INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Evaluation of delexicalization methods for research on emotional speech

Nicolas Audibert[1], *Francesca Carbone*[2,3], *Maud Champagne-Lavau*[2], *Aurélien Said Housseini*[1], *Caterina Petrone*[2]

[1]Laboratoire de Phonétique et Phonologie, UMR7018 CNRS/Sorbonne Nouvelle, Paris, France
[2]LPL, CNRS, Aix-Marseille University, Aix-en-Provence, France
[3]School of Psychology, University of Kent, Canterbury, United Kingdom

nicolas.audibert@sorbonne-nouvelle.fr, f.carbone@kent.ac.uk, maud.champagne-lavau@univ-amu.fr, aurelien.said-housseini@sorbonne-nouvelle.fr, caterina.petrone@univ-amu.fr

Abstract

Perceptual evaluation of non-controlled emotional speech requires delexicalization to neutralize semantic variation. However, most existing methods imply losing spectral cues crucial to emotional attribution, related to both laryngeal and supralaryngeal settings. We propose a method relying on voice morphing to retain part of the spectral information of the original stimuli, as an additional step to diphone synthesis delexicalization. After previous assessment of intelligibility loss, this study evaluates the naturalness of angry and neutral expressions in French films, delexicalized using low-pass filtering and the proposed method implemented with MBROLA and STRAIGHT. Results show that morphing-based delexicalization, which leads to accurate emotional attribution, is rated with a higher degree of naturalness than low-pass filtering. Implications for research in affective speech are discussed with regards to other delexicalization methods proposed in the literature.

Index Terms: delexicalization, emotion, prosody, voice quality, perception

1. Introduction

When perceiving emotional speech, listeners interpret prosodic information, voice quality and other spectral cues on the one hand, and semantic information on the other. Both channels contribute to the identification of the emotions expressed by the speaker in a mutually reinforcing way (see for example [1]).

Research on vocal expressions of emotions and other affects focuses mainly on prosodic variations as well as on voice quality cues and spectral changes that allow the listener to infer the speaker's affective state. Much of this research has used systematic variation by keeping verbal content constant in order to neutralize the influence of semantic content. Furthermore, emotional expressions are often elicited by actors in experimental settings [2, 3].

However, the naturalness of emotional expressions simulated in such settings have been questioned [4, 5], leading a growing number of researchers to opt for expressive speech extracts produced without control of the semantic content. Other authors focus on expressions of affects produced in contexts incompatible with a standardized semantic content (see e.g. [6] on the perception of charisma expressed by business speakers). In order to evaluate the affective information carried by prosodic and spectral variations, it is then necessary to abstract out the lexical content by presenting delexicalized stimuli to the listeners.

Various approaches have been proposed in the literature for the delexicalization of speech extracts, with the aim of isolating the contribution of prosody.

A classical method consists of modifying the spectral properties of the signal while preserving most of the suprasegmental variations via the spectral inversion technique [7], optionally combined with an additional filtering step [8]. The random splicing method [9] in which short extracts are shuffled in random order can also be mentioned.

The most commonly used method is to apply low-pass filtering to eliminate the segmental information present in the mid and high frequencies. The low-pass filtering allows retaining only the fundamental frequency (f0) modulations and part of the intensity modulations. In the vast majority of cases, a fixed cut-off frequency has been chosen, e.g. 400 Hz for [10] and [11] (the latter also retaining frequencies above 5 kHz) or 600 Hz for [6]. In some studies, filtering has been applied with a cut-off frequency dependent on the f0 register in delexicalized samples. For instance [12] opted for bandpass filtering designed to include the minimum value of f0 and the maximum value of its first harmonic. This choice is consistent with the recommendations of [13] who advocate a cut-off frequency one octave above f0.

[14] have evaluated several methods, including a time-dependent low-pass filtering with a cut-off frequency set just above f0 and inverse filtering, together with other candidate methods designed to retain only segmental information. All proposed methods achieved comparable performance for the identification of prosodic functions. In a preference rating task, listeners rejected spectral inversion and showed a slight preference for the modulation of a sinusoidal wave including the first two harmonics. This version of the minimal coding of suprasegmental information was therefore retained in the PURR delexicalization algorithm proposed by [14]. The PURR algorithm has been used by various authors to evaluate the role of prosody in the perception of foreign accent [15], emotion [16] or speaking styles [17]. Similarly, other studies have used delexicalization methods based on the modulation of f0 on pure tones [18], or on synthetic 'hum' stimuli [19] as implemented in Praat [20].

Although these methods can effectively reproduce f0 modulations, the resulting samples are highly unnatural. For this reason, [21] proposed a method based on diphone analysis/resynthesis combined with phone substitution and prosodic copy using the TD-PSOLA algorithm [22]. This method was then used by various authors, notably [23] for the study of language identification from suprasegmental cues. In order to produce stimuli more similar to the originals, [24]

10.21437/Interspeech.2023-1903

proposed to replace diphone synthesis by a delexicalized imitation produced by the same speaker, combined with prosodic copy to keep controlled f0 modulations. The main limitation of this analysis/resynthesis approach is that the voice quality and other spectral cues present in delexicalized stimuli are those of read speech, unless a diphone database (or imitation) produced by the target speaker under the same conditions as the utterance to be synthesized is available. This can be problematic for the study of expressive speech, especially emotional expressions for which voice quality cues can be particularly salient (e.g. [25]). To our best knowledge, the only method allowing the restitution of voice quality and its variations is based on the estimation of the glottal flow by inverse filtering and on a linear modeling of the vocal tract [26].

As an alternative to inverse filtering methods that can be error-prone and induce a loss of auditory quality, [27] replaced original vowels with a slice of a vowel produced by the same speaker in a comparable context and averaged the spectrum for the unvoiced segments. Stimuli generated using this 'surrogate vowels' method were found to be more natural than the baseline low-pass filtering. A potential limitation of the 'surrogate vowels' method proposed by [27] is that it implicitly assumes that the spectral characteristics of a speaker, a speech style or an affect can be considered constant within the same utterance.

For our part, we propose a simple method based on a first step of analysis/resynthesis with phone substitution, completed by the use of voice morphing to partially restore the variations over time of the spectral characteristics of the original. With a view to applying this method to the study of expressive speech, following a study of loss of intelligibility and emotional attribution, we evaluate the degree of naturalness of delexicalized neutral expressions and expressions of hot anger.

2. METHODS

2.1. Corpus

We used 54 auditory stimuli presented in 3 different versions detailed below: (1) as natural stimuli, i.e., in their original version; (2) as synthesized stimuli, via the speech synthesizer MBROLA [28], and combined with the vocal morphing system STRAIGHT [29]; and (3) as low-pass filtered stimuli. Fig. 1 presents the waveform and spectrogram of the three stimuli corresponding to an angry utterance.

2.1.1. Natural stimuli

We extracted 73 natural stimuli from French movies, with the help of a student in fine arts. The auditory stimuli were converted into .wav files (sampling rate: 48 kHz as in the original DVDs). The natural stimuli were produced by nine French actors (four women, five men) either with hot angry (nine stimuli) or neutral (nine stimuli) expressions. These stimuli were 6.54 syllables long on average (mean duration = 0.97 sec, SD = 0.34 sec). They could also contain words carrying negative or positive valence (e.g., *Je suis pas malade moi!*, 'I am not crazy!'). The natural stimuli were of good recording quality and contained very little to no background noise. None of them contained overlapping speech from other actors.

2.1.2. Synthesized stimuli: 'saltanaj_morphing' condition

We created a synthesized version for each natural stimulus, in which broad phonotactics, rhythm and intonation contour were preserved from the original sentences. Such delexicalized stimuli were further enriched by partially reconstructing the global spectral characteristics of the original stimuli. Delexicalization was made by means of MBROLA [28] which performs a resynthesis through the concatenation of diphones, using a database of French diphones selected according to the actor's gender. We adopted the condition 'saltanaj' [23] in which each phone in the original stimulus is substituted by a phone of the same broad phonological category (e.g., /s/ for fricatives and vowels with /a/). Phones duration were obtained from a hand-corrected segmentation of original stimuli.

The spectral characteristics of the original stimuli were partially reconstructed by using a vocal morphing technique through STRAIGHT [29] in Matlab [30]. To do so, we used the original stimuli down-sampled at 16 kHz as source and the corresponding delexicalized stimulus synthesized via MBROLA as target. Since the segmental durations of the original stimuli are preserved in the synthesized version, a Short-Term Fourier Transform procedure was used on 25ms overlapping frames of both source and target signal, the time-aligned morphing process with STRAIGHT being applied to each pair of frames. After initial tests with various rates, the morphing rate was set at 0.5 as a trade-off between a lower rate that would have discarded the largest part of voice quality information and a higher rate that would have also rebuilt the segmental information of original stimuli. At the end of the process, we reconstructed the fundamental frequency contours of the original stimuli using TD-PSOLA [22], and the intensity via the Vocal Toolkit in Praat [31]. We refer to this condition hereafter as 'saltanaj_morphing'.

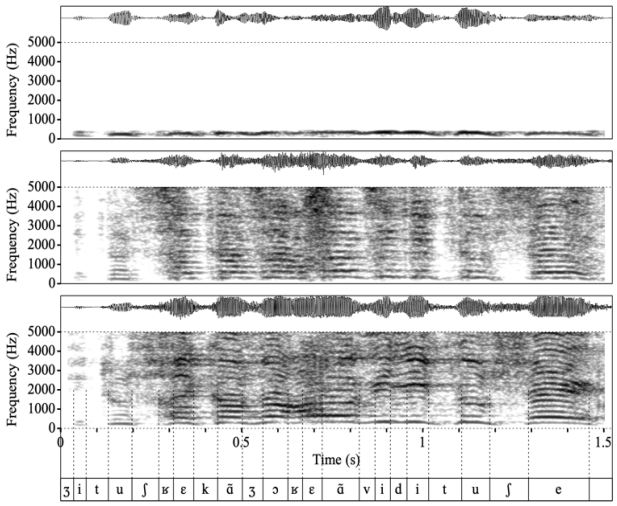

Figure 1: *Waveform and wide-band spectrogram of the three versions of the angry utterance 'J'y toucherai quand j'aurai envie d'y toucher !' (I'll touch it when I want to touch it!) produced by a male French actor in a film. Bottom: original stimulus; middle: stimulus delexicalized in condition 'saltanaj_morphing'; top: stimulus delexicalized in condition 'low-pass'.*

After their inspection to ensure the absence of obvious artifacts, the synthesized stimuli were validated through two tasks. An intelligibility task ensured that the addition of spectral information to the base 'saltanaj' condition did not lead to word recognition. Forty-seven native French listeners (38 F and 9 M, mean age = 24.7 y.o.) transcribed the auditory stimuli

orthographically. The intelligibility rate was computed for each provided transcription as the proportion of words matching the original stimulus. From this task, we retained only those stimuli which had a mean intelligibility rate equal or inferior to 30% (mean intelligibility rate in selection = 10%, SD = 9%). Furthermore, an identification task assessed whether the target emotional expressions (anger and neutral) were accurately recognized through prosody. Thirty-nine native French listeners who did not participate in the intelligibility task (21 F and 18 M, mean age = 30.3 y.o.) identified the emotional expressions by choosing between 'anger' and 'neutral'. From this task, we retained only those synthesized stimuli whose identification rate was equal or superior to 70%. Based on the intelligibility and identification tasks, eighteen stimuli were selected as they satisfied both selection criteria.

To ensure that the vocal morphing efficiently preserved emotional information in addition to the sole diphone-based resynthesis with phone substitution, the same identification task was performed without using STRAIGHT for vocal morphing. In this 'saltanaj' condition, the fundamental frequency contour and intensity of the original stimuli were reconstructed using the same method as for the 'saltanaj_morphing' condition. Twenty native French listeners (13 F and 7 M; mean age = 22 y.o.) participated in this complementary emotional identification task. Comparison of identification rates between tasks on the eighteen selected stimuli showed that listeners performed significantly better in the 'saltanaj_morphing' condition than in the 'saltanaj' condition (p=.003, mean rate 93% vs. 77%).

2.1.3. Filtered stimuli: 'low-pass' condition

Low-pass filtering of original stimuli was performed by means of a custom script written in Praat [20], using a Hann filter with a smoothing of 50 Hz. Since fundamental frequency is likely to reach high values in expressions of anger with high activation, the cutoff frequency has to be chosen carefully. In order to preserve f0 variations without including segmental information, the cutoff frequency was therefore set to two semitones over the maximum f0 in the stimulus. As a result, the applied cutoff frequency was between 136 Hz and 678 Hz (mean 373 Hz, SD 155 Hz) for the eighteen selected stimuli. We refer to this condition hereafter as 'low-pass'.

An intelligibility task was set-up following the same principles as in the 'saltanaj_morphing' condition. Twelve native French listeners (7 F and 5 M mean age = 21.9 y.o.) transcribed the low-pass filtered stimuli orthographically. For the eighteen selected stimuli, listeners could not provide a transcription in 74% of cases on average (SD: 19%). The analysis of proposed transcriptions indicated that none of them was related to the original stimulus, thus an intelligibility rate of 0%.

2.2. Perceptual evaluation of naturalness

A Qualtrics online survey collected responses from 39 French native speakers (34 F, 5 M; mean age = 21.7, SD = 5.74). Participants were asked to wear headphones or earbuds and to sit in a quiet room with no background noise. Prior to the task, participants responded to a short demographic questionnaire including questions about their language background and language use, age, educational level and current occupation. Participants were told that they were going to listen to angry or neutral stimuli which were less or more modified as for their words. They had to rate the naturalness of the emotional expressions, considering only the speaker's 'tone of voice' but disregarding the verbal level. We assume that collected ratings correspond to the similarity with expressions of the same emotion produced by human speakers. Each auditory stimulus was rated on a 5-point Likert scale ranging from *"not at all natural"* (corresponding to "1") to *"absolutely natural"* (corresponding to "5"). Angry and neutral stimuli were presented in two separate blocks, with the order of presentation being flipped for half of participants to prevent bias from block order. Stimuli in condition 'original', 'saltanaj_morphing' and 'low-pass' were randomly presented within each block.

2.3. Acoustic analysis

The overall acoustic difference between the original stimuli and their resynthesized or filtered versions was estimated by Euclidean distances in the 12-dimension space of MFCC coefficients. Using a custom Praat script, 13 coefficients were extracted on 15ms frames of speech signal with 5ms overlap (after excluding the initial and final silent part). Coefficient 0 related to the overall energy in the signal was dropped. The distance between the original and modified versions was then calculated on each frame.

3. Results

3.1. Ratings of naturalness

All participants were included in the data analysis as they completed 100% of the survey.

Fig. 2 shows the mean score for participants' judgments along the scale of naturalness. For both angry and neutral expression, the mean scores increase from stimuli with low-pass filtering (anger = 2.13; neutral = 1.95) to stimuli with both saltanaj and morphing manipulation (anger = 2.77; neutral = 2.87). As expected, original stimuli have the highest naturalness score (anger = 3.99; neutral = 4.09).

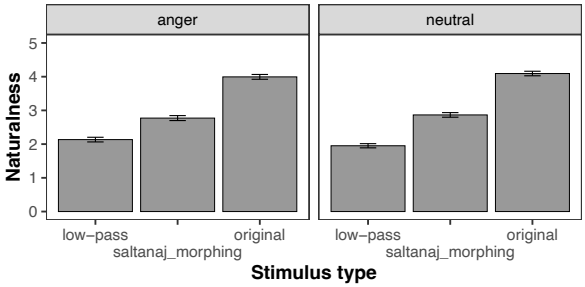

Figure 2: *Means and standard error for perceived naturalness, split by stimulus type and emotion.*

A linear mixed effects model fitted with R package lme4 [32] tested the effects of stimulus type (low pass/saltanaj_morphing/original) and emotional expression (angry/neutral) on the ratings of naturalness. Participants, actors and items within actors were included as random intercepts. The statistical analysis was based on 2016 observations (9 actors X 2 emotional expressions X 3 stimulus types X 39 participants).

We started the statistical analysis by fitting a model with a maximal random slope structure. Because of convergence issues, we reduced over-parametrization of the random structure by deleting interactions and random slopes with very

little variance. The simplification of the random structure of the model did not change the interpretation of the results. Significant interactions between fixed factors were analyzed using post-hoc tests via the function emmeans() of the emmeans R package [33], with Tukey correction for multiple comparisons [34]. The final model was:

lmer(ratings ~ stimulus_type * emotion + (1+stimulus_type + emotion|listeners) + (1|actors) + (1|actors:trial)

Results confirmed that the saltanaj_morphing stimuli received higher ratings of naturalness than low-pass stimuli, for both angry ($t=4.53$, $p<.001$) and neutral ($t=6.49$, $p<.001$) emotional states. On the other hand, the saltanaj_morphing stimuli were judged less natural than the original stimuli, for both angry ($t=-6.62$, $p<.001$) and neutral ($t=-6.65$, $p<.001$) emotional expression. Furthermore, for each stimulus type, naturalness' ratings were irrespective of emotional expression, i.e., the ratings were similar in angry and neutral stimuli ($p > .05$).

3.2. Acoustic distances to original stimuli

Fig. 3 shows the mean distance to the corresponding original stimulus in the MFCC space for both versions of delexicalized stimuli and both emotional categories, averaged for each corresponding original stimulus X stimulus type.

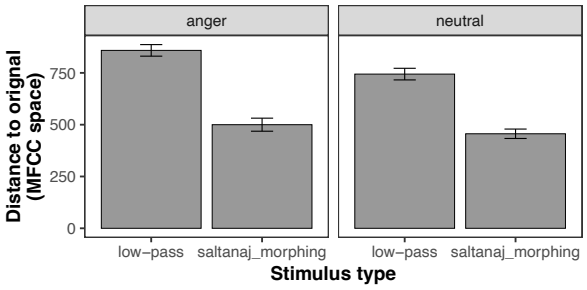

Figure 3: *Stimulus-wise means and standard error for acoustic distance to the corresponding original stimulus computed on 15ms frames in the 12-D space of MFCC, split by stimulus type and emotion.*

In order to evaluate the effect of the stimulus type, of the emotion and of the interaction between both factors, a linear mixed effects model was fitted with the following structure:

lmer(distToOriginalMFCC ~ stimulus_type * emotion + (1+stimulus_type+emotion|actor) + (1|frame_time)

As expected, results indicate a strong effect of the stimulus type ($t=35.68$, $p<.001$), the acoustic distance to original stimuli being significantly lower in 'saltanaj_morphing' condition than in 'low-pass' condition. A smaller effect of emotion is also found ($t=-4.8$, $p=.002$), as well as a strong effect of the interaction ($t=17.52$, $p<.001$).

Pairwise comparisons with Tukey correction show no difference in the acoustic distance to the original stimuli between neutral and angry stimuli in 'saltanaj_morphing' condition ($t=1.41$, $p=.528$). In the 'low-pass' condition, this distance is significantly higher in angry stimuli ($t=4.71$, $p=.007$), which may be explained by the higher f0 in angry expressions.

4. Discussion

One advantage of the proposed method based on 'saltanaj' phones substitution and vocal morphing over classical delexicalization methods (spectral inversion, filtering and other methods of minimal coding of suprasegmental information) is that it combines effective delexicalization with preservation of affective information. It also generates synthetic stimuli closer to the originals in terms of naturalness. In this respect, it is a promising method for the perceptual evaluation of emotions and other affects. Indeed, the higher degree of naturalness obtained minimizes the probability that delexicalized extracts are processed as non-speech stimuli by listeners (see e.g. [35] on brain processing of speech vs. non-speech sounds). Its application may therefore be extended to studies on perception of linguistic prosody in interaction with spectral variation.

The proposed 'saltanaj-morphing' delexicalization method can be complementary to the inverse filtering method proposed by [26], which accounts for voice quality using a model of the glottal flow wave in the original stimuli. This method does not require a neutral reference and is therefore also applicable to extracts for which direct comparison with other productions of the same speaker is not possible. Moreover, it allows the partial restitution of spectral changes not directly related to the glottal flow waveform but considered by some authors in a broader definition of voice quality [36].

With our proposed method, the degree of naturalness could be improved by fully exploiting the potential of a speech morphing tool such as STRAIGHT, which is not the case with this first version. In this study, we have chosen a process that can be fully automated provided that a phone alignment of original stimuli is available. This induces some audible spectral distortions in the synthesized stimuli, which could be reduced by manually fine tuning the temporal anchor points. When available, using neutral productions of the same speaker in the resynthesis process may improve the quality of the generated delexicalized stimuli. This can be achieved with a speaker-specific diphone database or as a replacement for it [24], or by using this neutral reference as target in a first step of morphing modification.

On the other hand, an obvious limitation is that, depending on the level of morphing applied, some segmental information may remain, for which reason we have chosen to apply only partial morphing. It is therefore recommended to check that the delexicalization process actually leads to a loss of intelligibility.

5. Conclusions

The delexicalization method we propose, based on the use of speech morphing in addition to phone substitution resynthesis, makes it possible to obtain delexicalized stimuli rated as more natural than those obtained by pass-down filtering and in which the identification of the emotional expression is preserved. We believe that this method, which is quite simple to implement, may prove useful for research on the perception of emotional speech and other expressions of affect.

6. Acknowledgements

This study was supported by an A*MIDEX (Aix-Marseille University) grant to Caterina Petrone and Maud Champagne-Lavau. It has been partially supported by the "Investissements d'Avenir" program ANR-10-LABX-0083 (Labex EFL).

7. References

[1] B. M. Ben-David, N. Multani, V. Shakuf, F. Rudzicz, and P. H. van Lieshout, "Prosody and semantics are separate but not separable channels in the perception of emotional speech: Test for rating of emotions in speech," *Journal of Speech, Language, and Hearing Research*, vol. 59, no. 1, pp. 72–89, 2016.

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

[3] P. N. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: Different channels, same code?," *Psychological Bulletin*, vol. 129, no. 5, pp. 770–814, 2003, doi: 10.1037/0033-2909.129.5.770.

[4] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "Desperately seeking emotions or: Actors, wizards, and human beings," in *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[5] J. Wilting, E. Krahmer, and M. Swerts, "Real vs. acted emotional speech.," in *Interspeech*, 2006, vol. 2006, p. 9th.

[6] O. Niebuhr, A. Brem, J. Michalsky, and J. Neitsch, "What makes business speakers sound charismatic?," *Cadernos de Linguística*, vol. 1, no. 1, pp. 01–40, 2020.

[7] I. Lehiste and W. S. Wang, "Perception of sentence boundaries with and without semantic information," *Phonologica*, pp. 277–83, 1976.

[8] J. Kreiman, "Perception of sentence and paragraph boundaries in natural conversation," *Journal of Phonetics*, vol. 10, no. 2, pp. 163–175, 1982.

[9] K. R. Scherer, "Randomized splicing: A note on a simple technique for masking speech content.," *Journal of Experimental Research in Personality*, 1971.

[10] H. S. Magen, "The perception of foreign-accented speech," *Journal of phonetics*, vol. 26, no. 4, pp. 381–400, 1998.

[11] J.-P. Goldman, T. Pršir, G. Christodoulides, A. C. Simon, and A. Auchlin, "Phonogenre identification: a perceptual experiment with 8 delexicalised speaking styles," *Cahiers de linguistique française*, vol. 31, pp. 51–62, 2014.

[12] N. Obin, P. Lanchantin, A. Lacheret, and X. Rodet, "Reformulating prosodic break model into segmental HMMs and information fusion," in *Proc. Interspeech 2011*, 2011, pp. 1829–1832. doi: 10.21437/Interspeech.2011-40.

[13] J. K. MacCallum, A. E. Olszewski, Y. Zhang, and J. J. Jiang, "Effects of low-pass filtering on acoustic analysis of voice," *Journal of Voice*, vol. 25, no. 1, pp. 15–20, 2011.

[14] G. P. Sonntag and T. Portele, "PURR—a method for prosody evaluation and investigation," *Computer Speech & Language*, vol. 12, no. 4, pp. 437–451, 1998.

[15] L. Rognoni, "The impact of prosody in foreign accent detection. A perception study of Italian accent in English," *Methodological perspectives on second language prosody. Papers from ML2P*, pp. 89–93, 2012.

[16] D. O. Peres, "Perception of emotional speech in Brazilian Portuguese: an intonational and multidimensional approach," *Nouveaux cahiers de linguistique française*, vol. 31, pp. 153–196, 2014.

[17] P. A. Barbosa, S. Madureira, and P. Boula de Mareüil, "Cross-Linguistic Distinctions Between Professional and Non-Professional Speaking Styles.," in *Interspeech*, 2017, pp. 3921–3925.

[18] D.-H. Kim-Dufor, E. Ferragne, O. Dufor, C. Astésano, and J.-L. Nespoulous, "Perception and comprehension of linguistic and affective prosody in children with Landau-Kleffner syndrome," in *Speech Prosody 2010-Fifth International Conference*, 2010.

[19] P. Boula de Mareüil, A. Rilliard, I. Lehka-Lemarchand, P. Mairano, and J.-P. Lai, "Falling yes/no questions in Corsican French and Corsican: evidence for a prosodic transfer," *Prosody and language in contact: L2 acquisition, attrition and languages in multilingual situations*, pp. 101–122, 2015.

[20] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot. Int.*, vol. 5, no. 9, pp. 341–345, 2001.

[21] V. Pagel, N. Carbonell, and Y. Laprie, "A new method for speech delexicalization, and its application to the perception of French prosody," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, 1996, vol. 2, pp. 821–824.

[22] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech communication*, vol. 9, no. 5–6, pp. 453–467, 1990.

[23] F. Ramus and J. Mehler, "Language identification with suprasegmental cues: A study based on speech resynthesis," *The Journal of the Acoustical Society of America*, vol. 105, no. 1, pp. 512–521, 1999.

[24] T. Dubeda, "Prosodic boundaries in Czech: an experiment based on delexicalized speech," in *Ninth International Conference on Spoken Language Processing*, 2006.

[25] C. Gobl and A. N. Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech communication*, vol. 40, no. 1–2, pp. 189–212, 2003.

[26] M. Vainio, A. Suni, T. Raitio, J. Nurminen, J. Järvikivi, and P. Alku, "New method for delexicalization and its application to prosodic tagging for text-to-speech synthesis," in *Proc. Interspeech 2009*, 2009, pp. 1703–1706. doi: 10.21437/Interspeech.2009-514.

[27] A. Kain and J. P. H. van Santen, "Frequency-domain delexicalization using surrogate vowels," in *Proc. Interspeech 2010*, 2010, pp. 474–477. doi: 10.21437/Interspeech.2010-201.

[28] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. Van der Vrecken, "The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, 1996, vol. 3, pp. 1393–1396.

[29] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoustical science and technology*, vol. 27, no. 6, pp. 349–353, 2006.

[30] MATLAB, version 9.13.0.2080170 (R2022b). Natick, Massachusetts: The MathWorks Inc., 2022.

[31] R. Corretge, "Praat vocal toolkit: A praat plugin with automated scripts for voice processing," *Software package (http://www. praatvocaltoolkit. com/index. html)*, 2012.

[32] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015, doi: 10.18637/jss.v067.i01.

[33] R. V. Lenth *et al.*, "emmeans: Estimated marginal means, aka least-squares means (1.8.4-1)[R-Package]." 2023.

[34] A. Kuznetsova, P. B. Brockhoff, and R. H. Christensen, "lmerTest package: tests in linear mixed effects models," *Journal of statistical software*, vol. 82, pp. 1–26, 2017.

[35] S. Palva *et al.*, "Distinct gamma-band evoked responses to speech and non-speech sounds in humans," *The Journal of Neuroscience*, vol. 22, no. 4, p. RC211, 2002.

[36] J. Laver, "The phonetic description of voice quality," *Cambridge Studies in Linguistics London*, vol. 31, pp. 1–186, 1980.