Perception of expressivity in TTS: linguistics, phonetics or prosody?

Marie Tahon, Gwénolé Lecorvé, Damien Lolive, Raheel Qader

Perception of expressivity in TTS: linguistics, phonetics or prosody?

Marie Tahon*, Gwénolé Lecorvé, Damien Lolive, and Raheel Qader

IRISA/University of Rennes 1
6 rue de Kérampont, 22300 Lannion, France
{marie.tahon,gwenole.lecorve,damien.lolive,raheel.qader}@irisa.fr
https://www-expression.irisa.fr/

Abstract. Actually a lot of work on expressive speech focus on acoustic models and prosody variations. However, in expressive Text-to-Speech (TTS) systems, prosody generation strongly relies on the sequence of phonemes to be expressed and also to the words below these phonemes. Consequently, linguistic and phonetic cues play a significant role in the perception of expressivity. In previous works, we proposed a statistical corpus-specific framework which adapts phonemes derived from an automatic phonetizer to the phonemes as labelled in the TTS speech corpus. This framework allows to synthesize good quality but neutral speech samples. The present study goes further in the generation of expressive speech by predicting not only corpus-specific but also expressive pronunciation. It also investigates the shared impacts of linguistics, phonetics and prosody, these impacts being evaluated through different French neutral and expressive speech collected with different speaking styles and linguistic content and expressed under diverse emotional states. Perception tests show that expressivity is more easily perceived when linguistics, phonetics and prosody are consistent. Linguistics seems to be the strongest cue in the perception of expressivity, but phonetics greatly improves expressiveness when combined with and adequate prosody.

Keywords: Expressive speech synthesis, Perception, Linguistics, Phonetics, Prosody, Pronunciation adaptation.

1 Introduction

Speech synthesis usually consists of the conversion of a written text to a speech sound, also named as Text-To-Speech (TTS) process. While TTS has reached a fairly acceptable level of quality and intelligibility on neutral speech in the last decades, the lack of expressivity is often criticized, as it usually sounds different from spontaneous human conversations [1]. The shift of TTS from read to spontaneous and expressive speech would greatly help to reproduce situations where the synthetic voice talks with a user, for instance in the field of human-machine interactions. As a result, there is a crucial need not only for just

* Corresponding author

intelligible speech carrying linguistic information, but also for expressive speech. The present study investigates affective speech for TTS and finds applications in many domains such as education and entertainment. According to Campbell [2], the main challenge in expressive TTS is to find the adequation of affective states in the input and the realization of prosodic characteristics to express them in the output speech. Undoubtedly, prosody is an important cue in the perception of expressivity in speech. However, in the framework of expressive speech synthesis, prosody is highly related to the sequences of phonemes to be expressed and to the words below these phonemes. Therefore, linguistic and phonetic cues also play a significant role in the perception of expressivity. The present work investigates the shared impacts of linguistics, phonetics and prosody in the perception of quality and expressivity of speech samples generated with a TTS system.

Two main data-driven approaches coexist for TTS [1], unit selection and statistical parametric systems, and both require variable affective speech data of good audio quality. In that sense, there is a growing interest for audio books as shown by the Blizzard Challenge 2016 [3]. They 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 [4]. In the present study, three speech corpora with different levels of expressivity are used, one being collected from an audio book, another from high quality speech for synthesis, and the last from TV commentaries. A solution to introduce some 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 [5, 6]. Expressivity can also be controlled in symbolic terms (diphone identity, position, etc.) [7] or in prosodic terms (fundamental frequency, energy, duration) [8]. Those elements are usually used in the speech synthesizer directly in the cost function or in the construction of the acoustic model [9]. In addition, voice transformation techniques can be applied to synthetic samples [10, 11]. The TTS used in this paper is a neutral unit selection system [7], expressivity being controlled with different types of text, pronunciation and speech databases.

While a lot of work on expressive speech focus on acoustic models and prosody variations, very few of them deal with pronunciation. A perception study [12] showed that samples synthesized with the *realized* pronunciation were preferred to those synthesized with the pronunciation derived from an automatic phonetizer – the *canonical* pronunciation. In previous works, we proposed a statistical framework which adapts the *canonical* pronunciation to a target pronunciation. This framework allows to predict phoneme sequences by using Conditional Random Fields (CRF) models trained on linguistic, phonological, articulatory and prosodic features. The framework was used to generate spontaneous English pronunciations and the results show that a trade-off between quality and intelligibility is necessary [13]. It was also used to predict a corpus-specific pronunciation, i.e. a pronunciation adapted to the TTS voice corpus, thus conducting to a significant improvement of the overall quality of synthesized speech [14, 15]. In the work realized in [14], we manage to synthesize good quality speech samples on a neutral voice. The present study goes further in the generation of expres-

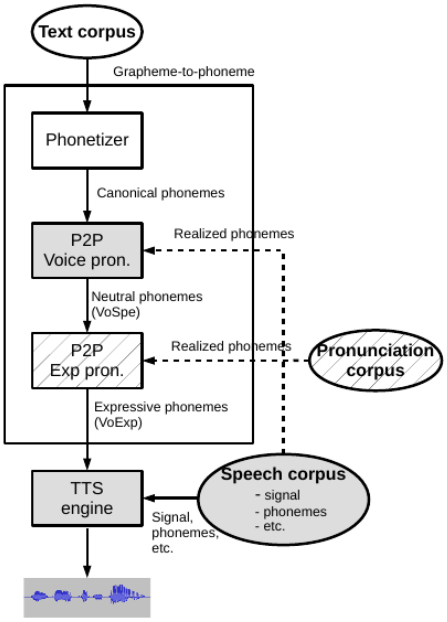

Fig. 1: General overview. Databases are symbolized with ellipses.

sive speech samples by predicting not only a corpus-specific pronunciation but also an expressive pronunciation. We also investigate the shared impacts of linguistics, phonetics and prosody on the perception of expressivity, as well as the best configuration towards an expressive synthesis system. In the remainder, an overview of the general process is presented in section 2. Speech, pronunciation and text databases are detailed in section 3. Features and models are exposed in section 4. Finally, section 5 presents the perception test protocol and results.

2 General overview

The process used in this study has been set up in order to study the impact of linguistic, phonetic and prosodic expressive variations on the perception of expressivity. Expressive variations of linguistic and prosodic features are easily managed through different corpora, whereas expressive pronunciation variants need to be generated with a pronunciation adaptation framework as illustrated in Figure 1. As detailed in [14], the goal of pronunciation adaptation is to reduce the differences between phonemes derived from a phonetizer (*canonical*) and phonemes as labelled in the pronunciation corpus (*realized*). To do so, the proposed method is to train CRFs phoneme-to-phoneme (P2P) models which predict adapted phonemes from *canonical* ones. To go further towards expressive pronunciation generation, this study combines two P2P models. The voice-specific pronunciation P2P model is trained on the TTS speech corpus with *canonical*

Table 1: Characteristics of each database. Mean (standard deviations) of fundamental frequency (F_0) in semitone and speech rate (SR) in syllable per seconds are given.

Corpus	Expressivity	# utt.	Dur.	# phon.	F_0(st)	SR
Speech corpora						
Telecom - train 70%	Neutral	5044	4h51'	151,945	89 (2.7)	4.7 (2.1)
Audiobook	Moderate	3339	10h45'	379,897	77 (3.2)	6.3 (1.2)
Commentary	Expressive	1631	5h25'	173,858	85 (5.0)	6.0 (1.7)
Pronunciation corpus						
Expressive	Expressive	6×47	0h41'	16,248	84 (7.1)	6.3 (1.8)
Text corpora						
Telecom - eval 30%	Neutral	2162	2h04'	64,960		
Expressive	Expressive	6×47	0h41'	16,248		

phonemes and predicts neutral VoSpe phonemes. The expressive pronunciation P2P model is trained on the pronunciation corpus with VoSpe phonemes and predicts expressive VoExp phonemes. One could argue that adaptation could have been realized without any voice-specific adaptation. Such a method could probably improve the expressiveness of the synthesized speech samples, but inconsistencies between speech and pronunciation corpora would remain, thus lowering the TTS quality. Overcoming the disadvantages of the aforementioned method, the protocol illustrated in Figure 1 was designed to generate expressive speech samples of good quality. Adapted VoSpe and VoExp pronunciations are evaluated with expressive and with neutral utterances. Such a protocol is of interest in evaluating the influence of words in the perception of expressivity. Finally, three different speech corpora are used to create TTS voices, each one having its own prosodic characteristics.

3 Databases

This section presents the databases used in the following experiments, which characteristics are given in Table 1. Three speech corpora are used for voice-specific pronunciation modelling and in the TTS voice creation. An emotional pronunciation corpus is used for expressive pronunciation modelling. Finally, utterances from two subcorpora of the aforementioned databases are used to evaluate the influence of linguistics.

3.1 Speech corpora

Speech corpora are used to train voice-specific pronunciation P2P models. They are also used to create TTS voices.

Telecom corpus features a French speech corpus dedicated to interactive vocal servers. As such, this corpus covers all diphonemes present in French. Phonemes and non speech sounds have been manually corrected. The *Telecom* corpus has been randomly split in two subsets. 70% are left for training purposes and the remaining 30% are kept for evaluations. This corpus comprises most

words used in the telecommunication field. Utterances are neutral such as: *"On nous invite à visiter les églises de onze heures à trois heures."* ("We are pleased to visit the churches from 11 a.m. to 3 p.m."). It features a neutral female voice which pitch is normal (170 Hz, 89 st) and pitch standard deviation is quite small. The speech rate is in the normal range according to [16]. According to these prosodic characteristics, this corpus can be considered as little expressive.

Audiobook corpus is extracted from a French audio book [17]. The reader is a professional male French actor with a very low voice (91 Hz, 77 st). The book "Albertine disparue" was written by the French author Marcel Proust. Data was automatically annotated using the process described in [18]. Since the main topic is an analysis of love pain, the tone is mainly serious, as this example suggests: *"Alors je pleurais ce que je voyais si bien et qui, la veille, n'était pour moi que néant."* ("Then I was crying what I was seeing so well, and what, before, was for me just a void"). Compared to the *Telecom* corpus, pitch variations are more important, speech rate is also faster. This corpus is then considered as moderately expressive.

Commentary corpus is extracted from commentaries which precede science-fiction French series. The male speaker presents the synopsis of each episode in a very expressive way. Data was also automatically annotated using the process described in [18]. The commentator often calls out to the audience, and gets it interested in viewing the episode. For example, he says: *"Qu'avez-vous pensé de ce géant qui s'avère être une tour humaine formée par trois acrobates ? Réalité, ou fiction ?"* ("What did you think of this giant who turns out to be a human tour made of three acrobats? Reality or fiction?"). In this corpus, the global pitch is quite high (136 Hz, 85 st) for a male speaker, and the variations are important, revealing a large diversity in prosody. The speech rate and its variations are at the same level as in the *Audiobook* corpus. For these reasons, this corpus is the most expressive.

3.2 Pronunciation corpus

The pronunciation expressive corpus is used to train expressive pronunciation models for each emotion.

Expressive corpus has been collected for expressive synthesis purposes. A male speaker recorded French sentences in various emotion styles with a high activation degree: anger, disgust, joy, fear, surprise and sadness. The speech material has been aligned to the corresponding text for prosodic analysis and alignment has been manually corrected [19]. The linguistic content of the sentences is informal and emotionally coloured language, as for example in the expression of anger: *"Oh ! Merde ! Il y a un bouchon, c'est pas vrai, on va encore louper le début du film !"* ("Oh! Shit! There is traffic, I can't believe it, we are going to miss the beginning of the film again!"). The choice of such sentences greatly helps the speaker to simulate an emotion while reading. Each of 6 expressive pronunciation model will be trained and evaluated in cross-validation using the 47 available utterances per emotion. Unsurprisingly, pitch and energy are highly variable throughout the corpus and the speech rate is as fast as in *Audiobook*.

The *Expressive* corpus offers the opportunity to study expressed pronunciations for different emotional states, this aspect being left for further studies.

3.3 Text corpora

120 utterances were randomly selected from *Telecom-eval* and *Expressive* corpora by sub-sampling the corpus according to the Phoneme Error Rate (PER) between *canonical* and *realized* pronunciations. These utterances will be used as neutral and expressive input text to evaluate the models. The 60 utterances selected from *Telecom-eval* differ from the utterances used to train the voice-specific pronunciation model and to create the TTS voice. On the contrary, due to the small size of the corpus, the 60 utterances selected from *Expressive* corpus are also used to train the expressive pronunciation model. Therefore, this corpus has been split in 5 folds and managed under cross-validation conditions.

4 P2P models

Voice-specific and expressive phoneme sequences are predicted using CRFs as pronunciation models. This section describes the features, then voice-specific and expressive pronunciation CRF models.

4.1 Features

CRFs are trained using the Wapiti toolkit [20] with the default BFGS algorithm on a speech or pronunciation corpus with different features. Precisely, as detailed in [14], four groups of features were investigated: 26 linguistic, 17 phonological, 9 articulatory and 8 prosodic features. Relevant features for pronunciation adaptation are then automatically selected according to a cross-validation protocol. Prosodic features are extracted in an oracle way, i.e., directly from the recorded utterances of the speech corpus. In the future, a prosodic model could be included in the synthesizer, thus making prosodic features available. Such a protocol allows to know to what extent prosody affects pronunciation models.

4.2 P2P voice-specific pronunciation model

The voice pronunciation model adapts *canonical* phonemes to phonemes as *realized* in the *speech corpus*. In previous work [14, 15], we have presented the training process of a P2P voice-specific model with the corpus *Telecom*. Table 2 shows the distribution of selected features within groups. Feature selection performed on the voice-specific model (VoSpe) excludes articulatory features. In the end, a set of 15 features including linguistic, phonological and prosodic features with a 5-phoneme window (two phonemes surrounding the current phoneme, named as W_2) were automatically selected. An optimal PER of 2.7% (baseline 11.2%) was reached when training models on the data. However, a perception test has shown that speech samples generated with the 15-feature set were perceived with the

Table 2: Number of selected features within groups with a W_0 phoneme window. Feature selection results are presented for adaptation to the voice pronunciation on *Telecom* (VoSpe) then to the expressive pronunciations on the *Expressive* corpus for each emotion (VoExp).

Feature group (# feat.)	VoSpe	VoExp					
		Anger	Disgust	Joy	Fear	Surprise	Sadness
Linguistic (26)	2	3	5	2	4	5	3
Phonological (17)	7	5	7	6	6	3	3
Articulatory (9)	0	3	4	1	1	2	2
Prosodic (8)	0 (removed)	6	6	7	5	7	8
Total (52)	9	17	22	16	16	17	16

same or a lower quality than samples generated with a 9-feature set excluding prosodic features. Since prosodic features are not generated from text yet but are estimated in an oracle way, only the selected linguistic and phonological 9-feature set is used. For the same reason, a 5-phoneme window (W_2) is applied to train voice-specific P2P models. The corpora used for training voice-specific pronunciation models are the three speech corpora described in section 3.1.

4.3 P2P expressive pronunciation model

The expressive pronunciation model adapts VoSpe phonemes which are predicted with the voice-specific pronunciation model described before, to phonemes as labelled in the *Expressive* pronunciation corpus. More precisely, 6 pronunciation models are trained for each emotion contained in the *Expressive* corpus. A greedy feature selection process is performed in 5-folds cross-validation conditions for each emotion separately starting from at least VoSpe phonemes and target *realized* expressive phonemes, then adding features one by one. Features are selected separately in the four groups and with three window sizes: W_0, W_1 and W_2. The window W_0 has shown to reach the best PER.

The number of selected features and its distribution within groups differ across emotions, as reported in Table 2, while applying W_0 on the phoneme sequence. According to Table 2, whatever the emotion, most of the prosodic features seem to be highly relevant for expressive pronunciation modeling, while articulatory features are not. Very few linguistic features were selected. Among them, word and stem are often selected, while word (disgust) and Part-of-Speech (fear) context and frequency (surprise) were selected for some emotions only. The case of sadness is interesting as all prosodic features were selected, and very few features from other groups are included in the final subset.

4.4 Objective evaluation of the models

Canonical phonemes extracted automatically from neutral (in cross-validation conditions) and expressive (in cross-corpus conditions) sentences are used as inputs to evaluate the models in terms of PER between realized expressive or neutral phonemes and predicted phonemes. The results are reported in Table 3.

Table 3: Average [standard deviation] PER (%) over emotions between canonical and predicted phonemes, with neutral and expressive text.

	Speech corpus →	Telecom		Audiobook		Comment.	
	Text corpus →	Neu	Exp	Neu	Exp	Neu	Exp
Pron.	VoSpe	3.0 [0]	16.0 [0.6]	6.9 [0]	15.6 [1.1]	6.4 [0]	16.3 [1.1]
	VoExp (W_0)	8.0 [0.8]	12.5 [2.3]	10.0 [0.5]	12.7 [0.7]	9.9 [0.5]	13.1 [1.5]
	VoExp (W_0+ sel. feat.)	9.0 [0.3]	5.1 [1.2]	10.0 [0.6]	6.0 [1.0]	9.9 [0.5]	5.1 [0.8]

In the case of neutral [resp. expressive] utterances in input, no realized expressive [resp. neutral] pronunciation is available since the corpus *Telecom* [resp. *Expressive*] was designed for neutral [resp. expressive] speech only. Therefore, with neutral [resp. expressive] text in input, the PER obtained with VoSpe is smaller [resp. higher] than the one obtained with VoExp whatever the *Voice* corpus, as shown in Table 3.

The combination of voice and expressive pronunciation models – which outputs VoExp phonemes – helps in reducing phonemic differences between the predicted and the realized expressive sequences when text is expressive. Furthermore, the addition of selected features is not of significant interest when the input text is neutral, but is when text is expressive. Average PER improvement reaches 6.7 pp. with *Audiobook*, 7.4 pp. with *Telecom* and 8.0 pp. with *Commentary* with the W_0+ selected features. A perception test will be able to evaluate the models in a similar way for both expressive and neutral text in input.

4.5 Example

Table 4 illustrates some differences which occur between a neutral and an expressive pronunciation. In this example, the *realized* pronunciation comes from *Expressive* corpus. Canonical pronunciation is adapted to the pronunciation of the speech corpus (VoSpe) then to the emotional pronunciation corpus (VoExp). Some deletions appear to be characteristic of an expressive pronunciation in French, for example deletion of the vowel /ø/ or the liquid /l/. Also the liaison /z/ is missing in the three adapted VoExp pronunciations as well as in the *realized* pronunciation. This example also presents an interesting case: the *canonical* pronunciation /ʒ ø n/ is subsituted by /ʒ/ in expressive pronunciations (see also /i n ø/. This is a regular case in French: first the deletion of /ø/ gives /ʒ n/ and the deletion of the negative *ne* gives the final pronunciation.

5 Perception test results

In this section, we present perception tests which evaluate the respective influences of linguistics, phonetics and prosody in terms of quality and expressivity.

5.1 Experimental set-up

Six perception AB tests were conducted independently. Each test combines a TTS voice built on one of the 3 speech corpora (*Telecom*, *Audiobook* or *Commentary*) and either neutral or expressive input text. Within a test, AB pairs

Table 4: Example of pronunciation adaptations. The input text is *Je suis dégoûtée, ils ne m'ont pas embauchée parce que je n'ai pas le diplôme.* "I am gutted, they did not hire me because I do not have the diploma:"

Cano:	ʒ ø s ɥ i d e g u t e i l n ø m ɔ̃ p a z ɑ̃ b o ʃ e p a ʁ s ə k ø ʒ ø n ɛ p a l ø d i p l o m ə
Real:	ʒ - s ɥ i d e g u t e i - - - m ɔ̃ p a - ɑ̃ b o ʃ e p a ʁ s - k ə ʒ - - e p a l - d i p l o m ə
Telecom	
VoSpe:	ʒ ø s ɥ i d e g u t e i l n ø m ɔ̃ p a z ɑ̃ b o ʃ e p a ʁ s - k ə ʒ ø n e p a l ø d i p l o m -
VoExp:	ʒ - s ɥ i d e g u t e i - - - m ɔ̃ p a - ɑ̃ b o ʃ e p a ʁ s - k ə ʒ - - e p a l - d i p l o m -
Audiobook	
VoSpe:	ʒ ø s ɥ i d e g u t e i l n - m ɔ̃ p a z ɑ̃ b o ʃ e p a ʁ s - k ə ʒ - n ɛ p a l ø d i p l o m -
VoExp:	ʒ - s ɥ i d e g u t e i - - - m ɔ̃ p a - ɑ̃ b o ʃ e p a ʁ s - k ə ʒ - - - p a l - d i p l o m -
Commentary	
VoSpe:	ʒ ø s ɥ i d e g u t e i l n ø m ɔ̃ p a z ɑ̃ b o ʃ e p a ʁ s - k ə ʒ ø n ɛ p a l ø d i p l o m -
VoExp:	ʒ - s ɥ i d e g u t e i - - - m ɔ̃ p a - ɑ̃ b o ʃ e p a ʁ s - k ə ʒ - - e p a l - d i p l o m -

combine 3 different pronunciations: *canonical* (Cano), adapted neutral voice-specific (VoSpe) and adapted expressive (VoExp). For each test, 11 participants were asked to judge 30 utterances per AB pair, consequently each AB pair was evaluated more than 300 times. The listeners had to answer to the questions reported in Figure 2 for the 30 utterances presented randomly. Speech samples were synthesized directly from the phoneme sequence (*canonical* or predicted with one of the 2 pronunciation models) derived from the tested input text. A TTS voice was created with the corpus-based unit selection TTS system described in [7] for each of the 3 speech corpora.

Between A and B, which sample reaches the best quality?
A, B, no differences
Between A and B, which sample is the most expressive?
A, B, no differences
For the most expressive sample, which emotion is expressed?
No emotion, an emotion that I do not recognize, anger, disgust, joy, fear, surprise, sadness, none.

Fig. 2: Perception test design.

5.2 Results

The results concerning quality are reported in Table 5, the ones concerning expressivity are reported in Table 6. The number of preferred samples in % is given for each pronunciation of the AB pair. Cases for which the two pronunciations are judged as similar are not reported. Significant preferences are annotated with a * according to the confidence interval used in [21]. VoSpe is globally preferred to Cano whatever the voice and the input text. This neutral adapted pronunciation is also judged with a better quality than the expressive adapted pronunciation (VoExp). Morover, it seems that VoExp reaches a better quality than Cano when input text is expressive rather than when the input text is neutral with *Telecom* and *Commentary* voices.

Table 5: Preferred samples (%) in terms of quality.

Text	Sample	Telecom		Audiobook		Comment.	
	A / B	A	B	A	B	A	B
Neutral	Cano / VoSpe	5.0	**65***	20	**27**	8.1	**27***
	Cano / VoExp	32	**48***	**47***	8.3	**35***	18
	VoExp / VoSpe	13	**41***	12	**34***	14	**35***
Express.	Cano / VoSpe	9.2	**54***	14	**26***	15	**28***
	Cano / VoExp	30	**39**	**46***	22	25	**32**
	VoExp / VoSpe	28	**47***	14	**44***	22	**25**

Table 6: Preferred samples (%) in terms of expressivity.

Text	Sample	Telecom		Audiobook		Comment.	
	A / B	A	B	A	B	A	B
Neutral	Cano / VoSpe	22	**30**	**22**	13	15	**26**
	Cano / VoExp	20	**30**	**23**	15	**23**	14
	VoExp / VoSpe	15	15	13	**18**	**17**	14
Express.	Cano / VoSpe	**28**	22	24	**30**	**29**	21
	Cano / VoExp	27	**43***	**34**	28	24	**29**
	VoExp / VoSpe	**41***	24	26	**33**	**26**	24

Interestingly, in cases where VoExp's quality is preferred to Cano's quality, expressivity of VoExp is also preferred: with *Telecom* voice whatever the text and with *Commentary* voice and expressive text. There is no differences in the perception of expressivity between VoSpe and VoExp while input text is neutral. With the *Audiobook* voice, Cano is preferred to VoExp both in terms of quality and expressivity. This result was expected since the phonetizer was tuned with *Audiobook* speech data. Obtained results show us that the expressive pronunciation adaptation framework improves the perception of expressivity especially as the speech corpus contains neutral speech (such as *Telecom*). Should the speech corpus be already expressive, the voice-specific pronunciation adaptation improves the global perceived quality, while expressive pronunciation adaptation does not achieve to improve the perceived expressivity, probably because expressivity is already contained in the prosody of the voice. Therefore, we show that the perception of expressivity relies on the adequation of phonetics and prosody.

With neutral text, whatever the prosody, participants are not able to recognize any emotion (correctly recognized emotions $< 5\%$ over all pronunciations). The perception of an emotional content is difficult when the expression of affect is conveyed by pronunciation only. However, they do when linguistic content is expressive (correctly recognized emotions $> 30\%$ over all pronunciations), thus meaning that emotion perception is strongly linked with linguistic content. Whilst *Commentary* has been characterized as the most expressive, the moderate voice *Audiobook* reaches the best recognition rate: average F_1 measure is 82% with *Audiobook* and only 67% with *Commentary*. Precisely, it seems that *Commentary* voice is not suitable for the expression of sadness ($F_1^{sad} = 34\%$), while *Audiobook* is ($F_1^{sad} = 74\%$). We have mentioned in section 4.3 that prosodic features were the most selected features in the expressive pronunciation model.

Same prosodic features can be used to model a sad pronunciation and can be reached by the TTS *Commentary* voice. However, even if the *Commentary* corpus is expressive, sadness is probably under-represented thus introducing the observed mismatch.

6 Conclusions

The present work evaluates the respective influence of linguistics, phonetics and prosody on the perception of quality and expressivity of synthetic speech samples. Neutral and expressive input texts, pronunciations and synthetic voices are used in a TTS system to evaluate the shared influences of these factors. The experiments confirm the interest of voice-specific adaptation for the perceived quality of TTS with different voices. Perception tests show that expressivity is better perceived when synthetic samples also have a good quality. While the perception of expressivity mainly relies on the adequation of phonetics and prosody, the perception of emotions is strongly linked with linguistics. The presented results open new perspectives in emotional data collection. In the framework of expressive speech synthesis, the use of a moderately expressive voice is of interest for the expression of affect and also for the quality of synthetic speech samples. In the future, prosodic features could be predicted directly from text, thus allowing to select appropriate speech units in the TTS voice. Further experiments are needed to label speech units according to their expressiveness, for instance with emotion recognition frameworks or speaking styles models.

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

References

1. M. Schröder: Expressive Speech Synthesis: Past, Present, and Possible Futures. In London: Springer London, pp. 111–126 (2009)
2. N. Campbell: Expressive/Affective Speech Synthesis. In Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 505–518 (2008)
3. S. King and V. Karaiskos: The Blizzard challenge 2016. In Proc. of Blizzard Challenge (satellite of Interspeech), Cupertino, USA (2016)
4. M. Charfuelan and I. Steiner: Expressive speech synthesis in maryTTS using audiobook data and emotionML. In Proc. of Interspeech, Lyon, France (2013)
5. H. Kanagawa, T. Nose, and T. Kobayashi: Speaker-independent style conversion for HMM-based expressive speech synthesis. In Proc. of ICASSP, Vancouver, Canada (2013)
6. 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 Proc. of Interspeech, San Fransisco, USA (2016)
7. P. Alain, J. Chevelu, D. Guennec, G. Lecorve, and D. Lolive: The IRISA text-to-speech system for the blizzard challenge 2016. In Proc. of the Blizzard Challenge (satellite of Interspeech), Cupertino, USA (2016)

8. I. Steiner, M. Schröder, M. Charfuelan, and A. Klepp: Symbolic vs. acoustics-based style control for expressive unit selection. In Proc. of ISCA Speech Synthesis Workshop (SSW7), Kyoto, Japan (2010)
9. S. Pammi and M. Charfuelan: HMM-based scost quality control for unit selection speech synthesis. In Proc. of ISCA Speech Synthesis Workshop (SSW8), Barcelona, Spain (2013)
10. O. Turk and M. Schröder. Evaluation of expressive speech synthesis with voice conversion and copy resynthesis techniques. In Audio, Speech and Language Processing (IEEE Trans.), vol. 18, no. 5, pp. 965–973 (2010)
11. L. Feugère, C. d'Alessandro, S. Delalez, L. Ardaillon, and A. Roebel: Evaluation of singing synthesis: methodology and case study with concatenative and performative systems. In Proc. of Interspeech, San Fransisco, USA (2016)
12. S. Brognaux, B. Picart, and T. Drugman: Speech synthesis in various communicative situations: Impact of pronunciation variations. In Proc. of Interspeech, Singapore (2014)
13. R. Qader, G. Lecorvé, D. Lolive, M. Tahon and P. Sébillot. Statistical Pronunciation Adaptation for Spontaneous Speech Synthesis. In Proc. of TSD, Pragua, Czech Republic (2017)
14. M. Tahon, R. Qader, G. Lecorvé, and D. Lolive. Improving TTS with corpus-specific pronunciation adaptation. In Proc. of Interspeech, San Fransisco, USA (2016)
15. M. Tahon, R. Qader, G. Lecorvé, and D. Lolive: Optimal feature set and minimal training size for pronunciation adaptation in TTS. In Proc. of SLSP, Pilzen, Czech Republic (2016)
16. F. Goldman-Eisler: The significance of changes in the rate of articulation. In Language and Speech, vol. 4(4), pp. 171–174 (1961)
17. D. Guennec and D. Lolive: Unit selection cost function exploration using an A* based text-to-speech system. In Proc. of TSD, Brno, Czech Republic (2014)
18. O. Boeffard, L. Charonnat, S. L. Maguer, D. Lolive, and G. Vidal. Towards fully automatic annotation of audiobooks for TTS. In Proc. of LREC, Istanbul, Turkey (2012)
19. K. Bartkova, D. Jouvet, and E. Delais-Roussarie: Prosodic parameters and prosodic structures of French emotional data. In Proc of Speech Prosody, Shanghai, China (2016)
20. T. Lavergne, O. Cappé, and F. Yvon: Practical very large scale CRFs. In Proc. of ACL, Uppsala, Sweden (2010)
21. G. Chollet and C. Montacié: Evaluating speech recognizers and databases. In Recent Adv. Speech Understand. Dialog Syst., NATO ASI F: Comput. Syst. Sci., vol. 46, pp. 345–348 (1988)