Discrete/Continuous Modelling of Speaking Style in HMM-based Speech Synthesis: Design and Evaluation

Nicolas Obin, Pierre Lanchantin, Anne Lacheret, Xavier Rodet

Discrete/Continuous Modelling of Speaking Style in HMM-based Speech Synthesis: Design and Evaluation

Nicolas Obin [1,2], *Pierre Lanchantin* [1]
Anne Lacheret [2], *Xavier Rodet* [1]

[1] Analysis-Synthesis Team, IRCAM, Paris, France
[2] Modyco Lab., University of Paris Ouest - La Défense, Nanterre, France
nobin@ircam.fr, lanchantin@ircam.fr, anne.lacheret@u-paris10.fr, rodet@ircam.fr

Abstract

This paper assesses the ability of a HMM-based speech synthesis systems to model the speech characteristics of various speaking styles[1]. A discrete/continuous HMM is presented to model the symbolic and acoustic speech characteristics of a speaking style. The proposed model is used to model the average characteristics of a speaking style that is shared among various speakers, depending on specific situations of speech communication. The evaluation consists of an identification experiment of 4 speaking styles based on delexicalized speech, and compared to a similar experiment on natural speech. The comparison is discussed and reveals that discrete/continuous HMM consistently models the speech characteristics of a speaking style.
Index Terms: speaking style, speech synthesis, speech prosody, average modelling.

1. Introduction

Each speaker has his own *speaking style* which constitutes his vocal signature, and a part of his identity. Nevertheless, a speaker continuously adapt his speaking style according to specific communication situations, and to his emotional state. In particular, each situational context determines a specific mode of production associated with it - a *genre* - which is defined by a set of conventions of form and content that are shared among all of its productions [1]. In particular, a specific discourse genre (DG) relates to a specific *speaking style*. Consequently, a speaker adapts his speaking style to each specific situation depending on the formal conventions that are associated with the situation, his a-priori knowledge about these conventions, and his competence to adapt his speaking style. Thus, each communication act instantiates a style which is composed of a style that depends on the speaker identity, and a conventional speaking style that is conditioned by a specific situation.
In speech synthesis, methods have been proposed to model and adapt the symbolic [2, 3] and acoustic speech characteristics of a speaking style, with application to emotional speech synthesis [4]. However, no study exists on the joint modelling of the symbolic and acoustic characteristics of speaking style, and speaking style acoustic modelling generally limits to the modelling of emotion, with rare extensions to other sources of speaking styles variations [5].

[1]This study was partially funded by "La Fondation Des Treilles", and supported by ANR Rhapsodie 07 Corp-030-01; reference prosody corpus of spoken French; French National Agency of research; 2008-2012.

This paper presents an average discrete/continuous HMM which is applied to the speaking style modelling of various discourse genres in speech synthesis, and assesses whether the model adequately captures the speech prosody characteristics of a speaking style. Incidentally, the robustness of the HMM-based speech synthesis is evaluated in the conditions of real-world applications. The paper is organized as follows: the speaking style corpus design is described in section 2; the average discrete/continuous HMM model is presented in section 3; the evaluation is presented and discussed in sections 4 and 5.

2. Speech & Text Material

2.1. Corpus Design

For the purpose of speaking style speech synthesis, a 4-hour multi-speakers speech database was designed. The speech database consists of four different DG's: catholic mass ceremony, political, journalistic, and sport commentary. In order to reduce the DG intra-variability, the different DGs were restricted to specific situational contexts (see list below) and to male speakers only.

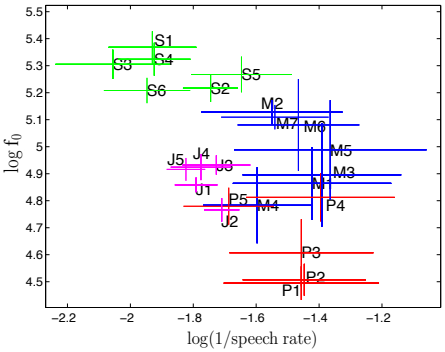


Figure 1: Prosodic description of the speaking styles depending on the speaker. Mean and variance of f_0 and speech rate (syllable per second).

The following is a description of the four selected DG's:

mass: Christian church sermon (pilgrimage and Sunday high-mass sermons); single speaker monologue, no interaction.

political: New Year's French president speech; single speaker monologue; no interaction.

journal: radio review (press review; political, economical,

technological chronicles); almost single speaker monologue with a few interactions with a lead journalist.

sport commentary: soccer; two speakers engaged in monologues with speech overlapping during intense soccer sequences and speech turn changes; almost no interaction.
The speech database consists of *natural speech* multi-media audio contents with strongly variable audio quality (background noise: crowd, audience, recording noise, and reverberation). The speech prosody characteristics of the speech databased are illustrated in figure 1.

3. Speaking Style Model

A speaking style model $\boldsymbol{\lambda}^{(style)}$ is composed of discrete/continuous context-dependent HMMs that model the symbolic/acoustic speech characteristics of a speaking style.

$$\boldsymbol{\lambda}^{(style)} = \left(\boldsymbol{\lambda}^{(style)}_{\text{symbolic}}, \boldsymbol{\lambda}^{(style)}_{\text{acoustic}}\right) \tag{1}$$

During the training, the discrete/continuous context-dependent HMMs are estimated separately. During the synthesis, the symbolic/acoustic parameters are generated in cascade, from the symbolic representation to the acoustic variations. Additionally, a rich linguistic description of the text characteristics is automatically extracted using a linguistic processing chain [6] and used to refine the context-dependent HMM modelling (see [7] and [8] for a detailed description of the enriched linguistic contexts).

3.1. Training of the Discrete/Continuous Models

3.1.1. Discrete HMM

For each speaking style, an average context-dependent discrete HMM $\boldsymbol{\lambda}^{(style)}_{\text{symbolic}}$ is estimated from the pooled speakers associated with the speaking style.

The prosodic grammar consists of a hierarchical prosodic representation that was experimented as an alternative to ToBI [9] for French prosody labelling [10]. The prosodic grammar is composed of major prosodic boundaries (F_M, a boundary which is right bounded by a pause), minor prosodic boundaries (F_m, an intermediate boundary), and prosodic prominences (P).

Let R be the number of speakers from which an average model $\boldsymbol{\lambda}^{(style)}_{\text{symbolic}}$ is to be estimated. Let $\mathbf{l} = (\mathbf{l}^{(1)}, \ldots, \mathbf{l}^{(R)})$ the total set of prosodic symbolic observations, and $\mathbf{l}^{(r)} = [l^{(r)}(1), \ldots, l^{(r)}(N_r)]$ the prosodic symbolic sequence associated with speaker r, where $l^{(r)}(n)$ is the prosodic label associated with the n-th syllable. Let $\mathbf{q} = (\mathbf{q}^{(1)}, \ldots, \mathbf{q}^{(R)})$ the total set of linguistic contexts observations, and $\mathbf{q}^{(r)} = [\mathbf{q}^{(r)}(1), \ldots, \mathbf{q}^{(r)}(N_r)]$ the linguistic context sequence associated with speaker r, where $\mathbf{q}^{(r)}(n) = [q_1^{(r)}(n), \ldots, q_L^{(r)}(n)]^\top$ is the (Lx1) linguistic context vector which describes the linguistic characteristics associated with the n-th syllable.

An average context-dependent discrete HMM $\boldsymbol{\lambda}^{(style)}_{\text{symbolic}}$ is estimated from the pooled speakers observations. Firstly, an average context-dependent tree $\text{T}^{(style)}_{\text{symbolic}}$ is derived so as to minimize the information entropy of the prosodic symbolic labels $\mathbf{l}$ conditionally to the linguistic contexts $\mathbf{q}$. Then, a context-dependent HMM model $\boldsymbol{\lambda}^{(style)}_{\text{symbolic}}$ is estimated for each terminal node of the context-dependent tree $\text{T}^{(style)}_{\text{symbolic}}$.

3.1.2. Continuous HMM

For each speaking style, an average acoustic model $\boldsymbol{\lambda}^{(style)}_{\text{acoustic}}$ that includes source/filter variations, f_0 variations, and state-durations, is estimated from the pooled speakers associated with the speaking style based on the conventional HTS system [11].

Let R be the number of speakers from which an average model is to be estimated. Let $\mathbf{o} = (\mathbf{o}^{(1)}, \ldots, \mathbf{o}^{(R)})$ the total set of observations, and $\mathbf{o}^{(r)} = [\mathbf{o}^{(r)}(1), \ldots, \mathbf{o}^{(r)}(T_r)]$ the observation sequences associated with speaker r, where $\mathbf{o}^{(r)}(t) = [o_t^{(r)}(1), \ldots, o_t^{(r)}(D)]^\top$ is the (Dx1) observation vector which describes the acoustical property at time t. Let $\mathbf{q} = (\mathbf{q}^{(1)}, \ldots, \mathbf{q}^{(R)})$ the total set of linguistic contexts observations, and $\mathbf{q}^{(r)} = [\mathbf{q}^{(r)}(1), \ldots, \mathbf{q}^{(r)}(T_r)]$ the linguistic context sequence associated with speaker r, where $\mathbf{q}^{(r)}(t) = [q_1^{(r)}(t), \ldots, q_L^{(r)}(t)]^\top$ is the (Lx1) linguistic context vector which describes the linguistic properties at time t.

An average context-dependent HMM acoustic model $\boldsymbol{\lambda}^{(style)}_{\text{symbolic}}$ is estimated from the pooled speakers observations. Firstly, a context-dependent HMM model is estimated for each of the linguistic contexts. Then, an average context-dependent tree $\text{T}^{(style)}_{\text{acoustic}}$ is derived so as to minimize the description length of the context-dependent HMM model $\boldsymbol{\lambda}^{(style)}_{\text{acoustic}}$.

The acoustic module models simultaneously source/filter variations, f_0 variations, and the temporal structure associated with a speaking style. Speakers f_0 were normalized with respect to the speaking style prior to modelling. Source, filter, and normalized f_0 observation vectors and their dynamic vectors are used to estimate context-dependent HMM models $\boldsymbol{\lambda}^{(style)}_{\text{acoustic}}$. Context-dependent HMMs are clustered into acoustically similar models using decision-tree-based context-clustering (ML-MDL [11]). Multi-Space probability Distributions (MSD) [12] are used to model continuous/discrete parameter f_0 sequence to manage voiced/unvoiced regions properly. Each context-dependent HMM is modelled with a state duration probability density functions (PDFs) to account for the temporal structure of speech [13]. Finally, speech dynamic is modelled according to the trajectory model and the global variance (GV) that model local and global speech variations over time [14].

3.2. Generation of the Speech Parameters

During the synthesis, the text is first converted into a concatenated sequence of context-dependent HMM models $\boldsymbol{\lambda}^{(style)}_{\text{symbolic}}$ associated with the linguistic context sequence $\mathbf{q} = [\mathbf{q}_1, \ldots, \mathbf{q}_N]$, where $\mathbf{q}_n = [q_1, \ldots, q_L]^\top$ denotes the (Lx1) linguistic context vector associated with the n-th phoneme.

Firstly, the prosodic symbolic sequence $\widehat{\mathbf{l}}$ is determined so as to maximize the likelihood of the prosodic symbolic sequence $\mathbf{l}$ conditionally to the linguistic context sequence $\mathbf{q}$ and the model $\boldsymbol{\lambda}^{(style)}_{\text{symbolic}}$.

$$\widehat{\mathbf{l}} = \underset{\mathbf{l}}{\text{argmax}}\, \text{p}(\mathbf{l}|\mathbf{q}, \boldsymbol{\lambda}^{(style)}_{\text{symbolic}}) \tag{2}$$

Then, the linguistic context sequence $\mathbf{q}$ augmented with the prosodic symbolic sequence $\widehat{\mathbf{l}}$ is converted into a concatenated

sequence of context-dependent models $\boldsymbol{\lambda}_{\text{acoustic}}^{(style)}$.

The acoustic sequence $\widehat{\mathbf{o}}$ is inferred so as to maximize the likelihood of the acoustic sequence $\mathbf{o}$ conditionally to the model $\boldsymbol{\lambda}_{\text{acoustic}}^{(style)}$.

$$\widehat{\mathbf{o}} = \underset{\mathbf{o}}{\text{argmax}} \max_{\mathbf{q}} \text{p}(\mathbf{o}|\mathbf{q}, \boldsymbol{\lambda}_{\text{acoustic}}^{(style)}) \text{p}(\mathbf{q}|\boldsymbol{\lambda}_{\text{acoustic}}^{(style)}) \quad (3)$$

First, the state sequence $\widehat{\mathbf{q}}$ is determined so as to maximize the likelihood of the state sequence conditionally to the model $\boldsymbol{\lambda}_{\text{acoustic}}^{(style)}$. Then, the observation sequence $\widehat{\mathbf{c}}$ is determined so as to maximize the likelihood of the observation sequence conditionally to the state sequence $\widehat{\mathbf{q}}$, the model $\boldsymbol{\lambda}_{\text{acoustic}}^{(style)}$ under dynamic constraint $\mathbf{o} = \mathbf{W}\mathbf{c}$.

$$\mathbf{R}_{\widehat{\mathbf{q}}}\widehat{\mathbf{c}} = \mathbf{r}_{\widehat{\mathbf{q}}} \quad (4)$$

where:

$$\mathbf{R}_{\widehat{\mathbf{q}}} = \mathbf{W}^\top \boldsymbol{\Sigma}_{\widehat{\mathbf{q}}}^{-1} \mathbf{W}. \quad (5)$$

$$\mathbf{r}_{\widehat{\mathbf{q}}} = \mathbf{W}^\top \boldsymbol{\Sigma}_{\widehat{\mathbf{q}}}^{-1} \boldsymbol{\mu}_{\widehat{\mathbf{q}}}. \quad (6)$$

and $\boldsymbol{\Sigma}_{\widehat{\mathbf{q}}}$ and $\boldsymbol{\mu}_{\widehat{\mathbf{q}}}$ are respectively the covariance matrix and the mean vector for the state sequence $\widehat{\mathbf{q}}$.

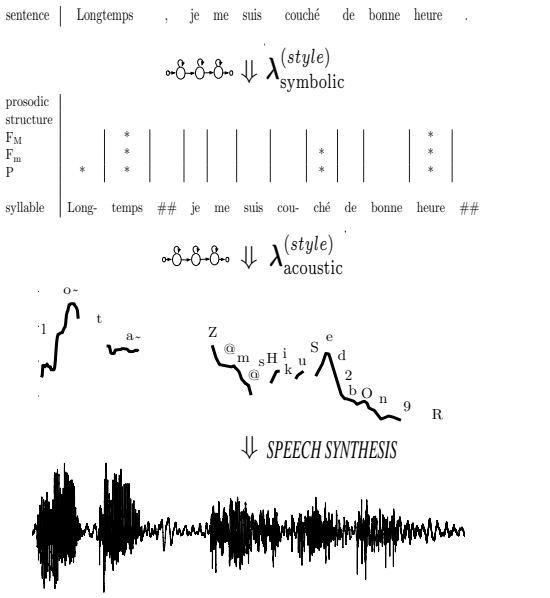


Figure 2: Generation of discrete/continuous speech parameters for the sentence: "*Longtemps, je me suis couché de bonne heure*" ("*For a long time I used to go to bed early*").

4. Evaluation

The proposed model has been evaluated based on a speaking style identification perceptual experiment, and compared to a speaking style identification experiment with natural speech [15]. For the purpose of such a comparison, it was necessary to provide a single evaluation scheme for both experiments. In particular, it was not possible to control the linguistic content of natural speech utterances which provides evident cues for DG's identification (a single keyword would be sufficient to identify a DG). Thus, such a comparison required to remove lexical access and to focus on the prosodic dimension only.

4.1. Experimental Setup

40 speech utterances (10 per DG) were selected in the speaking style corpus and removed from the training set. Lexical access was removed using a band-pass filter that insured that the lowest frequency of the fundamental frequency and the highest frequency of its first harmonic was included.

4.2. Subjective Evaluation

The evaluation consists of a multiple choice identification task from speech prosody perception. The evaluation was conducted according to crowd-sourcing technique using social networks. 50 subjects (including 25 native French speakers, 15 non-native French speakers, 10 non-French speakers; 34 expert and 16 naïve listeners) participated in this experiment. Participants were given a brief description of the different speaking styles. Then, they were asked to associate a speaking style to each of the speech utterances. For this purpose, participants were given three options:

total confidence: select only one speaking style when certain of the choice;

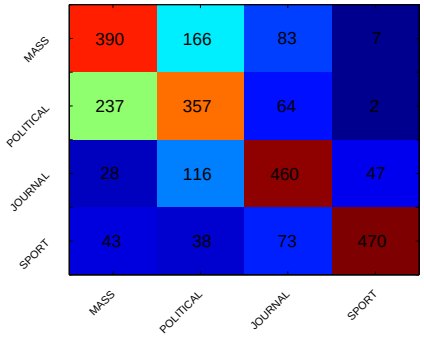
confusion: select two different speaking styles when two speaking styles are possible;

total indecision: select "indecision" when completely unsure. Subjects were asked to use this possibility only as a very last resort.

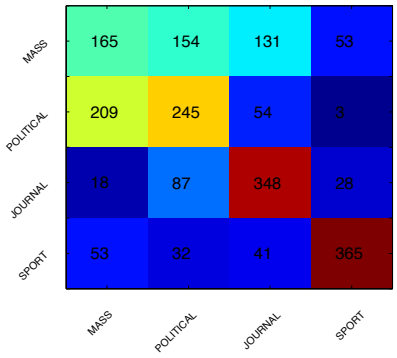
Additional informations were gleaned from the participants: speech expertise (expert, naïve), language (native French speaker, non-native French speaker, non-French speaker), age, and listening condition (headphones or not). *Expert* participants were actually coming from various domains (speech and audio technologies, linguistics, musicians). Participants were encouraged to use headphones.

5. Results & Discussion

Identification performance was estimated using a measure based on Cohen's Kappa statistic [16]. Cohen's Kappa statistic measures the proportion of agreement between two raters with correction for random agreement. Our measure monitors the agreement between the ratings of the participants and the ground truth. The measure varies from -1 to 1: -1 is perfect disagreement; 0 is chance; 1 is perfect agreement. *Confusion* ratings were considered as equally possible ratings. *Total indecision* ratings were relatively rare (3% of the total ratings) and removed. Figure 3 presents the identification confusion matrix. Overall score reveals fair identification performance ($\kappa = 0.38 \pm 0.04$) which is comparable to that observed for identification from natural speech ($\kappa_{\text{natural}} = 0.45 \pm 0.03$). The identification performance significantly depends on the speaking style (figure 4): sport commentary is substantially identified ($\kappa = 0.68 \pm 0.05$), journal fairly identified ($\kappa = 0.50 \pm 0.06$), political discourse moderately identified ($\kappa = 0.28 \pm 0.07$), and mass only slightly identified ($\kappa = 0.12 \pm 0.06$). In comparison with identification from natural speech, the identification is comparable in the case of the sport commentary and the journal speaking styles ($\kappa_{\text{natural}} = 0.70 \pm 0.03$ and $\kappa_{\text{natural}} = 0.54 \pm 0.05$, respectively). However, there is a drop in identification for the political and the mass speaking styles which is especially significant for the mass style ($\kappa_{\text{natural}} = 0.34 \pm 0.05$ and $\kappa_{\text{natural}} = 0.38 \pm 0.04$, respectively). This indicates that the model somehow failed to capture the relevant cues of the corresponding speaking style. Nevertheless, a large confusion



(a) natural speech



(b) synthetic speech

Figure 3: Identification confusion matrices. Rows represent synthesized speaking style. Columns represent identified speaking style.

exists between the political and the mass speech that is inherent to a similarity in the speaking style and the formal situation in which the speech occurs. Additionally, the conventional HMM-based speech synthesis system failed into modelling adequately the breathyness and the creakyness that is specific to the political speaking style, especially within unvoiced segments.

ANOVA analysis was conducted to assess whether the identification performance depends on the language of the participants. Analysis reveals a significant effect of the language ($F(2, 59) = 15, p < 0.001$) (F(48,2)=5.9, p-value=0.005), and confirms results obtained for natural speech. This confirms evidence that there exists variations of a speaking style depending on the language and/or cultural background.

Finally, an informal evaluation of the quality of the synthesized speech suggests that the speaking style modelling is robust to the large variety of audio quality.

6. Conclusion

In this study, the ability and the robustness of a HMM-based speech synthesis system to model the speech characteristics of various speaking styles were assessed. A discrete/continuous HMM was presented to model the symbolic and acoustic speech characteristics of a speaking style, and used to model the average characteristics of a speaking style that is shared among various speakers, depending on specific situations of speech communication. The evaluation consisted of an identification experiment of 4 speaking styles based on delexicalized speech, and compared with a similar experiment on natural speech. The evaluation showed that the discrete/continuous HMM consistently models the speech characteristics of a speaking style, and is robust to the differences in audio quality. This proves evidence that the discrete/continuous HMM speech synthesis system successfully models the speech characteristics of a speaking style in the conditions of real-world applications.

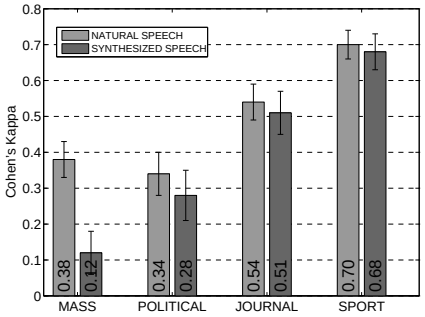


Figure 4: Mean identification scores and 95% confidence interval obtained for natural and synthesized speech.

7. References

[1] A.-C. Simon, A. Auchlin, M. Avanzi, and J.-P. Goldman, *Les voix des Français*. Peter Lang, 2009, ch. Les phonostyles: une description prosodique des styles de parole en français.

[2] H. Schmid and M. Atterer, "New statistical methods for phrase break prediction," in *International Conference On Computational Linguistics*, Geneva, Switzerland, 2004, pp. 659–665.

[3] P. Bell, T. Burrows, and P. Taylor, "Adaptation of prosodic phrasing models," in *Speech Prosody*, Dresden, Germany, 2006.

[4] J. Yamagishi, T. Masuko, and T. Kobayashi, "HMM-based expressive speech synthesis - Towards TTS with arbitrary speaking styles and emotions," in *Special Workshop in Maui*, Maui, Hawaï, 2004.

[5] S. Krstulović, A. Hunecke, and M. Schröder, "An HMM-based speech synthesis system applied to german and its adaptation to a limited set of expressive football announcements," in *Interspeech*, 2007.

[6] E. Villemonte de La Clergerie, "From metagrammars to factorized TAG/TIG parsers," in *International Workshop On Parsing Technology*, Vancouver, Canada, Oct. 2005, pp. 190–191.

[7] N. Obin, P. Lanchantin, A. Lacheret, and X. Rodet, "Towards improved HMM-based speech synthesis using high-level syntactical features," in *Speech Prosody*, Chicago, U.S.A., 2010.

[8] A. Lacheret, N. Obin, and M. Avanzi, "Design and Evaluation of Shared Prosodic Annotation for Spontaneous French Speech: From Expert Knowledge to Non-Expert Annotation," in *Linguistic Annotation Workshop*, Uppsala, Sweden, 2010.

[9] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: a standard for labeling english prosody," in *International Conference of Spoken Language Processing*, Banff, Canada, 1992, pp. 867–870.

[10] N. Obin, A. Lacheret, and X. Rodet, "Expectations for Speaking Style Identification: a Prosodic Study," in *Interspeech*, Makuhari, Japan, 2010, pp. 3070–3073.

[11] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *European Conference on Speech Communication and Technology*, Budapest, Hungary, 1999, pp. 2347–2350.

[12] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *International Conference on Audio, Speech, and Signal Processing*, Phoenix, Arizona, 1999, pp. 229–232.

[13] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-Markov model based speech synthesis," in *International Conference on Speech and Language Processing*, Jeju Island, Korea, 2004, pp. 1397–1400.

[14] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Transactions on Information and Systems*, vol. 90, no. 5, pp. 816–824, 2007.

[15] N. Obin, A. Lacheret, and X. Rodet, "HMM-based prosodic structure model using rich linguistic context," in *Interspeech*, Makuhari, Japan, 2010, pp. 1133–1136.

[16] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.