expressive speech synthesis: A review," *International Journal of Speech Technology*, vol. 16, pp. 237–260, 2013.

[13] G. Beller, X. Rodet, and C. Veaux, "IrcamCorpusExpressivity: Nonverbal words and restructurings," in *LREC Workshop on Emotions*, 2008.

[14] K. Bartkova, D. Jouvet, and E. Delais-Roussarie, "Prosodic parameters and prosodic structures of French emotional data," in *Speech Prosody*, Shanghai, China, 2016.

[15] S. King and V. Karaiskos, "The Blizzard Challenge 2016," in *Blizzard Challenge (satellite of Interspeech)*, 2016.

[16] S. King, L. Wihlborg, and W. Guo, "The Blizzard Challenge 2017," in *Blizzard Challenge (satellite of Interspeech)*, 2017.

[17] M. Charfuelan and I. Steiner, "Expressive speech synthesis in MARY TTS using audiobook data and EmotionML," in *Interspeech*, Lyon, France, August 2013.

[18] M. Schröder, *Expressive Speech Synthesis: Past, Present, and Possible Futures*. London: Springer London, 2009, pp. 111–126.

[19] H. Kanagawa, T. Nose, and T. Kobayashi, "Speaker-independent style conversion for HMM-based expressive speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 7864–7868.

[20] Y.-Y. Chen, C.-H. Wu, and Y.-F. Huang, "Generation of emotion control vector using MDS-based space transformation for expressive speech synthesis," in *Interspeech*, San Fransisco, USA, September 2016, pp. 3176–3180.

[21] L. Chen, N. Braunschweiler, and M. J. F. Gales, "Speaker and expression factorization for audiobook data: Expressiveness and transplantation," *Transactions on Audio, Speech and Language Processing (IEEE/ACM)*, vol. 23 (4), pp. 605–618, 2015.

[22] P. Alain, J. Chevelu, D. Guennec, G. Lecorvé, and D. Lolive, "The IRISA Text-To-Speech system for the Blizzard Challenge 2016," in *Blizzard Challenge (satellite of Interspeech)*, 2016.

[23] I. Steiner, M. Schröder, M. Charfuelan, and A. Klepp, "Symbolic vs. acoustics-based style control for expressive unit selection," in *ISCA Speech Synthesis Workshop (SSW7)*, Kyoto, Japan, 2010.

[24] M. S. Ribeiro, O. Watts, J. Yamagishi, and R. A. J. Clark, "Wavelet-based decomposition of f0 as a secondary task for dnn-based speech synthesis with multi-task learning," in *ICASSP (IEEE)*, 2016.

[25] S. Pammi and M. Charfuelan, "HMM-based sCost quality control for unit selection speech synthesis," in *ISCA Speech Synthesis Workshop*, Barcelona, Spain, September 2013, pp. 53–57.

[26] T. Fukada, T. Yoshimura, and Y. Sagisaka, "Automatic generation of multiple pronunciation based on neural networks." *Speech Communication*, vol. 27, pp. 63–73, 1999.

[27] R. Dall, S. Brognaux, K. Richmond, C. Valentini-Botinhao, G. E. Henter, J. Hirschberg, J. Yamagishi, and S. King, "Testing the consistency assumption: Pronunciation variant forced alignment in read and spontaneous speech synthesis," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, March 2016, pp. 5155–5159.

[28] P. Karanasou, F. Yvon, T. Lavergne, and L. Lamel, "Discriminative training of a phoneme confusion model for a dynamic lexicon in ASR," in *Interspeech*, Lyon, France, August 2013, pp. 1966–1970.

[29] L. Lu, A. Ghoshal, and S. Renals, "Acoustic data-driven pronunciation lexicon for large vocabulary speech recognition," in *Automatic Speech Recognition and Understanding Workshop (ASRU)*, Olomouc, Czech Republic, December 2013.

[30] G. Lecorvé and D. Lolive., "Adaptive statistical utterance phonetization for French." in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 4864–4868.

[31] T. J. Hazen, I. L. Hetherington, H. Shu, and K. Livescu, "Pronunciation modeling using a finite-state transducer representation," *Speech Communication*, vol. 46, no. 2, pp. 189–203, 2005.

[32] R. A. Bates, M. Osendorf, and R. A. Wright, "Symbolic phonetic features for modeling of pronunciation variation," *Speech Communication*, vol. 49, pp. 83–97, 2007.

[33] K. Livescu, P. Jyothi, and E. Fosler-Lussier, "Articulatory feature-based pronunciation modeling," *Computer Speech and Language*, vol. 36, pp. 212–232, 2016.

[34] B. Vazirnezhad, F. Almasganj, and S. Ahadi, "Hybrid statistical pronunciation models designed to be trained by a medium-size corpus," *Computer Speech and Language*, vol. 23, no. 1, pp. 1–24, 2009.

[35] A. Bell, D. Jurafsky, E. Fosler-Lussier, C. Girand, M. Gregory, and D. Gildea, "Effects of disfluencies, predictability, and utterance position on word form variation in English conversation," *J. Acoust. Soc. Am.*, vol. 113, pp. 1001–1024, 2003.

[36] A. Bell, J. Brenier, M. Gregory, C. Girand, and D. Jurafsky, "Predictability effects on durations of content and function words in conversational English," *Journal of Memory and Language*, vol. 60, no. 1, pp. 92–111, 2009.

[37] M. Adda-Decker, P. B. de Mareüil, G. Adda, and L. Lamel, "Investigating syllabic structures and their variation in spontaneous French," *Speech Communication*, vol. 46, pp. 119–139, 2005.

[38] P. B. de Mareüil and M. Adda-Decker, "Studying pronunciation variants in French by using alignment techniques," in *International Conference on Spoken Language Processing*, Denver, Colorado, USA, September 2002.

[39] C. L. Bennett and A. W. Black, "Prediction of pronunciation variation for speech synthesis: A data-driven approach," in *ICASSP*, 2005, pp. 297–300.

[40] S. Brognaux, B. Picart, and T. Drugman, "Speech synthesis in various communicative situations: Impact of pronunciation variations," in *Interspeech*, September 2014, pp. 1524–1528.

[41] R. Qader, G. Lecorvé, D. Lolive, and P. Sébillot, "Probabilistic speaker pronunciation adaptation for spontaneous speech synthesis using linguistic features," in *International Conference on Statistical Language and Speech Processing (SLSP)*, Budapest, Hungary, 2015.

[42] J. Chevelu, G. Lecorvé, and D. Lolive, "ROOTS: A toolkit for easy, fast and consistent processing of large sequential annotated data collections," in *9th International Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland, May 2014.
[43] M. Tahon, R. Qader, G. Lecorvé, and D. Lolive, "Optimal feature set and minimal training size for pronunciation adaptation in TTS," in *International Conference on Statistical Language and Speech Processing (SLSP)*, Pilzen, Czech Republic, 2016.
[44] F. Béchet, "LIA-PHON: Un système complet de phonétisation de texte," *Traitement Automatique des Langues (TAL)*, vol. 42, no. 1, pp. 47–67, 2001.
[45] Y. Lin, J.-B. Michel, E. L. Aiden, J. Orwant, W. Brockman, and S. Petrov, "Syntactic annotations for the google books ngram corpus," in *50th Annual Meeting of the Association for Computational Linguistics*, Jeju, Republic of Korea, July 2012, pp. 169–174.
[46] C. d'Alessandro, S. Rosset, and J.-P. Rossi, "The pitch of short-duration fundamental frequency glissandos," *J. Acoust. Soc. Am.*, vol. 104, pp. 2339–2348, 1998.
[47] T. Lavergne, O. Cappé, and F. Yvon, "Practical very large scale CRFs," in *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, Uppsala, Sweden, 2010, pp. 504–513.
[48] M. Adda-Decker and P. Hallé, "Bayesian framework for voicing alternation & assimilation studies on large corpora in French," in *15th international congress of phonetic sciences,*, Barcelona, Spain, 2007, pp. 613–616.
[49] N. B. Abdelli-Beruh, "Voicing and devoicing assimilation of French /s/ and /z/," *Journal of Psycholinguistic Research*, vol. 41, pp. 371–386, 2012.
[50] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24 (5), pp. 513–523, 1988.
[51] D. Guennec and D. Lolive, "Unit selection cost function exploration using an A* based Text-to-Speech system," in *17th International Conference on Text, Speech and Dialogue*, 2014, pp. 449–457.

Marie Tahon is currently Associate Professor at Le Mans University and conducts her research at LIUM (France). She graduated in engineering from the Ecole Centrale de Lyon (France) in 2007 and received the M.S. degree in acoustics from the Ecole Centrale de Lyon, in 2007. She received the Ph.D. degree in computer science from the University of Paris-Sud (Orsay, France) in 2012. She has been with the LIMSI-CNRS (Orsay, France) working on affective computing, with the LMSSC, CNAM (Paris, France) working on acoustics. Recently, she was with the IRISA (Lannion, France) in the team "Expression" where she conducted the present work. Her research interests concern automatic speech processing for expressive speech. She is a member of the French Association of Spoken Communication (AFCP) and of the French Acoustic Association (SFA).

Gwénolé Lecorvé is an Associate Professor at the applied science engineering school ENSSAT, University of Rennes 1, conducting his research as a member of the team "Expression" at IRISA (France). He graduated a B.Sc. degree in mathematics and applied sciences at the University of Bretagne Sud (Lorient, France) in 2004, a M.Eng. in computer science, and a M.Sc. in image processing and artificial intelligence at the Institut National des Sciences Appliquées (INSA Rennes, France) in 2007. He received his Ph.D. in computer science in 2010 from INSA Rennes as a member of the multimedia group of IRISA/INRIA (Rennes, France), working on linguistic adaptation of automatic speech recognition systems. His research interests are focused on variability in natural language and speech. He is a member of the directory board of the French Association of Spoken Communication (AFCP).

Damien Lolive is currently an Associate Professor at the applied science engineering school ENSSAT, University of Rennes 1. He conducts his research with the team "Expression" at the IRISA Laboratory (France). He graduated a M.Eng. in computer science at ENSSAT, France, and a M.Sc. in artificial intelligence at University of Rennes 1, France in 2005. He received his Ph.D. degree in computer science from University of Rennes 1, France, in 2008, as a member of the "Cordial" team, working on speech prosody and voice conversion systems. His main research interests include speech prosody, text-to-speech synthesis, natural language and speech interaction. He is a member of the directory board of the French Association of Spoken Communication (AFCP).