



# Comparison of chironomic stylization versus statistical modeling of prosody for expressive speech synthesis

Marc Evrard, Samuel Delalez, Christophe d'Alessandro, Albert Rilliard

LIMSI-CNRS, Audio & Acoustic Group  
 Rue John von Neumann, Campus Universitaire d'Orsay  
 Bât. 508, F-91405 ORSAY CEDEX, FRANCE  
 {evrard, delalez, cda, rilliard}@limsi.fr

## Abstract

Chironomic stylization is the process of real-time modification of intonation contours ( $f_0$  and tempo) using drawing/writing gestures with a stylus on a graphic tablet. The question addressed in this research is whether hand-made intonation stylization could improve or degrade expressivity and overall quality, compared to statistical modeling of prosody. A system for expressive TTS in French based on HMM was designed. A neutral corpus and six expressive speech corpora were used (*anger, fear, joy, sadness, sensuality, surprise*). Five sentences were synthesized with the six types of expressivity through CMLLR adaptation. Using a chironomic system, three trained subjects were asked to modify synthetic sentences, aiming at improving their expressive quality. Natural, HMM-TTS, and HMM-TTS-Chironomic sentences were evaluated in an expressivity recognition test and a MOS test. The results show that chironomic modification brings significant improvements in both recognition and MOS tests. These results are discussed in detail, together with the effects of voice quality on the perception of HMM-TTS expressive speech. The two main conclusions are: (i) intonation of HMM-TTS can be significantly improved; (ii) hand-corrected TTS improves expressivity and overall quality. Chironomic stylization is a powerful tool lying between fully automatic TTS and recorded speech.

**Index Terms:** Calliphony, chironomy, prosody, prosodic synthesis, expressive synthesis, adaptive training, HTS, HMM.

## 1. Introduction

This paper addresses the question of semi-automatic high quality expressive speech synthesis. In some situations (e.g., prompt preparation, audio books), automatic Text-To-Speech Synthesis (TTS) would benefit from some prosodic post-processing. A possible solution to this problem is to use chironomic stylization, i.e., real-time modification of intonation contours ( $f_0$  and tempo) using drawing/writing gestures with a stylus on a graphic tablet. Whether hand-made intonation stylization could improve or degrade expressivity and overall quality, compared to statistical modeling of prosody, is discussed here.

A first post-processing system was previously presented in the context of concatenative synthesis [1]. However, the modified speech was degraded, and voice quality aspects were not considered. In the present research, HMM-based synthesis and adaptive training techniques [2] were used for the generation of expressive speech. Adaptation allows for the training of a base model on few large corpora, and for the subsequent adaptation using smaller corpora (about 6 minutes of speech [3]) for other speakers or styles.

Adaptation — here applied to 6 small corpora of expressive speech — results in synthetic speech that exhibits clear vocal quality changes. In this experiment, a happy voice (*joy*) was produced with a formant shift that mimics the effect of smile [4], and *sadness* with a lower pitch and slower tempo [5]. However, expressive speech is produced partly via long-term change in the vocal tract settings [6], and partly via modifications of the dynamic strategy used by speakers — typically for their intonation settings [5]. Intonation, seen as phrase and sentence-long changes in pitch, requires large amount of data to be adequately captured and modeled: the size of the corpus appears to be an important parameter when suprasegmental features need to be modeled. Building multiple large training corpora is neither time nor cost-effective. A post-processing correction of the TTS output may provide a solution to this current limitation, by allowing a human operator to provide the contour adequate to the sentence being synthesized. It has been shown [7] that one can mimic natural intonation using a chironomic interface (i.e., a handwriting gesture driven interface allowing an operator to draw intended intonation and rhythmic patterns).

The questions investigated in this paper are twofold: (i) Could hand-made prosodic stylization improve expressivity compared to statistical model productions? (ii) Do the changes induced by this additional processing degrade the overall quality of the output? A chironomic interface, named *Calliphony*, was used in a post-production process on the output of an HMM-based speech synthesizer for French, to modify pitch and tempo of expressive speech sentences, in order to enhance their quality and prototypicality.

A description of the HMM-based speech synthesis platform is given in the next section. *Calliphony* is detailed in section 3. Section 4 describes the experimental use of chironomy for the expressive speech production. Section 5 presents the perceptual validation of the quality and prototypicality of chironomic performances, before the concluding discussion on the results.

## 2. The TTS platform

LIPS<sup>3</sup> (LIMSI Parametric Statistical Speech Synthesizer) was built around the HTS (HMM-based speech synthesis system) platform [8]. Text processing modules for French were locally developed and adapted to the natural language processing requirements of the HTS software package.

### 2.1. Natural language processing modules

The graphemes-to-phonemes (GP) conversion was developed on an already existing core set of rules (previously evaluated

in [9]). The linguistic features were extracted using an object oriented Python library developed *in situ*. Out of the standard 53 features proposed by the HTS working group [10] for the English Language, 19 were selected, based on the French language particularities, on the feature subset classification suggested in [11], and on our own experiments.

## 2.2. The synthesis software

The TTS synthesis platform used for this work made use of the STRAIGHT vocoder (STRAIGHT-v40-007-d) [12] and the Speech Signal Processing Toolkit (SPTK-3.8) [13] for extracting and synthesizing acoustic parameters. Statistical models of these parameters were trained using HTS (hts-2.3beta).

The scripts for style adaptation provided within HTS were modified to be applied to the case of expressive variations. The default parameters for 48 kHz signals' acoustic feature extraction and training were used, except for the fundamental frequency ( $f_0$ ) range of the STRAIGHT extractor, set to 120 Hz – 600 Hz, a Mel Generalized Cepstral (MGC) [14] order of 50 for SPTK, and a Minimum Description Length (MDL) tuning factor of 0.6 for the MGC stream (tree size control parameter for MDL) of HTS.

The *style*-adaptive training (SAT, *style* in place of *speaker*) was carried out using all corpora (neutral and expressive speech). Synthesis of each type of expressivity was conducted through the Constraint Maximum Likelihood Linear Regression (CMLLR) adaptation [2] procedure using the corresponding expressive speech corpus.

## 2.3. Corpus of expressive speech

The training corpus was based on a neutral set and 6 smaller sets with expressive variations. The speaker was a professional actress, L1 speaker of standard Parisian French. The recordings took place in a soundproof vocal booth, using a condenser microphone with an omnidirectional polar pattern, at a close distance. The recording was performed at 48 kHz, 24-bit and eventually converted to 48 kHz, 16-bit PCM files. The complete text corpus consisted of 1402 sentences, containing 10 313 words and 15 552 syllables. The corpus was phonetically labeled using the EHMM tool (Ergodic hidden Markov models) [15] from the Festvox tools suite, and manual corrections of phonetic label were applied.

The 1402 sentences composed the neutral set. The first 160 sentences of this set were used for the six expressive sets. The speaker was asked to perform the neutral style as a calm narrative expressivity. Six types of vocal expressivity, corresponding to six affective states were selected. The types of expressivity were chosen in order to be maximally different as far as their acoustic performances were concerned (high or low pitch mean and range, breathy or modal phonation, lip smile, etc.). Five French expressive labels (English translations are given here) were based on the GENEVA Multimodal Emotion Portrayals (GEMEP) corpus [16]: (hot) *anger*, *fear* (worry), (elated) *joy*, *sadness* (depression) and *surprise*. The expression of *sensuality* was added in order to complete the vocal inventory with a breathy phonation [17].

Five sentences were removed from each training corpora to serve for evaluation purposes, in order to have a natural performance as reference. Thus, 1397 sentences were kept for the main corpus as training set, and 155 sentences in each of the six expressive sets for the adaptation phase.

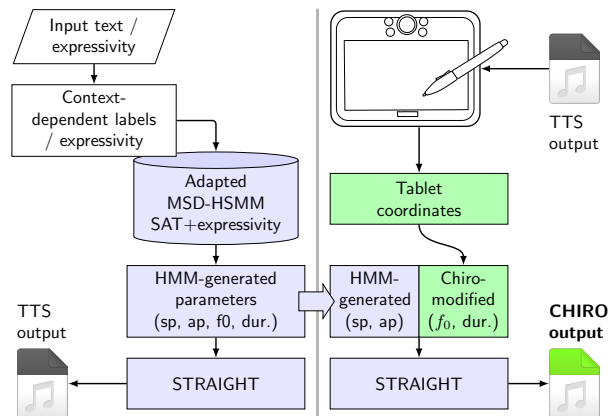


Figure 1: *Chironomic production of expressive speech process: Expressive speech synthesis (left), Calliphony operation (right).*

## 3. The Calliphony system

*Calliphony* is a system for computerized chironomic stylization of prosody [7]. It aims at modifying the fundamental frequency ( $f_0$ ) and tempo of speech utterances with the help of gestural control, in order to enhance the TTS output. The current system was built on a previously developed hand-gesture-controlled speech synthesis, which used a stylus on a graphic tablet to control pitch for musical purposes [18]. The use of a stylus allows one to reinvest the skills developed for handwriting. The interface used for this work was a WACOM® IntuosVR pen tablet. The  $(x, y)$  position and the pressure of the stylus on the tablet are sent in real-time to the computer. In this version of *Calliphony*, the  $x$ -axis of the tablet is mapped to the reading speed of the utterance, and the  $y$ -axis to a logarithmic scale of the  $f_0$ . This means that the current position of the stylus gives the actual absolute value of  $f_0$  on the  $y$ -axis, and the acceleration (or deceleration) rate on the  $x$ -axis. For instance, if the performers keep the stylus in the center of the tablet, the  $f_0$  would be constant at a fixed frequency, and the rhythm would be the same as in the input sentence. Other aspects of vocal quality and articulation are not affected by the process.

The  $f_0$  and tempo modifications obtained through the tablet interface are transferred onto the speech waveform in real-time, while the performer listens to the resulting sound. This is done using a real-time implementation of the Time-Domain Pitch Synchronous Overlap Add (TD-PSOLA [19]) algorithm [20]. The software was implemented in Java within the Max/MSP programming environment.

As the current implementation of STRAIGHT is not real-time, the system cannot achieve a real-time resynthesis at the output of the TTS system. Thus, as shown in Figure 1, the TTS output is currently recorded as a wave file, which is subsequently used in the *Calliphony* system, while the  $(x, y)$  positions of the stylus are recorded. Eventually, these  $(x, y)$  values are reused to modify the output parameters of the HMM-TTS system, and generate a high quality sound, enhanced by chironomic movements.

## 4. Chironomic production of expressive speech

The ability of the *Calliphony* interface to enhance the expressivity of a TTS system by modeling relevant contour of  $f_0$  and

tempo parameters was evaluated through a production task. A group of three subjects — the performers — were selected to enhance the expressive variations of speech samples through *Calliphony*. They were chosen based on their experience as musicians and users of the pen tablet as a musical instrument.

The production task consisted of manually modifying the expressivity of a set of recorded TTS sentences. At this step, the text of the five sentences extracted from the learning corpus (cf. section 2.3) was used. A set of audio files was first synthesized through the HMM models trained on the expressive speech corpora, using the CMLLR adaptation procedure (SAT+<expressivity>). Performers then tried to enhance the expressivity by changing the  $f_0$  and tempo contour. For example, they had to enhance the prototypicality of the expression of *fear* or *surprise* carried by a sentence already generated by the TTS system to express *fear* or *surprise* respectively. When satisfied with the resulting sentence, each performer selected its two best productions, for each of the five sentences and each of the six types of expressivity.

Among these performances, the best ones (one for each sentence and expressivity) were selected by the authors. These 30 chironomically-enhanced stimuli (the “CHIRO” stimuli) were then used during the listening tests described in the following section, together with the 30 wave files recorded at the output of the HMM-TTS system (the “TTS” stimuli), and the 30 sentences recorded by the speaker of the training corpus (the “NAT” stimuli).

## 5. Performance evaluations

Two perceptual experiments, a recognition test and a quality evaluation test, were set up to measure:

- The capability of listeners to recognize the 6 intended types of expressivity (i.e., *anger*, *fear*, *joy*, *sadness*, *sensuality*, and *surprise*).
- The perceived quality of the output speech signal.

21 subjects participated in both tests, starting with the recognition test. The same 5 sentences, as performed by the original speaker (NAT), the HMM-TTS (TTS) or the *Calliphony* system (CHIRO) in each of the 6 types of expressivity, were used for both tests.

### 5.1. Expressivity recognition

Expressive sentences were presented to listeners in random order. Subjects were instructed to judge which of the 6 types of expressivity they perceived, and to select it in a forced-choice list. Results were expressed as binary recognition scores, and stored in a contingency matrix. A logistic regression was used to analyze the relative influence of the following fixed factors: the targeted Expressivity (Expr: 6 levels), the Production system (Prod: 3 levels) and the Sentence (Sent: 5 levels). The individual influence of listeners was modeled as a random factor. The `lme4` library of R was used [21, 22]. Table 1 presents the results of the analysis.

All factors, and their interactions, have a significant effect on the recognition performance. Meanwhile, the different systems (i.e., the speaker (NAT), the HMM-TTS and the *Calliphony* (CHIRO) systems) that produced the stimuli explain most of the observed deviance, followed by the presented expressivity and the interaction between these two factors. The TTS-generated types of expressivity reach 58% of recognition, whereas the chironomic-enhanced version passes 75%, and the natural version 86%.

Table 1: Analysis of deviance applied on the output of the logistic regression run on the recognition results. Significance ( $p$ ) tested according to a Chi-square distribution ( $\chi^2$ ), depending on the degree of freedom ( $df$ ) of the factors (and their two-way interactions — see text for labels).

Factor	$\chi^2$	$df$	$p$
Expr	69.5	5	< 0.0001
Prod	164.8	2	< 0.0001
Sent	34.3	4	< 0.0001
Expr:Prod	111.2	10	< 0.0001
Expr:Sent	64.2	20	< 0.0001
Prod:Sent	19.2	8	< 0.05

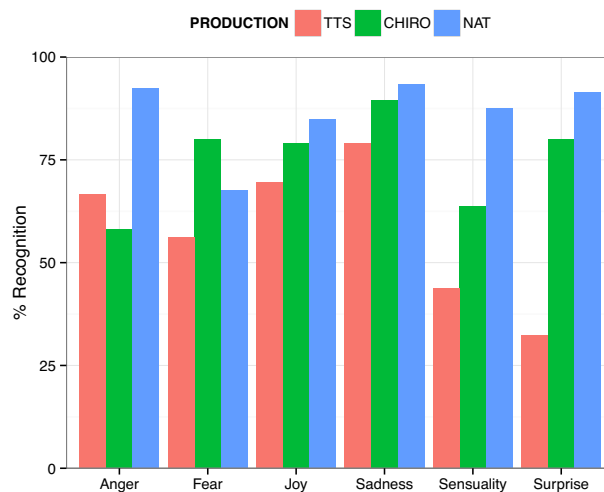


Figure 2: Percentage of recognition obtained for the 2-way interaction between the six targeted types of expressivity, and the three production systems.

The graph of Figure 2 shows that recognition scores also depend on the targeted type of expressivity. The chironomic performances mostly upgrade the recognition scores; meanwhile, in the case of *anger*, it did slightly degrade the recognition performances (by 9%). For the five other types of expressivity, the contour induced by gestures in the  $f_0$  and tempo parameters helps listeners to recognize the expression. Chironomic productions even surpass the recognition score of natural *fear*. The added value of chironomic augmentation of the TTS output is particularly noteworthy in the case of *surprise* (48%).

An analysis of the contingency matrix shows that listeners do few systematic recognition errors. A classification study shows the only substantial confusions concern the TTS-produced expressions. This can be explained by the acoustic proximity of stimuli, notably in terms of mean pitch level. For the expression of *sensuality* (performed with a low pitch), subjects tend to mix it with *sadness* — also performed with a low pitch. One may also note that the breathiness of the natural sensual sentences is not adequately reproduced by the synthesis system. *Fear* and *surprise*, which both exhibit a high pitch, were respectively mixed with *anger* and *fear* — but not with *joy*, which also exhibits a high pitch, certainly because of the characteristic voice quality of smile, which is well captured by the synthesis system.

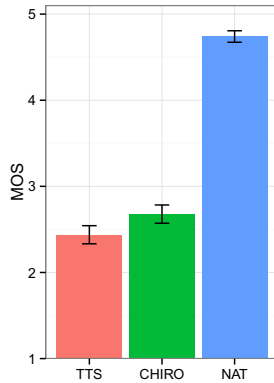


Figure 3: Mean MOS scores for the 2-way interaction between the targeted expressivity and the production systems.

## 5.2. Quality evaluation

Acoustic processing often degrades the perceived quality of speech. In order to evaluate the effect of a chironomic processing applied to the output of the TTS system on overall quality, subjects were asked to rate the perceived overall quality of the same 5 sentences, with 6 types of expressivity, on a 1 to 5 MOS scale (mean opinion score, 5 being the best score). Results were analyzed through an analysis of variance, with the targeted Expressivity (Expr: 6 levels), the Production system (Prod: 3 levels) and the Sentence (Sent: 5 levels) as fixed factors. Results are presented in Table 2.

Table 2: Analysis of the variance explained by each factor (see text for labels) and their two-way interactions on the MOS score. Results report the  $F$ -statistics for the factor's and the error's degrees of freedom ( $df$ ), the associated  $p$  level and the effect size ( $\eta_p^2$ ).

Class	$df$	$df$ error	$F$	$p$	$\eta_p^2$
Expr	5	1840	36.3	< 0.001	0.09
Prod	2	1840	1405.4	< 0.001	0.60
Sent	4	1840	7.8	< 0.001	0.02
Expr:Prod	10	1840	7.9	< 0.001	0.04
Expr:Sent	20	1840	1.6	< 0.05	0.02
Prod:Sent	8	1840	6.4	< 0.001	0.03

All factors have a significant effect on the MOS scores. Meanwhile, the type of production has a major explicative power ( $\eta_p^2 = 0.60$ ), followed by the targeted expressivity. The effect of the production factor is depicted in Figure 3. It shows increasing scores from TTS and *Calliphony* (CHIRO) to the natural performances (NAT). Each increment was tested as significant under a post-hoc Tukey test. The interaction between expressivity type and production type shows that this qualitative increment of chironomic augmentation is always observed. Typically, the synthetic expressions performed with a high  $f_0$  (*anger*, *fear*, *joy*, *surprise*) were rated with a lower MOS score than the two with a low  $f_0$ .

## 6. Discussion & conclusions

The main conclusion of this study of TTS enhancement through chironomy is that hand gestures can produce changes in the con-

tour of  $f_0$  and tempo, which improve both expressivity recognition and global quality. Thus, chironomy seems to be a promising way to improve naturalness, quality and expressivity of speech synthesis. The detailed results also shed light on some specific strengths and weaknesses of both the presented TTS system and the *Calliphony* system that are discussed here.

The greatest changes in recognition rate (cf. Figure 2) are observed for the TTS performances, with *surprise* and *sensuality* below 50%, while *sadness* exceeds 75%. These low scores could be explained by (i) the poor performance of the TTS system in capturing the characteristic breathiness of the sensual voice, and (ii) the poor modeling of intonation for *surprise*. Indeed, for these two categories (*sensuality* and *surprise*), the relative contribution of chironomy is contrasted. The recognition gain obtained for *surprise* is major, indicating a primary role of an adequate prosodic contour over the sentence for the performance of *surprise*. On the contrary, the chironomic gain for *sensuality* is modest, showing that the prosody contour may be of less perceptual pertinence, in this case, than an adequate voice quality. This relative importance of the prosody contour vs. the mean voice quality is also observed with *anger*, whose recognition is degraded by chironomy; on the contrary, the performance of *fear* produced with *Calliphony* is even greater than the natural ones.

The use of the *Calliphony* tool paired with a parametric speech synthesizer is also promising for research purposes, in the fields of prosody and expressive voice analysis — typically for research on affective expressions. Being able to disentangle the relative role of dynamic vs. static prosodic cues has been a long-standing debate for the understanding of affective speech (e.g., [23, 5, 24, 6, 25]) that has often been limited by the complexity of separately modifying the various components of prosody — and notably of voice quality (meanwhile, cf. [26, 27]). One can easily map other parameters of the TTS system onto the tablet, in addition, or in replacement of the current ( $f_0$ , *speed*) pair. For example, the bad result observed for *anger* would certainly benefit from a mapping of the parameters allowing the performer to manipulate the voice strength [28].

Regarding the general quality results, the MOS test summarized in Figure 3 shows that the supplementary manipulations introduced by the *Calliphony* interface don't degrade quality. On the contrary, the perceived quality is increased. This could be explained by two elements. On the one hand, the manipulation was applied to the parametric vector used at the vocoder input, and the sentence was subsequently synthesized so that no loss of quality occurs during processing. On the other hand, the original  $f_0$  produced by the HMM-TTS shows important and unnatural microprosodic changes, while the continuous movements of the hand provide smoother  $f_0$  variations, which may be perceived as more natural.

To conclude, the hand-controlled interface improves expressivity and overall quality of the expressive speech HMM-TTS. Chironomic stylization is a versatile tool that could be used for research purposes, such as to experiment with prosody modifications without loss of quality.

Future developments will focus on interfaces to control rhythm, to replace continuous speed control by a discrete, syllable-based control.

## 7. Acknowledgments

This work was supported by the ADN T-R Project (FUI-11 OSEO/DGCIS) granted by *la région Île-de-France*, *le conseil général de la Seine-Saint-Denis*, and *la ville de Paris*.

## 8. References

- [1] S. Le Beux, A. Riilliard, and C. d'Alessandro, "Calliphony: a real-time intonation controller for expressive speech synthesis," in *SSW*, 2007, pp. 345–350.
- [2] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings (ICASSP)*, vol. 2, 2001, pp. 805–808.
- [3] J. Yamagishi, H. Zen, T. Toda, and K. Tokuda, "Speaker-Independent HMM-based Speech Synthesis System: HTS-2007 System for the Blizzard Challenge 2007," in *Blizzard Challenge Workshop, Bonn, Germany*, 2007.
- [4] V. C. Tarter and D. Braun, "Hearing smiles and frowns in normal and whisper registers," *The Journal of the Acoustical Society of America*, vol. 96, no. 4, pp. 2101–2107, 1994.
- [5] T. Bänziger and K. R. Scherer, "The role of intonation in emotional expressions," *Speech communication*, vol. 46, no. 3, pp. 252–267, 2005.
- [6] M. Goudbeek and K. Scherer, "Beyond arousal: Valence and potency/control cues in the vocal expression of emotion," *The Journal of the Acoustical Society of America*, vol. 128, no. 3, pp. 1322–1336, 2010.
- [7] C. d'Alessandro, A. Riilliard, and S. Le Beux, "Chironomic stylization of intonation," *The Journal of the Acoustical Society of America*, vol. 129, no. 3, pp. 1594–1604, 2011.
- [8] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.
- [9] F. Yvon, P. B. De Mareüil, V. Aubergé, M. Bagein, G. Bailly, F. Béchet, S. Foukia, J.-F. Goldman, E. Keller, V. Pagel *et al.*, "Objective evaluation of grapheme to phoneme conversion for text-to-speech synthesis in French," *Computer Speech & Language*, vol. 12, no. 4, pp. 393–410, 1998.
- [10] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in *Proceedings of 2002 IEEE Workshop on Speech Synthesis*, 2002, pp. 227–230.
- [11] S. Le Maguer, N. Barbot, and O. Boeffard, "Evaluation of contextual descriptors for HMM-based speech synthesis in French," in *8th International Speech Communication Association (ISCA) Speech Synthesis Workshop*, 2013, pp. 153–158.
- [12] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.
- [13] K. Tokuda, K. Oura, A. Tamamori, S. Sako, H. Zen, T. Nose, T. Takahashi, J. Yamagishi, and Y. Nankaku, "Speech signal processing toolkit (SPTK)," *Online*, recent version, 2012.
- [14] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis – a unified approach to speech spectral estimation," in *ICSLP*, 1994.
- [15] K. Prahallad, A. W. Black, and R. Mosur, "Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis," in *Proceedings of 2006 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 2006.
- [16] T. Bänziger, M. Mortillaro, and K. R. Scherer, "Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception," *Emotion*, vol. 12, no. 5, p. 1161, 2012.
- [17] P. R. Léon, *Précis de phonostylistique: parole et expressivité*. Nathan, 1993.
- [18] C. d'Alessandro, L. Feugere, S. Le Beux, O. Perrotin, and A. Riilliard, "Drawing melodies: Evaluation of chironomic singing synthesis," *The Journal of the Acoustical Society of America*, vol. 135, no. 6, pp. 3601–3612, 2014.
- [19] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using di-phones," *Speech communication*, vol. 9, no. 5, pp. 453–467, 1990.
- [20] S. Le Beux, B. Doval, and C. d'Alessandro, "Issues and solutions related to real-time TD-PSOLA implementation," in *Audio Engineering Society Convention 128*, 2010.
- [21] D. Bates, M. Maechler, B. Bolker, and S. Walker, "lme4: Linear mixed-effects models using Eigen and S4," *ArXiv e-print; submitted to Journal of Statistical Software*, 2014.
- [22] R. H. Baayen, *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press, 2008.
- [23] D. R. Ladd, K. E. Silverman, F. Tolkmitt, G. Bergmann, and K. R. Scherer, "Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect," *The Journal of the Acoustical Society of America*, vol. 78, no. 2, pp. 435–444, 1985.
- [24] Y. Greenberg, M. Tsuzaki, H. Kato, and Y. Sagisaka, "A trial of communicative prosody generation based on control characteristic of one word utterance observed in real conversational speech," in *Proc. Speech Prosody*, 2006, pp. 37–40.
- [25] J. A. de Moraes and A. Riilliard, "Illocution, attitudes and prosody," *Spoken Corpora and Linguistic Studies*, vol. 61, p. 233, 2014.
- [26] C. Gobl, A. Ní Chasaide *et al.*, "The role of voice quality in communicating emotion, mood and attitude," *Speech communication*, vol. 40, no. 1, pp. 189–212, 2003.
- [27] N. Audibert, D. Vincent, V. Aubergé, and O. Rosec, "Expressive speech synthesis: Evaluation of a voice quality centered coder on the different acoustic dimensions," in *Proc. Speech Prosody*, 2006.
- [28] J.-S. Liénard and C. Barras, "Fine-grain voice strength estimation from vowel spectral cues," in *INTERSPEECH*, 2013, pp. 128–132.