


The evaluation of prosody in speech synthesis: a systematic review

Julio Cesar Galdino   [University of São Paulo | juliogaldino@usp.br]

Ariadne Nascimento Matos  [University of São Paulo | ariadnenmtos@usp.br]

Flaviane Romani Fernandes Svartman  [University of São Paulo | flavianesvartman@usp.br]

Sandra Maria Aluisio  [University of São Paulo | sandra@icmc.usp.br]

 Instituto de Ciências Matemáticas e de Computação (ICMC), University of São Paulo (USP), Av. Trabalhador São-carlense, 400, São Carlos, SP, 13566-590, Brazil.

Received: 22 January 2025 • Accepted: 29 April 2025 • Published: 09 July 2025

Abstract This paper presents a systematic review on the relationship between prosody and speech synthesis, focusing on the evaluation of prosodic parameters of synthesized speech. The relevance of the topic lies in the fact that the task of speech synthesis has not yet been resolved, therefore the information obtained in this review can contribute to knowledge and to the improvement of the methodologies used in evaluating the prosody of synthesized speech. To select studies, we used the Parsifal platform, including 100 studies published between 2020 and 2024, with the purpose of answering eight previously established research questions. The highlights of this systematic review are presented in the following. The main prosodic parameters considered in speech synthesis systems are fundamental frequency (F0), duration and intensity, with F0 standing out (95 studies). The metric most frequently used in studies belongs to the group of acoustic metrics — F0-RMSE (Root mean-squared error evaluation of F0). Lower values of this metric indicate greater proximity between the F0 of synthesized speech and that of natural speech. The most used dataset was LJ Speech, a public domain speech dataset consisting of English audio clips of a single speaker reading short excerpts from seven non-fiction books, reinforcing that the predominant language was English — 48 studies focus on English to evaluate speech description prosody, although there is a relevant number of studies in Mandarin Chinese (27 studies) and Japanese (15 studies). Most studies used models as a baseline to compare the performance of their methods or proposed new models in order to improve the prosody of synthesized speech. Each study presented different methods for this improvement, according to the objectives, such as learning prosodic features extracted from reference speech and adding auxiliary modules to existing model architectures. As highlighted baselines, there was recurrent use of Tacotron 2, which generates mel-spectrograms from text and then synthesizes speech from the generated mel-spectrograms using a separately trained vocoder, and FastSpeech 2, which can extract explicit prosodic features to be directly used as entry into training.

Keywords: Speech Synthesis, TTS, Prosody, Objective Evaluation, Subjective Evaluation

1 Introduction

Speech synthesis refers to the synthetic (computerized) generation of speech and the conversion of written text to speech (TTS – Text-to-Speech) [Taylor, 2009]. Text-to-speech synthesis is inherently multidisciplinary, whose areas involve electrical engineering (signal processing), linguistics, computer science, among others [Van Santen *et al.*, 1997]. Current advances in speech synthesis models are attributed to the great advances in deep learning research, which allowed the integration of specific modules from traditional synthesis systems into a single model, resulting in improvements in the quality of synthesis [Casanova *et al.*, 2024].

Several methods have been proposed for this task, but regardless of the method, the generated voice presents suprasegmental (prosodic) features of speech that need to be evaluated. These parameters, larger than the segments (vowels, consonants), can include elements of melody, such as tone and intonation, and speech dynamics, such as duration, pause, rhythm, among others [Cagliari, 1992].

The relationship between speech synthesis and prosody has been the subject of study for decades. For example, Sag-

isaka *et al.* [1997] present a collection of papers on computational approaches to processing the prosody of spontaneous speech, exploring the generation and modeling of prosody in computational synthesis. The evaluation of prosodic aspects of synthesized speech is also old — studies such as Jokisch *et al.* [2000] and Chen *et al.* [1998] already used objective metrics to evaluate prosodic parameters, such as Mean Squared Error (RMSE), while Hirst *et al.* [1998] already proposed subjective evaluations of prosody, which allowed untrained participants to subjectively evaluate TTS systems, pointing out which parts of a recording were unsatisfactory when clicking a mouse.

Surveys and systematic literature reviews on speech synthesis have been carried out in order to contribute to knowledge in the area.

Wagner *et al.* [2019] review the state of the art in TTS evaluation, proposing a user-centered research program. The authors highlight evaluation methods for different TTS applications, such as Virtual assistants, navigation, audio books, voice prostheses, that require specific criteria, such as intelligibility, comprehension, preference, similarity to original voice, among others. These evaluations are subjective or ob-

jective and are based on behavior (virtual assistants need to have clear and pleasant voices, robots need adequate voices, screen readers require intelligibility, etc.).

Cooper *et al.* [2024] analyzed objective and subjective evaluations, serving as a starting point for beginning researchers and speech synthesis professionals, offering a comprehensive historical view on advances and challenges in the field. The work offers a great contribution, discussing what studies evaluated in each decade, from 1980 to 2024, providing a broad view of each metric used, from initial intelligibility (comprehensibility) tests in the laboratory to more current metrics, such as large-scale crowdsourcing mean opinion score (MOS) tests.

Other reviews on speech synthesis focus on specific aspects, such as synthesis in less studied languages. Chemnad and Othman [2023] present a systematic literature review of 22 years (2000 to 2022) on speech synthesis in Arabic, evaluating synthesis methods, corpora used and evaluation methods. Analyzing 36 papers, the authors conclude that the language lacks free and open source datasets for training and that naturalness and comprehensibility criteria dominate system evaluations in this language.

The generation of emotional speech is also the subject of specific synthesis reviews. Alwaisi and Németh [2024] explore the methodologies used to improve the expressiveness of synthetic speech in objective metrics (such as Itakura-Saito (I-S) measure, Root mean square (RMSE), Gross pitch error (GPE)) and subjective metrics (MOS, AB Preference Test, ABX Preference Test and MUSHRA — Multiple Stimuli with Hidden Reference and Anchor). The review provides a comprehensive overview of historical and contemporary trends in synthesis and indicates future directions for advancement in expressive speech synthesis.

Do *et al.* [2021] present a systematic review based on 23 studies (2006-2020) that use multilingual data for TTS of low-resource languages. They analysed strategies used by those studies for incorporating multilingual data and how they affect output speech quality. Important to say that the review shows the evaluation metrics used in those studies, grouping them into five classes: acoustics, intelligibility, naturalness, quality, similarity, including both objective and subjective evaluation metrics. They bring 23 different evaluation metrics. Moreover, to investigate the difference in output quality between corresponding monolingual and multilingual models, they present a novel measure to compare this difference across the studies and evaluation metrics.

One of the existing specific systematic reviews aimed to review studies on speech synthesis and prosody. Galdino and Oliveira Jr [2023] identified 10 studies between 2010 and 2021, written in Portuguese, on the evaluation of the Portuguese language. The work showed that fundamental frequency was the most commonly used prosodic features to improve synthetic speech. The authors highlighted the importance of collaborations between linguists and researchers in the areas of computing and related areas, in order to bring better results to the area of research in speech synthesis.

Although there are significant surveys and systematic literature reviews on speech synthesis, and specifically on speech synthesis and prosody, such as those described above, this systematic review seeks to fill gaps left by previous works,

presenting the following contributions:

1. It deals with the most recent works (2020 to 2024) on the relationship between speech synthesis and prosody, to show recent trends in the area, including only publications written in English, expanding the existing review in Portuguese.
2. Covers a comprehensive analysis of 100 studies, which allows a remarkable study on objective and subjective metrics, grouping them into four categories: Acoustic, Intelligibility, Naturalness/Quality and Similarity.
3. Adopts criteria for selecting studies, from six databases, in order to provide information on eight research questions that detail the assessments carried out.
4. Provides information on prosodic parameters, metrics, terminologies, datasets, languages, synthesis models used, comparison between the prosody of synthetic speech and natural speech, limitations and challenges, encouraging future studies to fill research gaps and improve the methodologies used on synthetic speech, since prosody plays an important role in speech.

2 Methodology

2.1 Research Questions

To achieve the goal of this systematic review of the literature, eight research questions were formulated to guide the extraction of information from the studies and to support the discussion developed in this review:

1. Which prosodic parameters have been considered in speech synthesis systems?
2. What terminology is used for evaluation metrics?
3. What methodologies are employed to evaluate prosody in the studies? Are these methodologies predominantly manual, automatic, or hybrid?
4. What types of corpora/datasets have been used in research (read speech or spontaneous speech), how is the transcription process for these resources conducted, and how are they described in terms of quality of the audio recordings?
5. Which languages have been most frequently used in research focused on speech synthesis and prosody?
6. Do studies make comparisons between different speech synthesis models? If so, which models have been most used and which have the greatest potential?
7. Do studies make comparisons between the prosody of synthetic and natural speech? If so, what are the main conclusions of those studies?
8. What are the main challenges currently limiting the enhancement of prosody in speech synthesis systems?

2.2 Search Strategy

The review was carried out using the Parsifal¹ platform, an online tool that allows documentation of the entire review process, designed in the master's dissertation of Freitas e

¹<https://parsif.al/>

Souza [2015]. Its features allow users to input research questions, search strings, databases, and other relevant information necessary to perform a review with predefined criteria established by the researchers themselves. Additionally, the platform offers mechanisms to import studies from databases, initial screening, automatic identification of duplicates, and establishment of quality assessment criteria during the reading phase of the full text of the studies.

For this review, six scientific databases were used:

- Association for Computing Machinery (ACM) digital library
- Institute of Electrical and Electronics Engineers (IEEE) Xplore
- Google Scholar
- Journal of the Acoustical Society of America (JASA)
- Scopus
- SpringerLink

The research was conducted using the following search string: (prosody OR intonation OR prosodic OR “prosody modeling” OR “fundamental frequency”) AND (“text-to-speech” OR TTS OR “speech synthesis” OR “synthesis model*”) AND (“objective evaluation” OR “objective assessment” OR “evaluation metric*” OR “evaluation method*”) AND (“subjective evaluation” OR “subjective assessment”) AND (acoustic* OR intelligibility OR naturalness OR quality OR similarity OR “latency time” OR “latency test” OR “noise robustness” OR “background noise”). This search string was applied to the previously mentioned databases.

We included open access studies published between 2020 and 2024 (a period of five years), written in English and specifically addressing prosodic aspects. Studies were excluded if they merely mentioned prosody without evaluating it, focused exclusively on segmental production, or were dedicated solely to automatic speech recognition. We also excluded publications from journals and conferences without a peer-review process, as well as studies that provided only an abstract. Furthermore, we excluded literature reviews, opinion papers, duplicates across databases, books and book chapters, studies with unclear evaluation metrics, and those that did not describe the corpus used. The adopted inclusion (I) and exclusion (E) criteria for the selection of studies are summarized below:

- (I1) open access studies;
- (I2) studies only written in English;
- (I3) studies between 2020 and 2024 (5 years);
- (I4) published studies;
- (I5) studies that evaluate prosodic aspects;
- (E1) documents with only the abstract;
- (E2) documents that only mention prosody but do not evaluate it;
- (E3) literature review and opinion studies;
- (E4) duplicate studies;
- (E5) studies outside the scope of the theme;
- (E6) books, book chapters and entire proceedings;
- (E7) journals and conferences that do not have a peer review process;
- (E8) studies whose evaluation metrics are not clear.

The selected studies were imported into the Parsifal platform. However, to the best of our knowledge, the Google Scholar and JASA databases did not support exporting study lists into a single CSV file. Consequently, the subsequent steps were partially completed within Parsifal, while the steps involving these two databases were managed using external documents.

In the first stage, we conducted a title and abstract screening within Parsifal and externally for studies retrieved from Google Scholar and JASA. In the second stage, we performed a quality assessment, in which the studies selected based on their titles and abstracts were fully read and evaluated according to the pre-established criteria defined by the authors of this review:

- QA1. Are the objectives of the study clearly defined and aligned with the research question presented?
- QA2. Are the datasets/corpus used clearly described, including details about origin, size, characteristics, and justification for their selection?
- QA3. Are the experiments described in detail, allowing an understanding of their execution and the reproducibility of the results?
- QA4. Does the study use pre-trained models or methods? If so, are the pretraining configurations described in detail, including datasets, parameters, and objectives?
- QA5. Are the results and conclusions of the study useful and relevant to answering the research questions posed?
- QA6. Is prosody approached as a central element in evaluating the quality or naturalness of speech synthesis?

The evaluation process was as follows: each criterion was scored as “1.0 = Yes”, “0.5 = Partially” or “0.0 = No”. The sum of the six criteria served as the basis for the final decision on whether to include the study. Studies that scored more than 3 points, more than half of the total possible score (6.0), were definitively included in the review. Studies that scored 3 or less were excluded. The summary is shown in Figure 1.

After the aforementioned steps, studies relevant to the topic were added, which met all inclusion criteria, but which, however, had not been returned in the searches carried out.

2.3 Screening and inclusion

The database search stage returned a total of 1682 documents (**Table 1**). After identifying and excluding 36 duplicate studies, 1,646 records remained in the databases. During the screening stage, based on reading the title and abstract, 210 studies were selected for the quality assessment stage. Of these, 114 were discarded, generating a provisional total of 96 eligible studies. Subsequently, four more studies were added², resulting in 100 studies included in this review (**Figure 2**)³.

²These four studies were already being studied by the authors and appear with * in Table A.

³Figure based on the PRISMA flow diagram.

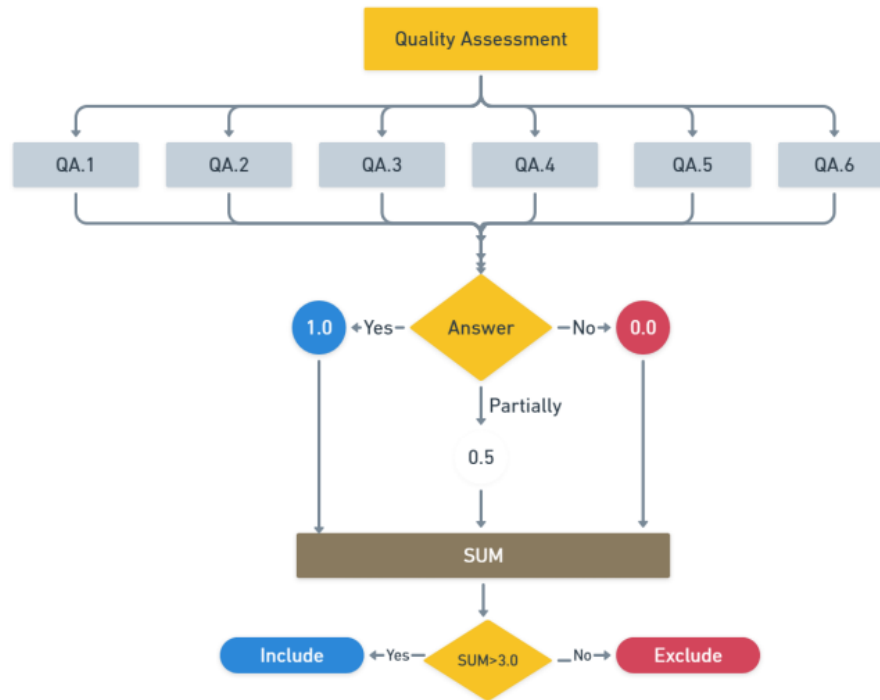


Figure 1. Quality Assessment Process

Table 1. Summary of the search string stage in the databases

Database	Total	By year	Language	Books and chapters	Open access	Final
Scopus	101	40	-	38	36	36
ACM	232	189	-	188	-	188
Springer	1,367	1,157	1,152	1,061	91	91
IEEE	115	47	-	-	-	47
Scholar	2,600	1,150	1,143	1,135	1,089	1,089
JASA	1,121	299	-	274	231	231
Total						1,682

3 Overall Results

The main aspects extracted from the accepted papers in this systematic review are presented in the Appendix A (Table A). Studies published in 2020 were small in number compared to subsequent years, maybe due to the pandemics, indicating an increase in the relationship between speech synthesis and prosody over time. In 2024, there was a reduction in the number of papers. However, it is important to note that this systematic review was prepared based on data collected until the end of October 2024, which may explain this apparent drop in published studies (Figure 3).

Regarding the type of evaluation, most of the studies analyzed carried out both objective and subjective assessments. Two of them were restricted to objective evaluations, while only one of them carried out a subjective evaluation (Table 2). Thus, there is a tendency in the field of speech synthesis to value complementarity between types of evaluations, indicating that the isolated use of one type or another usually serves specific purposes. For example, in Bauer *et al.* [2024], the paper intended to evaluate the impact of normal-

izing prosodic features on controllability in speech synthesis, leaving subjective evaluations as future research.

Table 2. Distribution of studies by evaluation type

Evaluation type	Quantity of studies
Only objective	2
Only subjective	1
Objective and subjective	97

In the analyzed studies, 41 TTS models were used for comparative purposes (Figure 4). Some of the studies trained more than one TTS model in their evaluations. Tacotron 2 was the most recurrent model, adopted in 24 studies, followed by FastSpeech 2, used in 22 studies.

4 Prosodic parameters in speech synthesis systems

Based on the studies included in this review, the main prosodic parameters considered in speech synthesis sys-

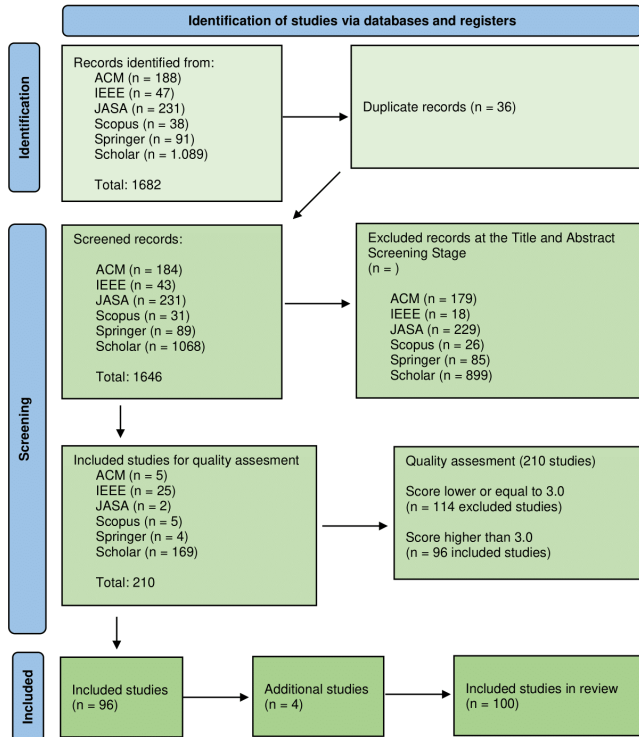


Figure 2. Summary of the identification, screening, and inclusion of studies

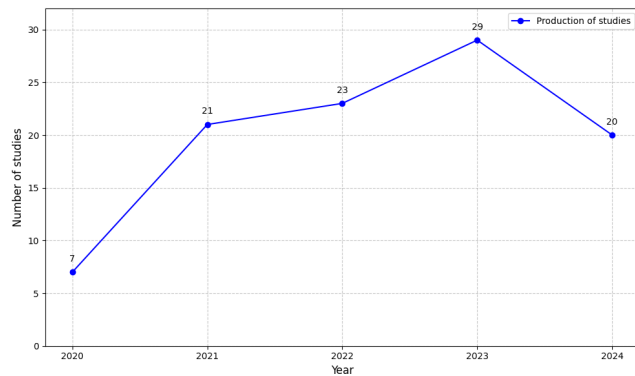


Figure 3. Studies published from 2020 to 2024

tems are fundamental frequency (F0), duration and intensity (Figure 5). Some studies evaluated more than one of these parameters, while others focused exclusively on a single parameter.

The fundamental frequency (F0) corresponds to the physical correlate of vocal fold vibration, determining the perception of a higher pitch (more vibrations) or lower pitch (less vibrations) [Moraes and Rilliard, 2022].

Most of the studies analyzed treat these terms as synonyms. Often, authors do not define the parameters, but it is common for them to alternate between the two terms or use “pitch (fundamental frequency)” or vice versa, suggesting that they do not make a conceptual distinction. Galdino and Oliveira Jr [2023] also observed this practice in studies carried out in Brazil on the relationship between speech synthesis and prosody. There are reasons why F0 is the main parameter for prosodic analysis of synthesis systems: the descending and ascending contours of F0 organize the speech intonation [Lucente, 2020], which directly impacts natural-

ness.

On the other hand, duration is a phenomenon that assigns values and weights to syllables, being important for the application (or not) of certain phonological processes [Cagliari, 1992]. Duration not only assigns values and weights to syllables, but is also related to prominence, in languages such as Portuguese, and marking boundaries of prosodic units, which directly impacts the perception of naturalness of speech and can differentiate natural from synthesized speech Santos *et al.* [2022].

Several approaches have been developed to model duration in TTS systems, but there are two main ones: rule-based methods (phoneme and/or syllable duration is measured using analytical formulas) and data-driven approaches (segment duration is obtained through training machine learning methods [Zangar *et al.*, 2021] Despite advances, duration-based methods often have difficulties in generating natural speech due to strict duration control [He *et al.*, 2022].

Another relevant prosodic parameter, intensity is a phenomenon used as a real physical measure of sound energy [Kent and Read, 2015]. This may explain why intensity is called “energy” in some of the studies in the systematic review. Most of these studies that evaluate intensity seek to improve the emotions produced by synthetic speech, as demonstrated by Xiao *et al.* [2023], Oh *et al.* [2024] and Luo *et al.* [2021]. In the literature, the intensity of anger and happiness usually reach a higher rate, while that of sadness and disgust a lower rate, and a normal rate for fear [Murray and Arnott, 1993]. Although intensity tends to be a less studied feature than duration and F0, it may be a clue that listeners use to identify stressed syllables [Kent and Read, 2015]. Despite the smaller number of studies that focused on this parameter, it is an important resource in the evaluation of TTS, due to both its role in speech and in the identification of emotions.

5 Evaluation of Prosody

5.1 Terminology for evaluation metrics

To evaluate synthetic speech, there are objective metrics, which involve calculations or unbiased analyses of the generated voice, and subjective metrics, which depend on human perception. The analysis of the studies included in this review reveals that emotion evaluation is often referred to as the objective and/or subjective assessment of the “expressiveness” of synthetic speech, as evidenced in studies as Liu *et al.* [2021b], Yang *et al.* [2020] and Kulkarni [2022]. Although the term “expressiveness” is widely used concerning emotion, there seems to be no consistent terminology for metrics used in other speech synthesis challenges. However, several studies evaluate the “quality” of synthetic speech, as observed in Ren *et al.* [2021a] and Jiang *et al.* [2024a]. In prosody research, the analysis of natural voice quality is common, which may explain the use of the term “quality” to refer to irregular voice synthesis, as noted in Mandeel *et al.* [2023b].

In other studies, there is an overlap of terms, such as the evaluation of “speech quality in terms of naturalness” [Bott, 2023; Hida *et al.*, 2022]. The term “naturalness” was more

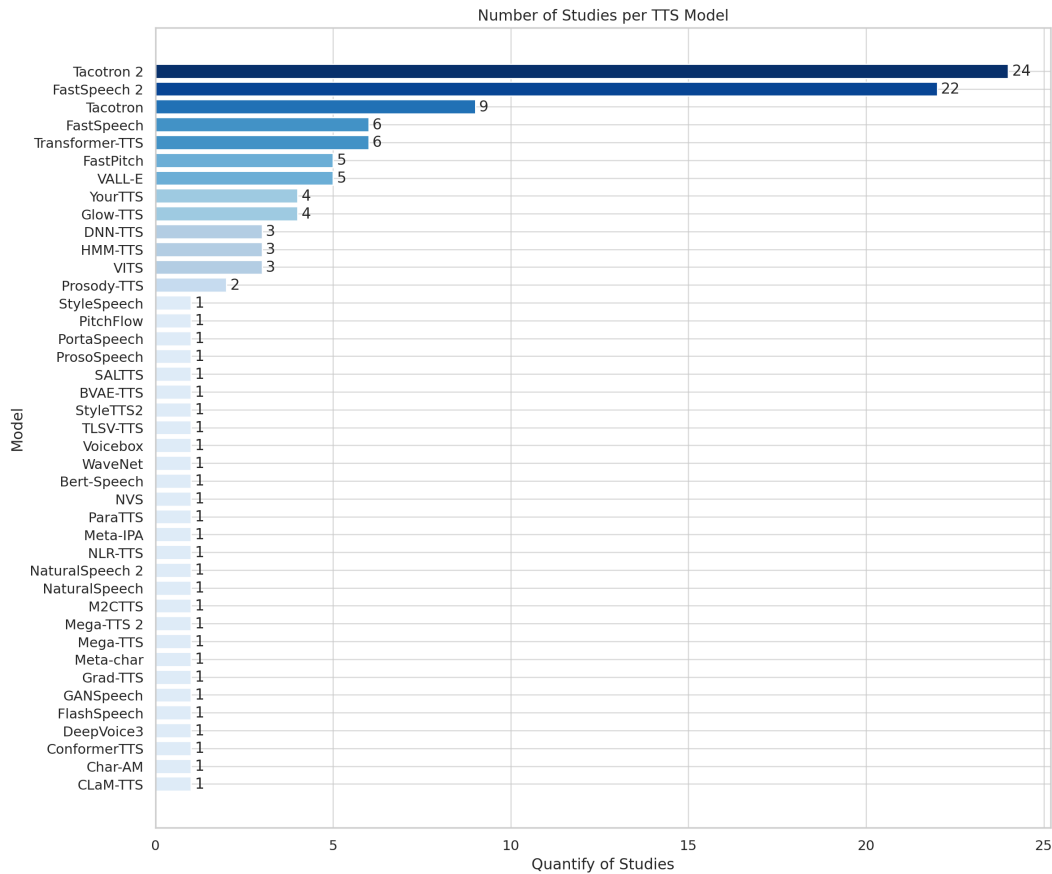


Figure 4. List of TTS models used in the studies

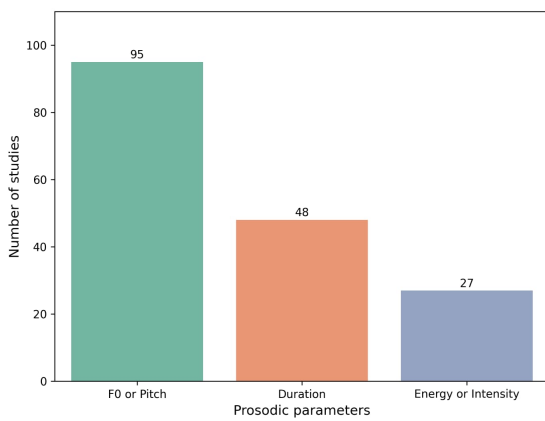


Figure 5. Distribution of studies by prosodic parameters.

frequently used in subjective metrics than in objective ones [Fujimaki et al., 2020; Zhang et al., 2021]. This may suggest that the perception of naturalness depends heavily on participants judgments. However, some studies have also applied the term in objective metrics, such as the calculation of Mel-Cepstral Distortion (MCD) to “objectively measure naturalness” [Xue et al., 2022].

According to Sini [2020], most subjective evaluations of speech synthesis systems focus on naturalness and intelligibility, while similarity has been adopted as an additional evaluation dimension due to the advancement of speaker adap-

tation and voice conversion techniques. Similarity was also a recurring term in the analyzed studies, particularly in the analysis of “speaker similarity” [Ogun et al., 2023]. In other works, specific metrics followed the terminology of evaluations such as QMOS (Quality, clarity, naturalness, and high-frequency details) and SMOS (Speaker similarity in terms of timbre reconstruction and prosodic pattern)[Jiang et al., 2024c].

Table 3 presents the categorization of commonly used evaluation metrics, grouped into four categories: Acoustic, Intelligibility, Naturalness/Quality and Similarity. Acoustic metrics are widely used to assess the fidelity and quality of synthesized speech. The F0-RMSE (Root Mean Square Error) is one of the most commonly used metrics to evaluate pitch. Several authors, including [Matsubara et al., 2023], Liu et al. [2021b], Wu et al. [2021a] and Zhou et al. [2024], have adopted this metric. The Mel-Cepstral Distortion (MCD), used by Wang et al. [2022b] and Tan et al. [2021], measures spectral distortions and is important for analyzing the quality of speech. The Gross Pitch Error (GPE) and F0 Frame Error (FFE) employed by Wang et al. [2022b] and Zhang et al. [2021] are used to identify pitch errors. There is a consistent use of F0-based metrics (such as F0-RMSE and GPE) in recent studies, indicating a clear focus on improving the pitch accuracy and expressiveness of synthesized speech.

For intelligibility metrics which measure how understandable synthetic speech is to listener, objective metrics such as Voice/Unvoiced Decision Error (V/UV), used by Mat-

subara *et al.* [2023], and F1-Score, adopted by Zou *et al.* [2021], quantitatively assess the accuracy of voicing decisions and classification performance. Furthermore, Pause-RMSE, used by Xue *et al.* [2022], objectively assesses the timing and placement of pause, which affects fluency. The analysis of intelligibility often blends both objective and subjective measures to provide a comprehensive assessment.

Evaluating the naturalness and quality of speech synthesis involves both subjective and objective approaches. MOS is widely used in studies like Mandeel *et al.* [2023b] and Wu *et al.* [2021a] and a subjective metric where human listeners rate the naturalness of speech samples. Similarly, MUSHRA, employed by Lameris *et al.* [2023], is a comparative subjective test designed to assess audio quality. Notably, while naturalness is often measured subjectively, some studies, like Xue *et al.* [2022], apply objective metrics such as MCD to quantify naturalness, illustrating an intersection between objective analysis and perceptual assessment.

For the similarity metrics, subjective tests like the AB Test and ABX Blind Test, used by Xin *et al.* [2023] and Wang *et al.* [2022b], involve human listeners comparing pairs of audio samples to judge similarity. Objective metrics such as Cosine similarity, applied by Sadekova *et al.* [2024], quantify the similarity between acoustic feature representations. Studies like Jiang *et al.* [2024c] also introduce integrated evaluations like QMOS and SMOS, combining quality and speaker similarity assessments.

The literature reveals that objective metrics are predominantly used in evaluating acoustic and intelligibility aspects, due to their reproducibility and precision. In contrast, subjective metrics are more common in evaluating naturalness and similarity, reflecting the importance of human perception in these dimensions. However, there is a growing trend to combine the two types of evaluation to provide more comprehensive and balanced assessments of speech synthesis.

5.2 Methodologies to evaluate prosody

The studies included in this systematic literature review did not explicitly use the terms “manual”, “automatic”, or “hybrid” to describe the methodologies employed for evaluating synthetic speech. Instead, the methodologies can be distinguished by the involvement or absence of human listening in the evaluating process, which indicates whether the assessments were “objective” or “subjective”. Table 4 presents the most commonly used metrics across the reviewed studies.

To evaluate prosody, the most commonly used objective metric was the Root Mean Square Error (RMSE) of the fundamental frequency (F0). This metric quantifies how closely the pitch contour of synthesized speech aligns with that of natural speech, with lower values indicating greater pitch accuracy [Ma *et al.*, 2024]. In some cases, prosodic features such as energy and duration were also measured using RMSE, along with additional metrics like Gross Pitch Error (GPE) and F0 Frame Error (FFE). These objective metrics are widely adopted due to their ability to automate the evaluation process, making analysis faster and more scalable. However, they also have limitations, particularly in capturing finer details of prosody, such as F0 variations and statistics like minimum, average, and maximum pitch values.

In contrast, the most frequently used subjective metrics were the Mean Opinion Score (MOS) [Mandeel *et al.*, 2023b; Wang *et al.*, 2022b; Al-Radhi *et al.*, 2023; Matsubara *et al.*, 2023], the AB Test [Xin *et al.*, 2023; Zhou *et al.*, 2024; Yang *et al.*, 2022; Yufune *et al.*, 2021], and the MUSHRA test [Mandeel *et al.*, 2023a; Mohan *et al.*, 2021; Prateek, 2023]. These methods rely on human listeners to assess the naturalness, quality, and expressiveness of synthetic speech. In particular, few studies explicitly stated that these subjective tests were designed specifically for prosody evaluation.

This is understandable, as prosodic parameters like pitch (F0), duration, and intensity are typically analyzed using acoustic software tools (e.g. Praat) rather than relying solely on human perception. Although subjective metrics require human listeners to evaluate the acceptability of synthetic speech, assessing prosody in detail would require the involvement of prosody experts specialized in the analysis of prosodic features.

The widespread use of F0-RMSE [Matsubara *et al.*, 2023; Liu *et al.*, 2021b; Yoneyama *et al.*, 2023] reflects the importance placed on pitch accuracy in prosody evaluation. Despite its strengths in automation, this metric struggles in capturing complex prosodic variations. For example, these metrics are blind to parts of the sentence where f0 deviations between natural and synthetic speech can occur. Deviations at sensitive points, such as points corresponding to stressed syllables or syllables produced in focus contexts, and prosodic boundaries, are probably more problematic than at other points in the sentence.

Similarly, the analysis of duration and intensity benefits from automation, but presents challenges when considering language-specific prosodic patterns. For example, in some languages, the duration of the syllables is organized systematically at the syllable level [Cagliari, 1992]. This opens opportunities to improve synthetic speech evaluation by analyzing the duration of syllables, particularly because some segments exhibit unique temporal characteristics. In Brazilian Portuguese, for example, the duration of syllables is one of the acoustic parameters that signals the position of stressed syllables, which tend to be longer in duration [Moraes, 1987]. Not only does accent affect the duration of segments, but, in fact, segments have specific duration characteristics, independent of prosodic properties. For example, in Brazilian Portuguese, nasal vowels are always longer than oral vowels, and open vowels are longer than closed vowels.

The data summarized in Table 4 further illustrates the predominance of certain metrics across the reviewed studies. The MOS was the most frequently employed metric, appearing in 73 studies, underscoring its relevance for subjective evaluation. The F0-RMSE was the most used objective metric, appearing in 40 studies, reflecting the priority given to pitch accuracy in prosody evaluation. Subjective metrics such as the AB Test and the MUSHRA test were used in 20 and 7 studies, respectively, confirming the importance of listener-based evaluations in assessing naturalness and expressiveness. On the other hand, metrics like Mel-Cepstral Distortion (MCD) (14 studies), Gross Pitch Error (GPE) (9 studies), and F0 Frame Error (FFE) (6 studies) were also prevalent, supporting a comprehensive evaluation of spectral and pitch-related prosodic features.

Table 3. List of evaluation metrics categorized by groups

Metric	Group	Studies
Creakiness percentage	Acoustics	Mandeeel <i>et al.</i> [2023b]
Harmonic-to-Noise Ratio (HNR)	Acoustics	Mandeeel <i>et al.</i> [2023b]
F0 RMSE	Acoustics	Matsubara <i>et al.</i> [2023], Liu <i>et al.</i> [2021b], Yoneyama <i>et al.</i> [2023], Wu <i>et al.</i> [2021a], Wu <i>et al.</i> [2021b], Xin <i>et al.</i> [2023], Lei <i>et al.</i> [2023], Fujimaki <i>et al.</i> [2020], Guo <i>et al.</i> [2023], Deng <i>et al.</i> [2024], Zhou <i>et al.</i> [2024], Wang <i>et al.</i> [2022a], Zhang <i>et al.</i> [2023b], Gong <i>et al.</i> [2021], Malviya <i>et al.</i> [2023], Yufune <i>et al.</i> [2021], Aso <i>et al.</i> [2020], Fujii <i>et al.</i> [2022], Kulkarni [2022], Liu <i>et al.</i> [2022], Liu <i>et al.</i> [2024], Zhang <i>et al.</i> [2023c], Liu <i>et al.</i> [2023], Li <i>et al.</i> [2023], Hodari [2022], Zhang <i>et al.</i> [2023a], The Nguyen <i>et al.</i> [2023]
F0 Frame Error (FFE)	Acoustics	Wang <i>et al.</i> [2022b], Zhang <i>et al.</i> [2021], Hamed and Lachiri [2024a], Bak <i>et al.</i> [2021], Hamed and Lachiri [2024b]
Gross Pitch Error (GPE)	Acoustics	Wang <i>et al.</i> [2022b], Xin <i>et al.</i> [2023], Zhang <i>et al.</i> [2021], Hamed and Lachiri [2024a], Hamed and Lachiri [2024b]
Mel-Cepstral Distortion (MCD)	Acoustics	Wang <i>et al.</i> [2022b], Tan <i>et al.</i> [2021]
Duration RMSE	Acoustics	Zhou <i>et al.</i> [2024], Zangar <i>et al.</i> [2021], Liu <i>et al.</i> [2024]
Mean Absolute Error (MAE)	Acoustics	Zangar <i>et al.</i> [2021], He <i>et al.</i> [2022], Pamisetty and Sri Rama Murty [2023], Ren <i>et al.</i> [2021a]
F0 voiced error (FVE)	Acoustics	Moon <i>et al.</i> [2022]
Voicing decision error (VDE)	Acoustics	Pamisetty and Sri Rama Murty [2023]
Frequency Weighted Segmental SNR (FwsNRseg)	Acoustics	Mandeeel <i>et al.</i> [2023a]
Pitch RMSE	Acoustics	Oh <i>et al.</i> [2024]
Dynamic Time Warping (DTW)	Acoustics	Jiang <i>et al.</i> [2022], Turkmen [2021], Zhang <i>et al.</i> [2023b], Jiang <i>et al.</i> [2024c], Bak <i>et al.</i> [2024], Zhang <i>et al.</i> [2023a]
Perceptual Evaluation of Speech Quality (PESQ)	Acoustics	Liu <i>et al.</i> [2021a]
Periodicity error (RMSEperiod)	Acoustics	Oh <i>et al.</i> [2024]
Kullback–Leibler (KL) divergence	Acoustics	Oh <i>et al.</i> [2024]
Wasserstein Distance	Acoustics	Wu <i>et al.</i> [2022]
Energy Distance	Acoustics	Wu <i>et al.</i> [2022]
CREPE	Acoustics	Bauer <i>et al.</i> [2024]
Legendre Polynomial (LP) decomposition	Acoustics	O’Mahony <i>et al.</i> [2023]
Mean value of pitch	Acoustics	Ju <i>et al.</i> [2022]
Distribution of log-F0 values	Acoustics	Ogun <i>et al.</i> [2023]
Voice/Unvoiced (V/UV) Decision Error	Intelligibility	Matsubara <i>et al.</i> [2023]
Unvoiced/voiced (U/V) errors of F0	Intelligibility	Yasuda [2021]
Pause RMSE	Intelligibility	Xue <i>et al.</i> [2022]
F1 score	Intelligibility	Zou <i>et al.</i> [2021]
Goodness of fit(R2)	Intelligibility	Jiang <i>et al.</i> [2024a]
Pearson correlation coefficient	Intelligibility	Peng and Ling [2022], Yang <i>et al.</i> [2020], Bak <i>et al.</i> [2024], Xue <i>et al.</i> [2022]
MOS	Naturalness/Quality	Mandeeel <i>et al.</i> [2023b], Wang <i>et al.</i> [2022b], Al-Radhi <i>et al.</i> [2023], Matsubara <i>et al.</i> [2023], Liu <i>et al.</i> [2021b], Yoneyama <i>et al.</i> [2023], Wu <i>et al.</i> [2021a], Wu <i>et al.</i> [2021b], Lei <i>et al.</i> [2023], Xin <i>et al.</i> [2023], Fujimaki <i>et al.</i> [2020], Inoue <i>et al.</i> [2024], Guo <i>et al.</i> [2023], Mandeeel <i>et al.</i> [2023a], Yanagita <i>et al.</i> [2023], Yasuda [2021], Lenglet <i>et al.</i> [2023], Jiang <i>et al.</i> [2024c], Moon <i>et al.</i> [2022], Bak <i>et al.</i> [2024], He <i>et al.</i> [2022], Xue <i>et al.</i> [2022], Liu <i>et al.</i> [2024], Kurihara [2024], Sadekova <i>et al.</i> [2024], Ren <i>et al.</i> [2021b], Lameris <i>et al.</i> [2023], Zhang <i>et al.</i> [2023c], Pamisetty and Sri Rama Murty [2023], Huang <i>et al.</i> [2023], Jiang <i>et al.</i> [2024b], Chen <i>et al.</i> [2021], Liu <i>et al.</i> [2023], Li <i>et al.</i> [2024b], Li <i>et al.</i> [2024a], Zhao <i>et al.</i> [2021], Ogun <i>et al.</i> [2023], Li <i>et al.</i> [2023], Hodari [2022], O’Mahony <i>et al.</i> [2023], Zhang <i>et al.</i> [2023a], Kaiki <i>et al.</i> [2021], The Nguyen <i>et al.</i> [2023], Kumar <i>et al.</i> [2022], Xu <i>et al.</i> [2024], Raitio <i>et al.</i> [2020], Ju <i>et al.</i> [2022], Łańcucki [2021], Ren <i>et al.</i> [2021a]
MUSHRA	Naturalness/Quality	Mandeeel <i>et al.</i> [2023a], Mohan <i>et al.</i> [2021], Prateek [2023], Hamed and Lachiri [2024b], Lameris <i>et al.</i> [2023]
MOSNET	Naturalness/Quality	Wang <i>et al.</i> [2022b],
DMOS	Naturalness/Quality	Li <i>et al.</i> [2022]
CMOS	Naturalness/Quality	Ye <i>et al.</i> [2024]
SMOS	Naturalness/Quality	Ye <i>et al.</i> [2024], Jiang <i>et al.</i> [2024c], Kumar <i>et al.</i> [2022],
UTMOS	Naturalness/Quality	Ye <i>et al.</i> [2024]
N-MOS	Naturalness/Quality	Bak <i>et al.</i> [2024], Li <i>et al.</i> [2024b], Li <i>et al.</i> [2024a], Ogun <i>et al.</i> [2023]
Perceptual-based measure	Naturalness/Quality	Yanagita <i>et al.</i> [2023]
Similarity-MOS (S-MOS)	Naturalness/Quality	Bak <i>et al.</i> [2024], Li <i>et al.</i> [2024a], Ogun <i>et al.</i> [2023]
Prosody Similarity-MOS (PS-MOS)	Naturalness/Quality	Bak <i>et al.</i> [2024],
ABX blind test	Similarity	Wang <i>et al.</i> [2022b]
AB test	Similarity	Xin <i>et al.</i> [2023], Zhou <i>et al.</i> [2024], Yang <i>et al.</i> [2022], Yufune <i>et al.</i> [2021], Aso <i>et al.</i> [2020], Fujii <i>et al.</i> [2022], Törö [2022], Yang <i>et al.</i> [2020], Zou <i>et al.</i> [2021], Zhang <i>et al.</i> [2023c], Jiang <i>et al.</i> [2024b], Kaiki <i>et al.</i> [2021], Liu <i>et al.</i> [2021c]
XAB test	Similarity	Yufune <i>et al.</i> [2021]
AXY	Similarity	Chien and Lee [2021], Chien and Lee [2021]
Cosine similarity	Similarity	Sadekova <i>et al.</i> [2024]
Wavelet Prosody Toolkit (WPT)	Similarity	Lameris <i>et al.</i> [2023]
Mean pitch distance	Similarity	Lee <i>et al.</i> [2022]

The methodologies for evaluating prosody in synthetic speech predominantly rely on a balance between objective and subjective metrics. Objective evaluations, particularly those measuring F0-RMSE, offer scalable and efficient assessments but may lack sensitivity to prosodic nuances. Subjective evaluations, including MOS and AB Tests, capture human perception but are resource-intensive and potentially inconsistent sometimes.

6 Corpora and languages evaluated

6.1 Corpora, transcription process, and quality of the recordings

The analysis of the studies included in this systematic review revealed significant variability in how datasets (corpora) are described, particularly regarding the type of speech used for training and evaluation, whether read or spontaneous speech.

Table 4. Metrics most commonly used in studies

Metric	Quantity
MOS (Mean Opinion Score)	73
F0 (Pitch) RMSE (Root Mean Square Error)	40
AB Test	20
MCD (Mel-Cepstral Distortion)	14
GPE (Gross Pitch Error)	9
Duration MSE (Mean Square Error)	5
MUSHRA	7
FFE (F0 Frame Error)	6
Duration and Energy MAE	4
DTW (Dynamic Time Warping)	4
F0 CORR (F0 Correlation)	4
Duration RMSE	3
V/UV (Voice/Unvoiced)	3
Duration Error	2
EN (Energy) RMSE	2
Goodness of fit	2
Perceptual-based measure	2

Additionally, details about the transcription process, the quality of the datasets employed, and the total number of hours applied were often not explicitly provided. However, in cases where datasets are publicly available, some of this missing information can be verified through dataset documentation on their respective websites. To address this, we compiled the most frequently used public datasets in the analyzed studies (**Table 5**). This section focuses on the types of datasets used in the reviewed studies, their transcription processes, and the quality of audio recordings.

Out of the analyzed studies, 56 exclusively utilized public datasets, 26 relied solely on internal datasets, and 18 employed a combination of both (Figure 6 and **Table A**, in Appendix A). Public datasets were more commonly used for pretraining due to their large scale and accessibility, while internal datasets were often preferred for fine-tuning because of their specialized and task-specific content.

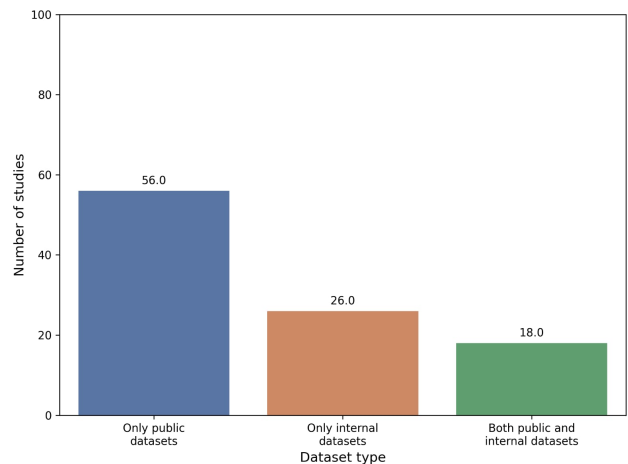
For example, Li *et al.* [2024b] used a 10,000-hour public speech corpus for pretraining and fine-tuned the model with a 5.4-hour internal dataset. Although the total amount of data was not always reported in hours, some studies provided the number of sentences used during training and testing, offering indirect insight into dataset size.

Most publicly available datasets consisted of read speech, whereas spontaneous or conversational speech was less commonly used and typically appeared in emotional or acted speech datasets. Internal datasets, on the other hand, frequently contained more spontaneous or acted speech tailored to specific research goals. For instance, one internal Mandarin dataset featured 16.7 hours of spontaneous speech.

Transcription methods varied considerably across datasets. In many cases, studies did not detail the transcription process, possibly assuming that such information was only necessary when directly relevant to the study’s objectives. However, when pertinent, transcription details were included. For example, Mandeel *et al.* [2023a] provided information on punctuation in transcriptions to support their analysis of interrogative sentences. Some datasets offered high-quality manual transcriptions [Wu *et al.*, 2022; Fujimaki *et al.*, 2020;

Wu *et al.*, 2021a] while others used automatic speech recognition (ASR) models for alignment and segmentation [Furukawa, 2022; O’Mahony *et al.*, 2023; Raitio *et al.*, 2020]. Notable examples include LibriLight, which employed a hybrid DNN-HMM ASR model pre-trained on 960 hours of LibriSpeech for phoneme alignment [Jiang *et al.*, 2024c], and SynPaFlex-Corpus, which was automatically annotated with phonetic labels, word boundaries, and morphosyntactic tags [Sini, 2020]. Similarly, CanTTS3 incorporated punctuation in its transcriptions to distinguish between statements and questions [Bai *et al.*, 2022].

In contrast, public datasets such as VCTK and CMU-ARTIC often lacked detailed descriptions of their transcription procedures. Moderate or unspecified quality datasets included Aishell-3 for Chinese language [Tan *et al.*, 2021]. Private datasets typically offered higher audio fidelity due to controlled recording environments. For instance, one internal studio-quality Mandarin dataset contained 251 hours of speech [Guo *et al.*, 2023], while another internal French dataset comprised 600 hours of high-quality read speech [Sini, 2020]. The studies highlight the predominance of read speech datasets and varied audio quality standards. While public datasets offer accessibility and scale, internal datasets provide tailored, often higher-quality data for specialized tasks.

**Figure 6.** Distribution of studies by dataset type

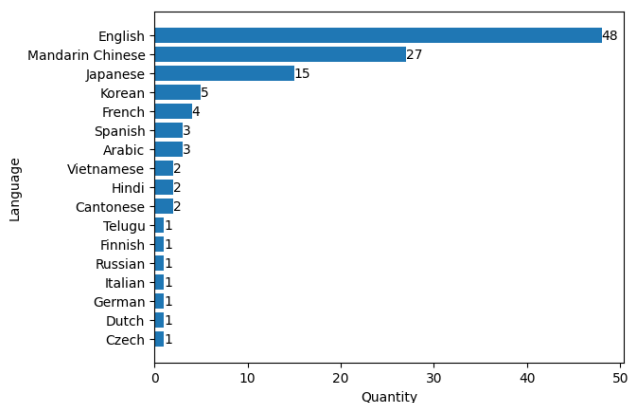
6.2 Languages

The predominant language in studies that evaluated speech synthesis prosody was English, with 48 studies (**Figure 7**). There was also a considerable amount of studies in Mandarin Chinese, (27 studies) and Japanese, with 15 studies. In smaller numbers, studies investigated Arabic (3), Cantonese (2), Czech (1), Dutch (1), French (4), German (1), Hindi (2), Italian (1), Korean (5), Russian (1), Spanish (3), Finnish (1), Telugu (1), and Vietnamese (2). Some works studied more than one language.

Although English already has a large number of large-scale datasets, it continues to be a predominant language in studies on speech synthesis systems. This indicates that other less explored languages, after facing the challenge of having

Table 5. List of the most commonly used public datasets in the studies reviewed

Dataset	Quantity	Speech type	Hours	Language
BiaoBei Dataset	4	Read speech and Acted speech	-	Mandarin Chinese
Blizzard Challenge	9	Read speech	Varies by year	Varies by year
CMU-ARCTIC	3	Read speech	-	Variety of English accents
Emotion Speech Database	4	Read speech	29	English and Mandarin Chinese
Hi-Fi Multi-Speaker	2	Read speech	292	English
IEMOCAP	2	Acted speech	12	American English
JSUT	9	Read speech	10	Japanese
LibriLight	2	Read speech	60 K	American English
LibriTTS	10	Read speech	585	American English
LJSpeech	21	Read speech	24	American English
Mandarin Chinese Read Speech Corpus	2	Read speech	775	Variety of Mandarin Chinese Accents
VCTK	14	Read speech	44	Variety of English accents
SynPaFlex-Corpus	2	Read speech	87	French

**Figure 7.** Distribution of languages by studies

a large-scale and good quality dataset, will still have to deal with the complexity of enriching the prosody of synthesized speech. This is relevant, considering that studies still seek to improve English prosody, despite the large availability of data in this language.

7 Speech synthesis models used in comparisons

Some studies included in this systematic review carried out comparisons between different models. However, most studies aimed to propose methods, techniques or improvements to existing model architectures, seeking to improve prosody or enrich the speech synthesis task as a whole. Therefore, it is common for studies to conclude that the proposed models have greater potential than the initial baseline model.

Although this tendency shows a limitation in directly answering our sixth research question (Do studies make comparisons between different speech synthesis models? If so, which models have been most used and which have the greatest potential?), the selected studies allowed us to identify the

models that were most used for comparisons, as well as the challenges faced by these models in generating prosodic characteristics. The comparison involves both traditional parametric models and modern end-to-end neural network approaches. The main models highlighted were Tacotron 2 (24), FastSpeech 2 (22), Tacotron (9), Transformer-TTS (6), FastSpeech (6), FastPitch (5), VALL-E (5), YourTTS (4), GloW-TTS (4) and VITS (3) (see Figure 4).

The most commonly used speech synthesis model identified in the reviewed studies is FastSpeech 2, which is frequently employed as a baseline due to its efficiency and high-quality output [Turkmen, 2021; Wu *et al.*, 2021a; Oh *et al.*, 2024; Bott, 2023]. Researchers have explored various enhancements to FastSpeech 2, such as integrating prosody predictors and hierarchical prosody modeling, to address its limitations and expand its capabilities.

Another widely used baseline is Tacotron 2, known for its ability to produce natural and expressive speech. This model is frequently compared with proposed systems in studies aimed at improving prosody representation and expressiveness, serving as a benchmark for evaluating advancements.

In addition to these, HiFi-GAN and WaveNet are prominent neural vocoders commonly used in conjunction with text-to-speech systems to enhance the quality of synthesized speech [Matsubara *et al.*, 2023]. These vocoders play a critical role in improving the naturalness and clarity of output by refining the generated waveforms. Finally, Transformer-based models, such as TransformerTTS and its variants, have been evaluated for their ability to capture fine-grained prosodic details [Wang *et al.*, 2022a; Liu *et al.*, 2021c; Ren *et al.*, 2021a]. These models leverage the transformer architecture to better represent prosodic nuances, which are essential to produce expressive and natural-sounding speech.

Several studies have proposed models that outperform baseline systems in specific tasks, showcasing advancements in synthesis quality, efficiency, and prosody control. For in-

stance, Harmonic-Net [Matsubara *et al.*, 2023] demonstrated superior synthesis quality, real-time performance, and enhanced controllability of prosodic features. Similarly, DiffProsody [Oh *et al.*, 2024] utilized diffusion-based adversarial training to achieve faster and higher-quality prosody generation. Another innovative model, Ctrl-P [Mohan *et al.*, 2021], offered precise and disentangled control over prosodic features, significantly improving the naturalness of synthesized speech. Additionally, models like FlashSpeech [Ye *et al.*, 2024] stood out for their efficiency in zero-shot synthesis, combining high speaker similarity and synthesis speed.

In the domain of enhanced prosody control, ProsoSpeech and Prosody-TTS [Huang *et al.*, 2023] introduced advanced methods, such as masked autoencoders, to improve prosody representation. Quasi-Periodic Models [Wu *et al.*, 2021a] enhanced pitch controllability by leveraging quasi-periodic convolutional networks.

Some models have also addressed challenges in multilingual synthesis and low-resource scenarios. For example, the Meta-Learning-Based Multilingual Model [Peng and Ling, 2022] achieved significant improvements in multilingual synthesis by addressing unique linguistic challenges. Additionally, zero-shot and few-shot techniques, as demonstrated by models like YourTTS and MultiVerse [Bak *et al.*, 2024], achieved high-quality speech synthesis with notable prosody similarity, even with minimal training data.

The Tacotron 2 architecture has a state-of-the-art TTS performance, but lacks clear modeling of prosodic attributes, which makes accurate representation of accents (linguistic variations) difficult [Zhou *et al.*, 2024]. FastSpeech 2 can extract duration, pitch and energy from audio data during training, and use predicted values in inference [Wu *et al.*, 2022]. FastPitch can generate pitch-manipulated speech, but pitch expressiveness and speaker similarity may decrease due to its decoder being prone to learning the relationship between text and pitch [Bak *et al.*, 2021]. In turn, YourTTS typically performs well on speaker similarity, but faces the challenge of still generating a systematic average pitch contour [Huang *et al.*, 2023].

8 Limitations, Challenges and Conclusions

8.1 Limitations and Challenges

The primary limitations identified in the studies include data scarcity and annotation difficulties.

Additional constraints were related to the complexity of tonal and multilingual language processing and the integration of prosody with other components of speech synthesis.

To improve the prosody of speech generated by speech synthesis models, several challenges need to be faced, as this task is not completely solved. Hodari [2022] highlights three major challenges. Synthesizing prosody is difficult because it is incorporated into the speech signal, which makes its manipulation complex. Additionally, there is no clear spelling for prosody, which means it is underspecified in the input text, making direct control difficult. Ultimately, prosody is

determined by the context of a speech act, which TTS systems do not have complete access to.

An analysis of the reviewed studies highlights a wide range of challenges that extend beyond the core task of speech synthesis. Among these, certain gaps appear recurrently, with the most prominent being the lack of prosodic variations in synthesized speech. This limitation is particularly problematic in contexts requiring significant variation, such as audiobook narration [Xin *et al.*, 2023] and maintaining coherent prosody in long sentences or extended utterances [Guo *et al.*, 2023]. Similarly, the inability of TTS systems to effectively generate prosodic variation in spontaneous speech make the problem more challenge, limiting the ability to replicate natural conversational dynamics [Li *et al.*, 2024a]. This lack of variation can result in inappropriate prosodies, potentially causing misunderstandings [Zou *et al.*, 2021].

One of the causes of this problem is the dependence on textual input to define prosody. Many models rely exclusively on text, leading to monotonous or averaged prosodic contours that fail to reflect desired expressiveness or emotional nuances [Raitio *et al.*, 2020; Zou *et al.*, 2021]. Furthermore, most current models lack explicit and interpretable prosody controls, making it difficult for users to adjust speech delivery to suit specific contexts or emotions [Takumi *et al.*, 2024; O'Mahony *et al.*, 2024].

The challenge of insufficient variability is compounded by the oversimplified representations of prosody in many models, which reduce it to single variables, such as mean pitch, failing to capture its multimodal and dynamic nature [Huang *et al.*, 2023; Mohan *et al.*, 2021]. Additionally, there is a persistent trade-off between compactness and richness in prosodic representations, as achieving both efficient computation and nuanced expressiveness remains difficult [Zhang *et al.*, 2021].

Another critical issue is the lack of contextual integration in prosody modeling. Many systems overlook cross-utterance dependencies, which reduces the naturalness and coherence of synthesized speech in long-form or multi-sentence contexts [Wu *et al.*, 2022; Xiao *et al.*, 2023]. This challenge is particularly relevant for spontaneous speech, where handling disfluencies and spontaneity, such as fillers, interruptions, and variations in speech rhythm is crucial for achieving naturalness [Lameris *et al.*, 2023; Li *et al.*, 2024b]. Moreover, the integration of syntactic and semantic context into prosody modeling is still underexplored, despite its importance for generating coherent and expressive speech [Liu *et al.*, 2021c; Furukawa, 2022].

A consequence of the challenge in reproducing prosodic variety is the limitation in expressing emotions in synthesized speech. One of the challenges in expressive speech synthesis is that different prosodization possibilities, related to different styles, can be associated with the same sequence of phonemes in expressive speech Wu *et al.* [2022]. Some methods still have difficulty explicitly incorporating expressivity, such as autoregressive speech synthesis models [Hamed and Lachiri, 2024a], non-autoregressive neural techniques [Turkmen, 2021], and newer TTS models combined with neural vocoders [Lenglet *et al.*, 2023]. In order to solve this problem, initiatives seek to propose methods that can predict and

control emotional speech [Inoue *et al.*, 2024].

A significant challenge in text-to-speech (TTS) synthesis is the scarcity of high-quality speech data required for training. Generating realistic prosody often requires large, high-quality datasets, which are resource-intensive to create and annotate [Liu *et al.*, 2021a; Peng and Ling, 2022]. The manual annotation of prosodic features is costly, while automatic methods frequently introduce errors, compromising the synthesis quality [Fujimaki *et al.*, 2020].

This data scarcity is especially problematic for low-resource languages and dialects, where limited datasets hinder accurate prosody modeling [Yufune *et al.*, 2021; Zhou *et al.*, 2024]. For example, generating natural-sounding prosody in resource-scarce languages like Arabic remains a challenge due to the lack of sufficient data [Al-Radhi *et al.*, 2023]. Similarly, multilingual TTS systems face difficulties because they require extensive speech corpora for all supported languages, which are often unavailable [Peng and Ling, 2022]. This issue extends to the generation of speech in different linguistic varieties of a language, where limited data availability restricts the diversity of synthesized outputs [Zhou *et al.*, 2024; Zhang *et al.*, 2023b].

Finally, each language has its own prosodic peculiarities. Therefore, the prosody and controllability of synthesized speech are still insufficient, especially in tonal languages such as Mandarin [Gong *et al.*, 2021]. Prosodic modeling of these languages becomes difficult, due to prosodic changes in the “cadence” (or modulation) of [Liu *et al.*, 2021a] sentences. Declarative questions in Cantonese are asked with rising intonation, but some neural TTS systems are not capable of synthesizing this type of intonation [Bai *et al.*, 2022]. Furthermore, the particularities of each language lead to other problems, such as the need for professional annotators who are familiar with the language’s tonal system, for example, which also makes it difficult to create linguistic varieties of this type of language [Yufune *et al.*, 2021].

8.2 Comparison between the prosody of synthetic speech and natural speech: conclusions of the studies

Most studies aimed to propose methods and techniques to improve the architecture of the synthesis model. Thus, the results of the studies indicated that there was an improvement in the prosody of voices generated by speech synthesis.

Despite this uniformity in conclusions, studies point out some gaps that separate the prosody of synthetic speech from natural speech. For example, Inoue *et al.* [2024] report that some emotions, such as “anger” and “happiness” showed a close proximity to natural speech, while “sadness” had a negative correlation. These inconsistencies demonstrate the need to more precisely control prosody related to emotions expressed in speech, simulating natural speech patterns, which may be possible through the use of pre-trained language models such as HuBERT [Oh *et al.*, 2024].

Although progress has been made in aligning prosody with dialogue context or paragraph reading, such as in ContextSpeech, synthetic speech still struggles with dynamic emotional prosody and fine-grained adaptations in a variety

of linguistic contexts [Xiao *et al.*, 2023; Zhang *et al.*, 2023c]. More sophisticated techniques are needed to capture dynamic emotional shifts and spontaneous prosody, especially in contexts requiring high emotional variability [Huang *et al.*, 2023; Hamed and Lachiri, 2024a].

This control can be performed at the phoneme level on parameters such as F0, duration and intensity, increasing the naturalness of speech, but it may be better to provide more abstract controls such as “emphasize this word” or “create rising intonation associated with question” [Mohan *et al.*, 2021]. At the same time, this type of control can present challenges, especially for tonal languages such as Japanese, where emphasis is typically conveyed through increasing F0 [Takumi *et al.*, 2024].

For tonal languages such as Mandarin and Japanese, advances in pitch and duration control, as demonstrated by models like FastMandarin and the approach proposed by Yanagita *et al.* [2023], have improved prosodic alignment. Despite these improvements, these systems still struggle to replicate the nuanced tonal patterns and precise timing variations inherent in natural speech [Jiang *et al.*, 2024a; Yanagita *et al.*, 2023]. Similarly, approaches like Decoupled Pronunciation and Prosody Modeling have enhanced prosody in multilingual contexts, but they continue to fall short in reproducing the richness of natural prosody, particularly in low-resource languages and dialects [Peng and Ling, 2022; Zhou *et al.*, 2024].

Many models have significantly improved pitch accuracy, rhythm, and overall naturalness, narrowing the gap between synthetic and natural prosody. For example, Harmonic-Net provides improved control over prosodic parameters, closely mirroring the natural variations in human speech [Matsubara *et al.*, 2023]. Similarly, models like SiFi-GAN and FastPitch-Formant enhance pitch controllability and robustness, bringing synthetic speech closer to natural prosody [Yoneyama *et al.*, 2023; Bak *et al.*, 2021]. Techniques like hierarchical modeling and discourse-level linguistic features have enhanced the coherence and expressiveness of long-form synthetic speech, though fine intonation control in extended contexts remains superior in natural speech [Guo *et al.*, 2023; Wu *et al.*, 2022].

Synthetic speech often lacks the nuanced variability and spontaneous shift characteristic of natural prosody. For instance, models like FlashSpeech and StyleTTS 2 achieve high audio quality and diverse prosody but still fall short in capturing subtle context-driven variations [Ye *et al.*, 2024; Li *et al.*, 2023]. Models such as Quasi-Periodic WaveNet and CampNet achieve lower pitch-related errors (e.g., RMSE), better pitch alignment, and improved temporal accuracy [Wu *et al.*, 2021b; Wang *et al.*, 2022a]. Studies show that while synthetic speech can achieve high scores for naturalness and intelligibility, it often falls short in speaker likeness and perceived emotional depth. For instance, synthetic sentences with creaky voice patterns were judged lower in speaker likeness and overall naturalness compared to natural speech [Mandee *et al.*, 2023b].

Significant advancements have been made in improving the prosody of synthetic speech, with many models achieving better pitch accuracy, rhythm, and overall naturalness, thereby narrowing the gap with natural prosody. For in-

stance, Harmonic-Net offers enhanced control over prosodic parameters, effectively replicating the natural variations observed in human speech [Matsubara *et al.*, 2023]. Similarly, models like SiFi-GAN and FastPitchFormant improve pitch controllability and robustness, bringing synthetic speech closer to the prosodic qualities of natural speech [Yoneyama *et al.*, 2023; Bak *et al.*, 2021]. Techniques such as hierarchical modeling and the incorporation of discourse-level linguistic features have further enhanced the coherence and expressiveness of long-form synthetic speech. However, achieving fine-grained intonation control in extended contexts remains a domain where natural speech still excels [Guo *et al.*, 2023; Wu *et al.*, 2022].

Despite these improvements, synthetic speech often lacks the nuanced variability and spontaneous shifts that define natural prosody. For example, models such as FlashSpeech and StyleTTS 2 deliver high audio quality and diverse prosodic features but struggle to capture subtle, context-driven variations that are characteristic of natural speech [Ye *et al.*, 2024; Li *et al.*, 2023]. Meanwhile, models such as Quasi-Periodic WaveNet and CampNet demonstrate lower pitch-related errors (e.g., RMSE), better pitch alignment, and improved temporal accuracy, contributing to more natural-sounding prosody [Wu *et al.*, 2021b; Wang *et al.*, 2022a].

However, subjective evaluations reveal persistent gaps in speaker likeness and emotional depth. For example, synthetic sentences incorporating creaky voice patterns were rated lower in terms of speaker resemblance and overall naturalness compared to natural speech [Mandee *et al.*, 2023b]. These findings underscore that while synthetic speech has achieved notable advancements in prosodic accuracy and expressiveness, further innovations are needed to fully replicate the richness and emotional subtleties of natural prosody.

In general, the conclusions of the studies in relation to the comparison between the prosody of synthetic speech and that of natural speech show advances in a systematic way, since the results indicated that each new technique proposed contributed to bringing synthesized speech closer to natural speech. However, there is room to deepen the analysis, not only mentioning a good performance or a good performance of a proposed model, but also describing details of the prosodic parameters evaluated. For example, it would be useful to describe patterns of different syllable durations, emphasize measures of F0 variation, and explore intensity variants.

9 Conclusions

This review aimed to explore the relationship between speech synthesis and prosody, emphasizing the evaluation methodology used in the studies. Based on the eight prepared questions, we arrived at the following final considerations:

Prosodic parameters: The prosodic parameters most used in the studies were F0, duration and intensity, indicating that future research can focus on these aspects.

Terminological instability: The terminology in the evaluations proved to be unstable in relation to the term “quality”,

but there is a preference for the term “naturalness” in subjective evaluations and a standard in the term “expressiveness” for analyzing emotion.

Metrics: The metrics used in the studies allow for a quick assessment, but it is possible to enrich the assessment methods, based on more detailed analyzes of F0, the duration of each type of syllable and the intensity variants.

Datasets: The identified datasets can provide an overview of the most used datasets. However, it is useful for future research to highlight important issues, such as the number of hours used, the quality of the audios and the transcription process.

Predominance of English: English was the predominant language for evaluating speech synthesis prosody, which suggests that understudied languages will face challenges until large language datasets are created to improve prosody.

Speech synthesis models: Information about the models that were most considered for comparison indicates what research has considered to be the state of the art in the field of synthesis, in addition to highlighting the prosodic problems faced by these models. Future research could explore these adopted models, such as FastSpeech 2, with a focus on prosodic analysis, especially in less represented languages or linguistic varieties, such as Brazilian Portuguese. It is worth noting that, among the 100 studies analyzed, none carried out training and assessment of prosody in this language, highlighting the need for investigations.

Synthetic speech and natural speech: The conclusions of the included studies revealed that the methods proposed by each of the studies presented good performance, in terms of bringing synthetic speech closer to natural speech. However, there is room to further detail these results, including analyzes that allow for a more robust assessment. For example, it would be relevant to present the average difference in F0 variation and check whether the models can reproduce language patterns, such as the typical intonation of neutral declarative and interrogative sentences. Duration analyzes may indicate the normalization of syllable values or the comparison of the duration of different types of syllables, such as tonic, attack⁴ and unstressed.

Challenges of synthetic speech prosody: The studies included indicate that there is a need to generate more prosodic variation in the speech of models and that the lack of large-scale datasets also hinders progress in the field of synthesis. Furthermore, the particularities of each language present specific challenges that demand attention. In this sense, it is necessary to expand the scope of research to less explored languages.

In general, the number of studies included in our review reveals that there is a great concern about improving the

⁴Attack (or onset) is the first syllable of the statement.

prosody of speech synthesis. Despite the challenges, research has demonstrated progress in this field and can be enriched by more robust and more detailed analyzes of each of the prosodic parameters, which can be provided based on the methodology already adopted by research in Linguistics. For example, software, such as Praat [Boersma and Weenink, 2024], has scripts that can provide details of these parameters, improving the evaluation of prosody in speech synthesis, whose methodology requires continuous improvement.

Future research should aim to tackle the challenges outlined in this review, provide a comprehensive evaluation of prosodic parameters in synthesized speech, and support the development and accessibility of large-scale datasets for languages that remain underrepresented.

Declarations

Authors' Contributions

All authors contributed to the conceptualization of this systematic review. SMA contributed to project administration. JCG and ANM contributed to data curation, formal analysis, investigation, methodology, software, and writing - original draft. All authors contributed to writing - review & editing and also read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work was supported by the Center for Artificial Intelligence (C4AI-USP), with funding from the São Paulo Research Foundation (FAPESP Grant No. 2019/07665-4) and IBM Corporation. This project was also supported by the Ministry of Science, Technology and Innovation, with resources of Law No. 8.248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by SofTex and published Residence in TIC 13, DOU 01245.010222/2022-44. We thank the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES), through the Academic Excellence Program (PROEX), for the grants 88887.920395/2023-00 and 88887.976154/2024-00. We also thank the reviewers for reading the article and for their comments. FRFS also thanks the National Council for Scientific and Technological Development (CNPq) for the research productivity grant 304961/2021-3.

Availability of data and materials

The datasets (and/or softwares) generated and/or analysed during the current study will be made upon request.

References

- Al-Radhi, M. S., Ibrahim, O., Mandeel, A. R., Csapó, T. G., and Németh, G. (2023). Advancing limited data text-to-speech synthesis: Non-autoregressive transformer for high-quality parallel synthesis. In *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 152–157. IEEE. DOI: 10.1109/SpeD59241.2023.10314948.
- Alwaisi, S. and Németh, G. (2024). Advancements in expressive speech synthesis: A review. *IN-FOCOMMUNICATIONS JOURNAL*, 16(1):35–46. DOI: 10.36244/ICJ.2024.1.5.
- Aso, M., Takamichi, S., Takamune, N., and Saruwatari, H. (2020). Acoustic model-based subword tokenization and prosodic-context extraction without language knowledge for text-to-speech synthesis. *Speech Communication*, 125:53–60. DOI: 10.1016/j.specom.2020.09.003.
- Bae, J.-S., Bak, T., Joo, Y.-S., and Cho, H.-Y. (2021). Hierarchical context-aware transformers for non-autoregressive text to speech. In *Interspeech 2021*, pages 3610–3614. DOI: 10.21437/Interspeech.2021-471.
- Bai, Q., Ko, T., and Zhang, Y. (2022). A study of modeling rising intonation in cantonese neural speech synthesis. In *Interspeech 2022*, pages 501–505. DOI: 10.21437/Interspeech.2022-11173.
- Bak, T., Bae, J.-S., Bae, H., Kim, Y.-I., and Cho, H.-Y. (2021). Fastpitchformant: Source-filter based decomposed modeling for speech synthesis. In *Interspeech 2021*, pages 116–120. DOI: 10.21437/Interspeech.2021-866.
- Bak, T., Eom, Y., Choi, S., and Joo, Y.-S. (2024). MultiVerse: Efficient and expressive zero-shot multi-task text-to-speech. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9130–9147, Miami, Florida, USA. Association for Computational Linguistics. DOI: 10.18653/v1/2024.findings-emnlp.533.
- Bauer, J., Zalkow, F., Müller, M., and Dittmar, C. (2024). Evaluating the impact of prosody feature normalization on the controllability of pitch in speech synthesis. In Baumann, T., editor, *Elektronische Sprachsignalverarbeitung 2024, Tagungsband der 35. Konferenz, Regensburg, 6.-8. März 2024*, pages 188–195. TUDpress. DOI: 10.35096/othr/pub-7097.
- Boersma, P. and Weenink, D. (2024). Praat: doing phonetics by computer [Computer program]. Version 6.3.10. Available at: <http://www.praat.org/>.
- Bott, T. (2023). Content-aware text-to-speech with prompt-based prosody control. Masterarbeit, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart. Available at: <https://elib.uni-stuttgart.de/items/d2396d8f-70e4-4713-8a20-e2c2dc714dfd>.
- Cagliari, L. C. (1992). Prosódia: algumas funções dos suprasegmentos. *Cadernos de estudos linguísticos*, 23:137–151. DOI: 10.20396/cel.v23i0.8636850.
- Casanova, E., Santos, V. G., Svartman, F. R. F., Leite, M. Q., Candido Junior, A., Marcacini, R. M., Rezende, S. O., and Aluísio, S. M. (2024). Recursos para o processamento de fala. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, book chapter 3. BPLN, 3 edition. Available at: <https://brasileiraspln.com/livro-pln/3a-edicao/parte-fala/cap-recursos-fala/cap->

- recursos-fala.html.
- Chemnad, K. and Othman, A. (2023). Advancements in arabic text-to-speech systems: a 22-year literature review. *IEEE Access*, 11:30929–30954. DOI: 10.1109/ACCESS.2023.3260844.
- Chen, L., Deng, Y., Wang, X., Soong, F. K., and He, L. (2021). Speech bert embedding for improving prosody in neural tts. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6563–6567. IEEE. DOI: 10.1109/ICASSP39728.2021.9413864.
- Chen, S.-H., Hwang, S.-H., and Wang, Y.-R. (1998). An rnn-based prosodic information synthesizer for mandarin text-to-speech. *IEEE transactions on speech and audio processing*, 6(3):226–239. DOI: 10.1109/89.668817.
- Chien, C.-M. and Lee, H.-y. (2021). Hierarchical prosody modeling for non-autoregressive speech synthesis. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 446–453. IEEE. DOI: 10.1109/SLT48900.2021.9383629.
- Cooper, E., Huang, W.-C., Tsao, Y., Wang, H.-M., Toda, T., and Yamagishi, J. (2024). A review on subjective and objective evaluation of synthetic speech. *Acoustical Science and Technology*, 45(4):161–183. DOI: 10.1250/ast.e24.12.
- Deng, Y., Xue, J., Jia, Y., Li, Q., Han, Y., Wang, F., Gao, Y., Ke, D., and Li, Y. (2024). Concss: Contrastive-based context comprehension for dialogue-appropriate prosody in conversational speech synthesis. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10706–10710. IEEE. DOI: 10.1109/icassp48485.2024.10446506.
- Do, P., Coler, M., Dijkstra, J., and Klabbers, E. (2021). A systematic review and analysis of multilingual data strategies in text-to-speech for low-resource languages. In *Interspeech 2021*, pages 16–20. DOI: 10.21437/Interspeech.2021-1565.
- Freitas e Souza, V. (2015). ECOS PL-Science: Uma Arquitetura para Ecossistemas de Software Científico Apoiada por uma Rede Ponto a Ponto. Master’s thesis, Universidade Federal de Juiz de Fora. Available at: <https://repositorio.ufjf.br/jspui/handle/ufjf/4834>.
- Fujii, K., Saito, Y., and Saruwatari, H. (2022). Adaptive end-to-end text-to-speech synthesis based on error correction feedback from humans. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1702–1707. IEEE. DOI: 10.23919/APSIPAASC55919.2022.9979876.
- Fujimaki, D., Nose, T., and Ito, A. (2020). Integration of accent sandhi and prosodic features estimation for japanese text-to-speech synthesis. In *2020 IEEE 9th Global Conference on Consumer Electronics (GCCE)*, pages 358–359. IEEE. DOI: 10.1109/GCCE50665.2020.9291906.
- Furukawa, K. (2022). Applying syntax-prosody mapping hypothesis, prosodic wellformedness constraints, and boundary-driven theory to neural sequence-to-sequence speech synthesis. DOI: 10.48550/arXiv.2203.15276.
- Galdino, J. C. and Oliveira Jr, M. (2023). Prosódia e síntese da fala: uma revisão integrativa da literatura. *Revista da ABRALIN*, pages 1–15. DOI: 10.25189/rabralin.v22i1.2130.
- Gong, C., Wang, L., Ling, Z., Guo, S., Zhang, J., and Dang, J. (2021). Improving naturalness and controllability of sequence-to-sequence speech synthesis by learning local prosody representations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5724–5728. IEEE. DOI: 10.1109/ICASSP39728.2021.9414720.
- Guo, D., Zhu, X., Xue, L., Li, T., Lv, Y., Jiang, Y., and Xie, L. (2023). Hignn-tts: Hierarchical prosody modeling with graph neural networks for expressive long-form tts. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7. IEEE. DOI: 10.1109/ASRU57964.2023.10389629.
- Hamed, M. and Lachiri, Z. (2024a). Expressivity transfer in transformer-based text-to-speech synthesis. In *2024 IEEE 7th International Conference on Advanced Technologies, Signal and Image Processing (ATSIP)*, volume 1, pages 443–448. IEEE. DOI: 10.1109/ATSIP62566.2024.10638975.
- Hamed, M. and Lachiri, Z. (2024b). Fine-grained prosody transfer text-to-speech synthesis with transformer. In *2024 5th International Conference in Electronic Engineering, Information Technology & Education (EEITE)*, pages 1–7. IEEE. DOI: 10.1109/EEITE61750.2024.10654450.
- He, Y., Luan, J., and Wang, Y. (2022). Pama-tts: Progression-aware monotonic attention for stable seq2seq tts with accurate phoneme duration control. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7467–7471. IEEE. DOI: 10.1109/ICASSP43922.2022.9746202.
- Herrmann, B. (2023). The perception of artificial-intelligence (ai) based synthesized speech in younger and older adults. *International Journal of Speech Technology*, 26(2):395–415. DOI: 10.1007/s10772-023-10027-y.
- Hida, R., Hamada, M., Kamada, C., Tsunoo, E., Sekiya, T., and Kumakura, T. (2022). Polyphone disambiguation and accent prediction using pre-trained language models in japanese tts front-end. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7132–7136. IEEE. DOI: 10.1109/ICASSP43922.2022.9746212.
- Hirst, D., Rilliard, A., and Aubergé, V. (1998). Comparison of subjective evaluation and an objective evaluation metric for prosody in text-to-speech synthesis. In *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, pages 1–4. Available at: https://www.researchgate.net/publication/265414079_Comparison_of_Subjective_Evaluation_and_an_Objective_Evaluation_Metric_for_Prosody_in_Text-to-Speech_Synthesis.
- Hodari, Z. (2022). *Synthesising prosody with insufficient context*. PhD thesis, The University of Edinburgh - School of Informatics. DOI: 10.7488/era/2654.
- Huang, R., Zhang, C., Ren, Y., Zhao, Z., and Yu, D. (2023). Prosody-TTS: Improving prosody with masked autoencoder and conditional diffusion model for expressive text-to-speech. In *Findings of the Association for Computa-*

- tional Linguistics: ACL 2023*, pages 8018–8034, Toronto, Canada. Association for Computational Linguistics. DOI: 10.18653/v1/2023.findings-acl.508.
- Iliescu, D. A., Mohan, D. S. R., Teh, T. H., and Houdari, Z. (2024). Controllable prosody generation with partial inputs. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11916–11920. IEEE. DOI: 10.1109/ICASSP48485.2024.10446859.
- Inoue, S., Zhou, K., Wang, S., and Li, H. (2024). Hierarchical emotion prediction and control in text-to-speech synthesis. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10601–10605. IEEE. DOI: 10.1109/ICASSP48485.2024.10445996.
- Jiang, C., Gao, Y., Jin, H., Pan, L., and Ng, W. W. (2024a). Fastmandarin: Efficient local modeling for natural mandarin speech synthesis. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 461–465. IEEE. DOI: 10.1109/ICASSP48485.2024.10446112.
- Jiang, C., Gao, Y., Ng, W. W., Zhou, J., Zhong, J., Zhen, H., and Hu, X. (2024b). Semantic dependency and local convolution for enhancing naturalness and tone in text-to-speech synthesis. *Neurocomputing*, 608:128430. DOI: 10.1016/j.neucom.2024.128430.
- Jiang, Z., Liu, J., Ren, Y., He, J., Ye, Z., Ji, S., Yang, Q., Zhang, C., Wei, P., Wang, C., et al. (2024c). Mega-tts 2: Boosting prompting mechanisms for zero-shot speech synthesis. In *The Twelfth International Conference on Learning Representations*. DOI: 10.48550/arXiv.2307.07218.
- Jiang, Z., Su, Z., Zhao, Z., Yang, Q., Ren, Y., Liu, J., and Ye, Z. (2022). Dict-TTS: learning to pronounce with prior dictionary knowledge for text-to-speech. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, pages 11960–11974, Red Hook, NY, USA. Curran Associates Inc.. DOI: 10.48550/arXiv.2206.02147.
- Jokisch, O., Mixdorff, H., Kruschke, H., and Kordon, U. (2000). Learning the parameters of quantitative prosody models. In *6th International Conference on Spoken Language Processing (ICSLP 2000)*, pages 645–648. DOI: 10.21437/ICSLP.2000-160.
- Ju, Y., Kim, I., Yang, H., Kim, J.-H., Kim, B., Maiti, S., and Watanabe, S. (2022). Trinitts: Pitch-controllable end-to-end tts without external aligner. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 16–20. DOI: 10.21437/Interspeech.2022-925.
- Kaiki, N., Sakti, S., and Nakamura, S. (2021). Using local phrase dependency structure information in neural sequence-to-sequence speech synthesis. In *2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 206–211. IEEE. DOI: 10.1109/O-COCOSDA202152914.2021.9660456.
- Kent, R. D. and Read, C. (2015). *Análise acústica da fala*. Cortez Editora, 1a edition. Book. ISBN-13 978-8524923319.
- Kulkarni, A. (2022). *Expressivity transfer in deep learning based text-to-speech synthesis*. PhD thesis, Université de Lorraine. Available at: https://hal.univ-lorraine.fr/tel-03844914v1/file/DDOC_T_2022_0122_KULKARNI.pdf.
- Kumar, N., Narang, A., and Lall, B. (2022). Zero-shot normalization driven multi-speaker text to speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1679–1693. DOI: 10.1109/TASLP.2022.3169634.
- Kurihara, K. (2024). Phonetic and prosodic features for sequence-to-sequence acoustic modeling on japanese text-to-speech and their estimation. *University of Tsukuba*. DOI: 10.15068/0002012966.
- Lameris, H., Mehta, S., Henter, G. E., Gustafson, J., and Székely, É. (2023). Prosody-controllable spontaneous tts with neural hmms. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE. DOI: 10.1109/ICASSP49357.2023.10097200.
- Łańcucki, A. (2021). Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6588–6592. IEEE. DOI: 10.1109/ICASSP39728.2021.9413889.
- Lee, J., Lee, J. Y., Choi, H., Mun, S., Park, S., Bae, J.-S., and Kim, C. (2022). Into-TTS: Intonation template based prosody control system. DOI: 10.48550/arXiv.2204.01271.
- Lei, S., Zhou, Y., Chen, L., Wu, Z., Kang, S., and Meng, H. (2023). Context-aware coherent speaking style prediction with hierarchical transformers for audiobook speech synthesis. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE. DOI: 10.1109/ICASSP49357.2023.10095866.
- Lenglet, M., Perrotin, O., and Bailly, G. (2023). Local style tokens: Fine-grained prosodic representations for tts expressive control. In *12th ISCA Speech Synthesis Workshop (SSW2023)*, pages 120–126. ISCA. DOI: 10.21437/SSW.2023-19.
- Li, H., Zhu, X., Xue, L., Song, Y., Chen, Y., and Xie, L. (2024a). SponTTS: modeling and transferring spontaneous style for TTS. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12171–12175. IEEE. DOI: 10.1109/ICASSP48485.2024.10445828.
- Li, T., Wang, X., Xie, Q., Wang, Z., and Xie, L. (2022). Cross-speaker emotion disentangling and transfer for end-to-end speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1448–1460. DOI: 10.1109/TASLP.2022.3164181.
- Li, W., Yang, P., Zhong, Y., Zhou, Y., Wang, Z., Wu, Z., Wu, X., and Meng, H. (2024b). Spontaneous style text-to-speech synthesis with controllable spontaneous behaviors based on language models. In *Interspeech 2024*, pages 1785–1789. DOI: 10.21437/Interspeech.2024-1989.
- Li, Y. A., Han, C., Raghavan, V., Mischler, G., and

- Mesgarani, N. (2023). Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 19594–19621. Curran Associates, Inc. Available at: https://proceedings.neurips.cc/paper_files/paper/2023/file/3eaad2a0b62b5ed7a2e66c2188bb1449-Paper-Conference.pdf.
- Liu, J., Xie, Z., Zhang, C., and Shi, G. (2021a). A novel method for mandarin speech synthesis by inserting prosodic structure prediction into tacotron2. *International Journal of Machine Learning and Cybernetics*, 12:2809–2823. DOI: 10.1007/s13042-021-01365-x.
- Liu, R., Sisman, B., Gao, G., and Li, H. (2021b). Expressive tts training with frame and style reconstruction loss. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:1806–1818. DOI: 10.1109/TASLP.2021.3076369.
- Liu, R., Sisman, B., and Li, H. (2021c). Graphspeech: Syntax-aware graph attention network for neural speech synthesis. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6059–6063. IEEE. DOI: 10.1109/ICASSP39728.2021.9413513.
- Liu, Z., Wu, N., Zhang, Y., and Ling, Z. (2022). Integrating discrete word-level style variations into non-autoregressive acoustic models for speech synthesis. In *Interspeech 2022*, pages 5508–5512. DOI: 10.21437/Interspeech.2022-984.
- Liu, Z.-C., Chen, L., Hu, Y.-J., Ling, Z.-H., and Pan, J. (2024). Pe-wav2vec: A prosody-enhanced speech model for self-supervised prosody learning in tts. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:4199–4210. DOI: 10.1109/TASLP.2024.3449148.
- Liu, Z.-C., Ling, Z.-H., Hu, Y.-J., Pan, J., Wang, J.-W., and Wu, Y.-D. (2023). Speech synthesis with self-supervisedly learnt prosodic representations. In *Interspeech 2023*, pages 7–11. DOI: 10.21437/Interspeech.2023-1292.
- Lucente, L. (2020). Função comunicativa e alinhamento de contorno entoacional descendente. In *Anais do I Congresso Brasileiro de Prosódia*, volume 1, pages 31–34. Available at: http://www.periodicos.letras.ufmg.br/index.php/anais_coloquio/article/view/16762/1125613194.
- Luo, X., Takamichi, S., Koriyama, T., Saito, Y., and Saruwatari, H. (2021). Emotion-controllable speech synthesis using emotion soft labels and fine-grained prosody factors. In *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 794–799. IEEE. DOI: 10.1561/103.000000003.
- Ma, F., Li, Y., Xie, Y., He, Y., Zhang, Y., Ren, H., Liu, Z., Yao, W., Ren, F., Yu, F. R., and Ni, S. (2024). A review of human emotion synthesis based on generative technology. DOI: 10.48550/arXiv.2412.07116.
- Malviya, S., Mishra, R., Barnwal, S. K., and Tiwary, U. S. (2023). A framework for quality assessment of synthesised speech using learning-based objective evaluation. *International Journal of Speech Technology*, 26(1):221–243. DOI: 10.1007/s10772-023-10021-4.
- Mandeel, A. R., Al-Radhi, M. S., and Csapó, T. G. (2023a). Enhancing end-to-end speech synthesis by modeling interrogative sentences with speaker adaptation. In *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 158–163. IEEE. DOI: 10.1109/SpeD59241.2023.10314910.
- Mandeel, A. R., Al-Radhi, M. S., and Csapó, T. G. (2023b). Modeling irregular voice in end-to-end speech synthesis via speaker adaptation. In *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 170–175. IEEE. DOI: 10.1109/SpeD59241.2023.10314920.
- Matsubara, K., Okamoto, T., Takashima, R., Takiguchi, T., Toda, T., and Kawai, H. (2023). Harmonic-net: Fundamental frequency and speech rate controllable fast neural vocoder. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1902–1915. DOI: 10.1109/TASLP.2023.3275032.
- Mohan, D. S. R., Hu, V., Teh, T. H., Torresquintero, A., Wallis, C. G., Staib, M., Foglianti, L., Gao, J., and King, S. (2021). Ctrl-p: Temporal control of prosodic variation for speech synthesis. In *Interspeech 2021*, pages 3875–3879. DOI: 10.21437/Interspeech.2021-1583.
- Moon, S., Kim, S., and Choi, Y.-H. (2022). Mist-tacotron: End-to-end emotional speech synthesis using mel-spectrogram image style transfer. *IEEE Access*, 10:25455–25463. DOI: 10.1109/ACCESS.2022.3156093.
- Moraes, J. A. d. and Rilliard, A. (2022). Entoação. In Jr., M. O., editor, *Prosódia, Prosódias: uma introdução*, pages 45–66. Editora Contexto. Book.
- Moraes, J. d. (1987). Correlatos acústicos de l’accent de mot en portugais brésilien. In *Proceedings XIth ICPhs: The Eleventh International Congress of Phonetic Sciences, August 1-7, 1987, Tallinn, Estonia, U.S.S.R.*, volume 3, pages 313–316. Academy of Sciences of the Estonian S.S.R. Available at: https://www.coli.uni-saarland.de/groups/FK/speech_science/icphs/ICPhS1987/11_ICPhS_1987_Vol1_3/p11.3_313.pdf.
- Murray, I. R. and Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, 93(2):1097–1108. DOI: 10.1121/1.405558.
- Ogun, S., Colotte, V., and Vincent, E. (2023). Stochastic pitch prediction improves the diversity and naturalness of speech in glow-tts. In *Interspeech 2023*, pages 4878–4882. DOI: 10.21437/Interspeech.2023-1673.
- Oh, H.-S., Lee, S.-H., and Lee, S.-W. (2024). Diff-prosody: Diffusion-based latent prosody generation for expressive speech synthesis with prosody conditional adversarial training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2654–2666. DOI: 10.1109/TASLP.2024.3395994.
- O’Mahony, J., Lai, C., and King, S. (2023). Synthesising turn-taking cues using natural conversational data. In *12th ISCA Speech Synthesis Workshop (SSW2023)*, pages 75–80. DOI: 10.21437/SSW.2023-12.

- O'Mahony, J., Corkey, N., Lai, C., Klabbbers, E., and King, S. (2024). Hierarchical intonation modelling for speech synthesis using legendre polynomial coefficients. In *Proc. SpeechProsody 2024*, pages 1030–1034. DOI: 10.21437/SpeechProsody.2024-208.
- Pamisetty, G. and Sri Rama Murty, K. (2023). Prosody-tts: An end-to-end speech synthesis system with prosody control. *Circuits, Systems, and Signal Processing*, 42(1):361–384. DOI: 10.1007/s00034-022-02126-z.
- Peng, Y. and Ling, Z. (2022). Decoupled pronunciation and prosody modeling in meta-learning-based multilingual speech synthesis. In *Interspeech 2022*, pages 4257–4261. DOI: 10.21437/Interspeech.2022-831.
- Prateek, N. (2023). Data efficiency in neural stylistic speech synthesis. Master's thesis, International Institute of Information Technology - Hyderabad, INDIA. Available at: https://web2py.iiit.ac.in/research_centres/publications/view_publication/mastersthesis/1342.
- Přibil, J., Přibilová, A., and Matoušek, J. (2020). Automatic statistical evaluation of quality of unit selection speech synthesis with different prosody manipulations. *Journal of Electrical Engineering*, 71(2):78–86. DOI: 10.2478/jee-2020-0012.
- Raitio, T., Rasipuram, R., and Castellani, D. (2020). Controllable neural text-to-speech synthesis using intuitive prosodic features. In *Interspeech 2020*, pages 4432–4436. DOI: 10.21437/Interspeech.2020-2861.
- Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. (2021a). Fastspeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations*. DOI: 10.48550/arXiv.2006.04558.
- Ren, Y., Liu, J., and Zhao, Z. (2021b). Portaspeech: portable and high-quality generative text-to-speech. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, volume 34 of *NIPS '21*, pages 13963–13974, Red Hook, NY, USA. Curran Associates Inc.. DOI: <https://doi.org/10.48550/arXiv.2109.15166>.
- Sadekova, T., Kudinov, M., Popov, V., Yermekova, A., and Khrapov, A. (2024). Pitchflow: adding pitch control to a flow-matching based tts model. In *Proc. Interspeech 2024*, pages 4418–4422. DOI: 10.21437/interspeech.2024-1023.
- Sagisaka, Y., Campbell, N., and Higuchi, N. (1997). *Computing prosody: computational models for processing spontaneous speech*. Springer Science & Business Media. Book.
- Santos, V. G., Alves, C., Carlotto, B., Dias, B., Gris, L., Izaias, R., Morais, M. L., Oliveira, P., Sicoli, R., Svartman, F. R. F., et al. (2022). CORAA NURC-SP minimal corpus: a manually annotated corpus of brazilian portuguese spontaneous speech. In *6th International Conference on Speech and Language Technologies on Iberian languages, IberSPEECH*, pages 161–165. DOI: 10.21437/IberSPEECH.2022-33.
- Sini, A. (2020). *Caractérisation et génération de l'expressivité en fonction des styles de parole pour la construction de livres audio*. PhD thesis, Rennes I. Available at: <https://syntheses.univ-rennes1.fr/search-theses/notice/view/rennes1-ori-wf-1-14015>.
- Takumi, W., Sunao, H., and Masanobu, A. (2024). Explicit prosody control to realize discourse focus in end-to-end text-to-speech. In *2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE. DOI: 10.1109/MLSP58920.2024.10734738.
- Tan, D., Deng, L., Yeung, Y. T., Jiang, X., Chen, X., and Lee, T. (2021). Editspeech: A text based speech editing system using partial inference and bidirectional fusion. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 626–633. IEEE. DOI: 10.1109/ASRU51503.2021.9688051.
- Taylor, P. (2009). *Text-to-speech synthesis*. Cambridge University Press. DOI: 10.1017/CBO9780511816338.
- The Nguyen, L., Pham, T., and Nguyen, D. Q. (2023). Xphonebert: A pre-trained multilingual model for phoneme representations for text-to-speech. In *Interspeech 2023*, pages 5506–5510. DOI: 10.21437/Interspeech.2023-444.
- Turkmen, T. (2021). *Duration modelling for expressive text to speech*. PhD thesis, Politecnico di Milano. Available at: https://www.politesi.polimi.it/retrieve/a0ad5454-4e82-4041-a757-5660d683016e/Executive_Summary_Duration_Modelling_for_Expressive_TTS.pdf.
- Törö, T. (2022). Analysis of a latent prosody space for controlling speaking styles in finnish end-to-end speech synthesis. Master's thesis, Faculty of Arts - University of Helsinki. Available at: <https://helda.helsinki.fi/server/api/core/bitstreams/dc1257c7-1b96-4db4-9741-ab676c15c0c7/content>.
- Van Santen, J. P., Sproat, R., Olive, J., and Hirschberg, J., editors (1997). *Progress in speech synthesis*. Springer Science & Business Media. DOI: 10.1007/978-1-4612-1894-4.
- Wagner, P., Beskow, J., Betz, S., Edlund, J., Gustafson, J., Eje Henter, G., Le Maguer, S., Malisz, Z., Éva Székely, Tännander, C., and Voße, J. (2019). Speech synthesis evaluation — state-of-the-art assessment and suggestion for a novel research program. In *10th ISCA Workshop on Speech Synthesis (SSW 10)*, pages 105–110. DOI: 10.21437/SSW.2019-19.
- Wang, T., Yi, J., Fu, R., Tao, J., and Wen, Z. (2022a). Campnet: Context-aware mask prediction for end-to-end text-based speech editing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2241–2254. DOI: 10.1109/TASLP.2022.3190717.
- Wang, Y., Xie, Y., Zhao, K., Wang, H., and Zhang, Q. (2022b). Unsupervised quantized prosody representation for controllable speech synthesis. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE. DOI: 10.1109/ICME52920.2022.9859946.
- Wu, N.-Q., Liu, Z.-C., and Ling, Z.-H. (2022). Discourse-level prosody modeling with a variational autoencoder for non-autoregressive expressive speech synthesis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7592–7596. IEEE. DOI: 10.1109/ICASSP43922.2022.9746238.
- Wu, Y.-C., Hayashi, T., Okamoto, T., Kawai, H., and

- Toda, T. (2021a). Quasi-periodic parallel wavenet: A non-autoregressive raw waveform generative model with pitch-dependent dilated convolution neural network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:792–806. DOI: 10.1109/TASLP.2021.3061245.
- Wu, Y.-C., Hayashi, T., Tobing, P. L., Kobayashi, K., and Toda, T. (2021b). Quasi-periodic wavenet: An autoregressive raw waveform generative model with pitch-dependent dilated convolution neural network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1134–1148. DOI: 10.1109/TASLP.2021.3061245.
- Xiao, Y., Zhang, S., Wang, X., Tan, X., He, L., Zhao, S., Soong, F. K., and Lee, T. (2023). Contextspeech: Expressive and efficient text-to-speech for paragraph reading. In *Interspeech 2023*, pages 4883–4887. DOI: 10.21437/Interspeech.2023-122.
- Xin, D., Adavan, S., Ang, F., Kulkarni, A., Takamichi, S., and Saruwatari, H. (2023). Improving speech prosody of audiobook text-to-speech synthesis with acoustic and textual contexts. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE. DOI: 10.1109/ICASSP49357.2023.10096247.
- Xu, C., Moore, B. C., Diao, M., Li, X., and Zheng, C. (2024). Predicting the intelligibility of mandarin chinese with manipulated and intact tonal information for normal-hearing listeners. *The Journal of the Acoustical Society of America*, 156(5):3088–3101. DOI: 10.1121/10.0034233.
- Xue, L., Soong, F. K., Zhang, S., and Xie, L. (2022). Paratts: Learning linguistic and prosodic cross-sentence information in paragraph-based tts. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2854–2864. DOI: 10.1109/TASLP.2022.3202126.
- Yanagita, T., Sakti, S., and Nakamura, S. (2023). Japanese neural incremental text-to-speech synthesis framework with an accent phrase input. *IEEE Access*, 11:22355–22363. DOI: 10.1109/ACCESS.2023.3251657.
- Yang, F., Luan, J., and Wang, Y. (2022). Improving emotional speech synthesis by using sus-constrained vae and text encoder aggregation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8302–8306. IEEE. DOI: 10.1109/ICASSP43922.2022.9746994.
- Yang, F., Yang, S., Wu, Q., Wang, Y., and Xie, L. (2020). Exploiting deep sentential context for expressive end-to-end speech synthesis. In *Interspeech 2020*, pages 3436–3440. DOI: 10.21437/Interspeech.2020-2423.
- Yasuda, Y. (2021). *Lexical pitch accent and duration modeling for neural end-to-end text-to-speech synthesis*. PhD thesis, Graduate University for Advanced Studies, Japan. Available at: <https://ir.soken.ac.jp/record/6414/files/A2241%E6%9C%AC%E6%96%87.pdf>.
- Ye, Z., Ju, Z., Liu, H., Tan, X., Chen, J., Lu, Y., Sun, P., Pan, J., Bian, W., He, S., et al. (2024). Flashspeech: Efficient zero-shot speech synthesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6998–7007, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3664647.3681044.
- Yoneyama, R., Wu, Y.-C., and Toda, T. (2023). High-fidelity and pitch-controllable neural vocoder based on unified source-filter networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. DOI: 10.1109/TASLP.2023.3313410.
- Yufune, K., Koriyama, T., Takamichi, S., and Saruwatari, H. (2021). Accent modeling of low-resourced dialect in pitch accent language using variational autoencoder. *Proc. SSW*, pages 189–194. DOI: 10.21437/SSW.2021-33.
- Zangar, I., Mnasri, Z., Colotte, V., and Jouviet, D. (2021). Duration modelling and evaluation for arabic statistical parametric speech synthesis. *Multimedia Tools and Applications*, 80:8331–8353. DOI: 10.1007/s11042-020-09901-7.
- Zhang, G., Qin, Y., Tan, D., and Lee, T. (2021). Applying the information bottleneck principle to prosodic representation learning. In *Interspeech 2021*, pages 3156–3160. DOI: 10.21437/Interspeech.2021-1049.
- Zhang, M., Zhou, X., Wu, Z., and Li, H. (2023a). Towards zero-shot multi-speaker multi-accent text-to-speech synthesis. *IEEE Signal Processing Letters*, 30:947–951. DOI: 10.1109/LSP.2023.3292740.
- Zhang, M., Zhou, Y., Wu, Z., and Li, H. (2023b). Zero-shot multi-speaker accent tts with limited accent data. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1931–1936. IEEE. DOI: 10.1109/APSIPAASC58517.2023.10317526.
- Zhang, Y.-J., Zhang, C., Song, W., Zhang, Z., Wu, Y., and He, X. (2023c). Prosody modelling with pre-trained cross-utterance representations for improved speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2812–2823. DOI: 10.1109/TASLP.2023.3278184.
- Zhao, Z., Chen, X., Liu, H., Wang, X., Yang, L., and Wang, J. (2021). SPTTS: Parallel speech synthesis without extra aligner model. In *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 864–869. IEEE. Available at: <http://www.apsipa.org/proceedings/2021/pdfs/0000864.pdf>.
- Zhou, X., Zhang, M., Zhou, Y., Wu, Z., and Li, H. (2024). Accented text-to-speech synthesis with limited data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1699–1711. DOI: 10.1109/TASLP.2024.3363414.
- Zou, Y., Liu, S., Yin, X., Lin, H., Wang, C., Zhang, H., and Ma, Z. (2021). Fine-grained prosody modeling in neural speech synthesis using tobi representation. In *Interspeech 2021*, pages 3146–3150. DOI: 10.21437/Interspeech.2021-883.

A Summary of included studies

Table A. Summary of selected studies. The last 4 studies (97-100) were already being studied by the authors and were included in the systematic review.

Number	Reference	Parameter(s)	Metrics	Dataset(s)	Language(s)	Comparison between models (Y/N)	Models
1	Mandee <i>et al.</i> [2023b]	F0, jitter, shimmer	Creakiness percentage, Harmonic-to-Noise Ratio (HNR), Surfboard, MOS, MUSHRA	LJSpeech (public), Hi-Fi multi-speaker (public)	English	N	FastSpeech 2
2	Wang <i>et al.</i> [2022b]	Pitch, speaking velocity	GPE, FFE, MCD, AB, MOSNET	Blizzard 2013 (public), Chinese Standard Mandarin Speech Corpus (public)	Mandarin Chinese	Y	Tacotron
3	Al-Radhi <i>et al.</i> [2023]	Pitch	Praat script, MUSHRA, MOS	Arabic speech corpus (public)	Arabic	Y	Tacotron 2, FastSpeech 2
4	Matsubara <i>et al.</i> [2023]	F0, speech rate	MCD, F0 RMSE, Voice/Unvoiced (V/UV) Decision Error, MOS, AB	JVS Corpus (public), JSUT (public)	Japanese	Y	FastSpeech
5	Liu <i>et al.</i> [2021b]	F0, duration, spectral features	MCD, F0 RMSE, Frame disturbance (FD), MOS	IEMOCAP (public), LJ-Speech (public)	English	Y	Tacotron
6	Yoneyama <i>et al.</i> [2023]	F0	MCD, F0 RMSE, MOS, AB	VCTK (public)	English	N	-
7	Wu <i>et al.</i> [2021a]	Pitch (F0)	MCD, F0 RMSE, MOS, AB	CMU-ARCTIC (public) Voice Conversion Challenge (VCC) 2018 (public)	English	N	-
8	Wu <i>et al.</i> [2021b]	Pitch (F0)	MCD, F0 RMSE, MOS	CMU-ARCTIC (public), VCC2018 (public)	English	Y	WaveNet
9	Xin <i>et al.</i> [2023]	F0	F0 RMSE, GPE, MOS, AB	J-MAC (public)	Japanese	Y	FastSpeech 2
10	Lei <i>et al.</i> [2023]	F0	F0 RMSE, Duration MSE, MOS	Internal audiobook corpus (internal)	Mandarin	Y	FastSpeech 2
11	Fujimaki <i>et al.</i> [2020]	F0	F0 RMSE, Comparison test	Speech data spoken by a single female speaker (internal)	Japanese	N	-
12	Inoue <i>et al.</i> [2024]	Pitch (F0), duration, energy	OpenSMILE, MOS	Blizzard Challenge (public), Emotion Speech Dataset (ESD) (public)	English	Y	FastSpeech 2
13	Guo <i>et al.</i> [2023]	Pitch, duration	MCD, Duration MSE, F0 RMSE, MOS	Internal audiobook corpus (internal), Internal studio-quality dataset (internal)	Mandarin	Y	FastSpeech 2
14	Mandee <i>et al.</i> [2023a]	Pitch	Frequency Weighted Segmental SNR (FwSNRseg), MUSHRA	LJSpeech (public), Hi-Fi multi-speaker dataset (public)	English	N	FastSpeech 2
15	Yanagita <i>et al.</i> [2023]	Pitch (F0)	Perceptual-based measure (pitch), MOS	JSUT (public)	Japanese	N	-
16	Deng <i>et al.</i> [2024]	F0	F0 RMSE, Naturalness MOS, CMOS	Chinese Conversational Speech Corpus (public)	Mandarin Chinese	N	M2CTTS
17	Zhou <i>et al.</i> [2024]	F0, duration	F0 RMSE, Phoneme duration RMSE, AB test	VCTK (Pre-training) (public), CMU-ARCTIC (public)	English	Y	Char-AM
18	Hida <i>et al.</i> [2022]	Pitch	Pitch of all moras (Snt-Exact), Mora-Accuracy, MOS	JSUT (public)	Japanese	N	-
19	Wang <i>et al.</i> [2022a]	F0	F0 RMSE, MOS	VCTK (public), LibriTTS (public)	English	Y	Tacotron, TransformerTTS
20	Zhang <i>et al.</i> [2023b]	Pitch (F0)	F0 RMSE, DTW, MOS	VCTK (public)	English	N	-
21	Yang <i>et al.</i> [2022]	F0, duration, energy	AB test	Mandarin corpora from a male speaker (internal)	Mandarin	Y	Tacotron 2
22	Liu <i>et al.</i> [2021c]	F0	RMSE, MOS	LJSpeech (public)	English	Y	TransformerTTS
23	Gong <i>et al.</i> [2021]	Pitch (F0)	F0 RMSE, MOS	Publicly available single-speaker Chinese standard Mandarin speech corpus (public)	Mandarin	N	-
24	Tan <i>et al.</i> [2021]	Duration	Mel-cepstral distortion (MCD), ABX	LJSpeech (public), VCTK (public), MST (internal), Aishell-3 (public)	English, Chinese	N	-
25	Zangar <i>et al.</i> [2021]	Duration	Duration RMSE, MAE	Arabic Speech Corpus (public)	Arabic	N	-
26	Herrmann [2023]	F0, duration	Praat for sentence duration and median F0	-	English	N	WaveNet
27	Malviya <i>et al.</i> [2023]	F0 (pitch)	F0 RMSE, MOS	CMU-INDIC Hindi TTS (public), IITM Hindi speech dataset (public)	Hindi	Y	HMM, DNN
28	Liu <i>et al.</i> [2021a]	Prosodic word (PW), prosodic phrase (PPH), and intonational phrase (IPH)	PESQ (Perceptual Evaluation of Speech Quality), Mel-SD	BiaoBei Technology Company Dataset (public)	Mandarin	Y	Tacotron2
29	Bai <i>et al.</i> [2022]	Pitch (F0)	FFE, MOS	CanTTS3 (internal)	Cantonese	Y	Tacotron 2
30	Yufune <i>et al.</i> [2021]	F0	F0 RMSE, XAB test	Subset of the JSUT corpus (public)	Japanese	N	-
31	Aso <i>et al.</i> [2020]	F0	F0 RMSE, AB test	JSUT Basic5000 Subcorpus (public), JNAS (public)	Japanese	Y	HMM, DNN
32	Fujii <i>et al.</i> [2022]	Pitch, duration	F0 RMSE, AB test	JSUT (public)	Japanese	Y	FastSpeech 2
33	Törö [2022]	F0, duration, pauses	F0 measurements and spectral tilt using Praat, AB test, MOS	One female native Finnish professional speaker (internal)	Finnish	Y	Tacotron 2, WaveNet
34	Furukawa [2022]	Pitch, duration	FallSize	Speech database created by Kaiki et al (internal)	Japanese	N	-
35	Zhang <i>et al.</i> [2021]	F0	GPE, FFE, MCD22	Blizzard 2013 (public)	English	N	-
36	Přibil <i>et al.</i> [2020]	F0, signal energy (Enc0), differential F0 (F0DIFF), jitter, shimmer, zero-crossing period, zero-crossing frequency	Calculation of TDUR, Auditory preference test	Speech database performed by four speakers (internal)	Czech	N	-
37	Sini [2020]	Intensity, F0, duration, pauses	Pitch extraction algorithms, temporal segmentation of the speech units, absence of acoustic signal between speech units	SynPaFlex-Corpus (public), Multispeaker French Audiobooks corpus dedicated to expressive read Speech Analysis (MUFASA) Corpus (internal)	French	Y	HMM, DNN
38	Bott [2023]	Pitch (F0), energy, duration	MOS	LJSpeech (public), LibriTTS (public), ESD (Emotional Speech Database) (public), RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) (public), TESS (Toronto Emotional Speech Set) (public)	English	Y	FastSpeech 2
39	Xiao <i>et al.</i> [2023]	Pitch, intensity, duration, pause	MOS	Audiobook corpus with an expressive Chinese male voice (internal)	Mandarin Chinese	Y	ConformerTTS
40	Oh <i>et al.</i> [2024]	F0	Pitch RMSE, MOS	LibriTTS (public), LJSpeech (public), Three individual speaker datasets constructed in a similar way from LibriVox recordings (internal)	English	N	-
41	Iliescu <i>et al.</i> [2024]	F0, energy, duration	RMSE, a/b/r	A proprietary Latin American Spanish corpus (internal)	Spanish	N	-

42	Li et al. [2022]	Pitch	Pitch trajectories plots, DMOS	DB1 (public), Children imitated by adults (AIC) (internal), DB6 (internal)	Mandarin	Y	Tacotron2, GST-TTS
43	Mohan et al. [2021]	F0, energy, duration	MUSHRA	A proprietary, multi-speaker, Mexican-Spanish corpus (internal) Spanish	Mexican-Spanish	Y	Tacotron 2
44	Prateek [2023]	F0 (pitch)	MUSHRA, FRMSE, FCORR, GPE, FPE (cents)	Internal dataset (internal), VCTK (public)	English	N	-
45	Peng and Ling [2022]	F0, energy	F0-RMSE, Pearson correlation coefficient of F0 (F0-CORR), EN-RMSE	CSS (public), Common Voice (public)	Mandarin Chinese, German, French, Dutch, Russian	Y	Tacotron2, Meta-char, Meta-IPA
46	Jiang et al. [2022]	Pitch, duration	DTW, MOS-P (Prosody: naturalness of pitch, energy, and duration), MOS-Q (Quality: clarity, high-frequency, and original timbre reconstruction)	Biaobei (public), JSUT (public), Common Voice (HK) (public)	Japanese, Cantonese	Y	NLR-TTS
47	Oh et al. [2024]	Pitch, energy, duration	Average differences in utterance duration (DDUR), Pitch error (RMSE _{F0}) in cents, Periodicity error (RMSE _{period}), Kullback-Leibler (KL) divergence of log f_0 , Real-time factor (RTF)	VCTK (public) + LibriTTS (public)	English	Y	FastSpeech 2, ProsoSpeech
48	Wu et al. [2022]	F0, duration, energy	Wasserstein Distance, Energy Distance, MOS	Audiobook (Chinese) (internal)	Mandarin Chinese	Y	FastSpeech2
49	Turkmen [2021]	Pitch, duration, speech rate, energy	Mean Absolute Duration Error, Mean Absolute Speech Rate Error, Absolute Difference in Standard Deviation, Absolute Difference in Skewness, Absolute Difference in Kurtosis, DTW, MOS	An internal dataset (internal), LibriSpeech (public)	Italian, English	Y	Tacotron 2, FastSpeech 2
50	Luo et al. [2021]	Pitch, energy	MSE, MOS	IEMOCAP (public), Blizzard2013 (public)	English	N	-
51	Bauer et al. [2024]	Pitch	CREPE: A convolutional representation for pitch estimation	Hi-fi multi-speaker english tts dataset (public), Male 1 (public), Female 2 (internal), Male 2 (internal)	English	Y	Tacotron
52	Takumi et al. [2024]	F0	Average differences of F0, MOS	Dataset for news sentences (JNAS) (public), Dataset for novels (J-KAC): Japanese Kamishibai and Audiobook Corpus (public)	Japanese	Y	VITS
53	Yang et al. [2020]	Energy, duration, F0	Pearson correlation coefficient, MOS, AB test	Publicly-available Blizzard Challenge 2019 corpus (public), Voice assistant (VA) corpus (internal), DB1 corpus (public)	Mandarin	Y	Tacotron2
54	Kulkarni [2022]	F0	F0 RMSE, MOS	Emotional Voice Database 2 (public), LJSpeech (public), Canadian French emotional dataset (public), Caroline dataset (internal), Synpaflex dataset (public), Lisa dataset (internal), Simple4All Tundra dataset (public), Siwis dataset (public)	French	Y	Tacotron 2, WaveNet
55	Hamed and Lachiri [2024a]	Pitch	F0 frame error (FFE), Gross Pitch Error (GPE), MOS	LJSpeech (public), Emotional Speech (ESD) (public)	English	N	-
56	Jiang et al. [2024a]	F0	Goodness of fit for F0, MOS, ABX	BIAOBEI (public)	Mandarin	Y	FastSpeech2, Tacotron
57	Bak et al. [2021]	Pitch, duration	FFE, MOS	An internal Korean speaker dataset (internal)	Korean	Y	FastPitch
58	Zou et al. [2021]	Pitch, Pause	F1 score, AB test	An internal English dataset (internal)	English	Y	Tacotron
59	Hamed and Lachiri [2024b]	Pitch	FFE, GPE, MOS, MUSHRA	LJSpeech (public), ESD (public)	English	N	-
60	Ye et al. [2024]	Pitch, duration	Prosody JS Divergence metric, CMOS, SMOS, UTMOS	Subset of Multilingual LibriSpeech (MLS) (public)	English	Y	VALL-E, NaturalSpeech 2, Voicebox, Mega-TTS, CLaM-TTS, FlashSpeech, YourTTS
61	Bae et al. [2021]	Pitch, duration	FFE, MCD, MOS	An internal Korean female speech dataset (internal)	Korean	Y	FastPitch
62	O'Mahony et al. [2024]	F0, duration	RMSE	LJSpeech (public)	English	Y	FastPitch
63	Chien and Lee [2021]	F0	GPE, FFE, MOS, AXV	LJSpeech (public)	English	N	-
64	Liu et al. [2022]	F0	F0 correlation, F0 RMSE, F0 Standard deviation (SD), MOS	Blizzard Challenge 2019 dataset (public)	Mandarin Chinese	Y	FastSpeech 2, Tacotron 2
65	Lee et al. [2022]	F0	Euclidean distance of the F0 values	An internal database (internal), VCTK (public)	English	Y	Tacotron
66	Yanagita et al. [2023]	F0	Perceptual-based measure, MOS	JSUT (public)	Japanese	N	-
67	Yasuda [2021]	F0, duration	RMSE, Correlation, Unvoiced/voiced (U/V) errors of F0, MOS	ATR Ximera (-), Chinese Mandarin speech corpus (public)	Japanese, Mandarin Chinese	Y	Tacotron, WaveNet
68	Lenglet et al. [2023]	Duration, pitch, energy	Duration Error (ms) Pitch Error (Semitones), Energy Error (dB)	Internal French dataset (internal)	French	Y	FastSpeech 2
69	Jiang et al. [2024c]	Pitch, duration	DTW, Duration error (DE), MOS	LibriLight (public)	English	Y	VALL-E, Mega-TTS 2
70	Moon et al. [2022]	Pitch	F0 voiced error (FVE), GPE, MOS	Single speaker voice data (internal), Multi-speaker dataset by Pitchtron (-)	Korean	Y	Tacotron
71	Bak et al. [2024]	Duration, pitch	DTW, F0 Pearson correlation coefficient (F0 PCC), Dur. RMSE, Similarity-MOS (S-MOS), Prosody Similarity-MOS, (PS-MOS)	LibriTTS (public), VCTK (public), Internal dataset (internal), AI-Hub (public)	English, Korean	Y	GANSpeech, YourTTS
72	He et al. [2022]	Duration	MAE, MOS	An internal corpus (internal)	Mandarin Chinese	Y	Tacotron 2, PAMA-TTS
73	Xue et al. [2022]	Pitch, energy, duration, pause	Pearson correlation coefficient, Pause RMSE, Logarithmic fundamental frequency (LF0) with Python library of Parselmouth, MOS	Audiobook corpus (internal)	Mandarin Chinese	Y	Tacotron 2, ParaTTS
74	Liu et al. [2024]	Pitch, energy, duration	F0 Corr, F0 RMSE, Energy Corr, Energy RMSE, Duration Corr, Duration RMSE, MOS	An iFLYTEK internal speech dataset (internal), An internal audiobook dataset (internal), Blizzard Challenge 2019 (public)	Mandarin Chinese	Y	FastSpeech 2, SAL-TTS
75	Kurihara [2024]	Pitch	F0 Correlation, MOS	JSUT (public)	Japanese	Y	Tacotron 2, Transformer TTS

76	Sadekova et al. [2024]	Pitch	Cosine similarity, MOS	LibriTTS (public), LibriLight (public)	English	Y	YourTTS, VALL-E
77	Ren et al. [2021b]	Pitch, energy, duration	The MOS and CMOS in two aspects: prosody (naturalness of pitch, energy and duration) and audio quality (clarity, high-frequency and original timbre reconstruction), and score MOS-P/CMOS-P and MOS-Q/CMOS-Q corresponding to the MOS/CMOS of prosody and audio quality	LJSpeech (public)	English	Y	Tacotron 2, TransformerTTS, FastSpeech, FastSpeech 2, Glow-TTS, BVAE-TTS, PortaSpeech
78	Lameris et al. [2023]	F0, Speech rate	Wavelet Prosody Toolkit (WPT), CMOS, MUSHRA	RyanSpeech (public), Trinity Speech and Gesture Dataset (public), LJSpeech (public)	English	Y	HMM
79	Zhang et al. [2023c]	F0 (pitch), duration	F0 RMSE, F0 CORR, MOS, AB test	A Mandarin audiobook dataset (internal), Blizzard Challenge (public)	English	Y	Fastspeech2
80	Pamisetty and Sri Rama Murty [2023]	F0, duration	RMSE, MAE, (Pearson's correlation coefficient (PCC), Voicing decision error (VDE), GPE, FFE	IndicTTS (-)	Telugu	Y	Tacotron, FastSpeech
81	Huang et al. [2023]	Pitch, energy, duration	MCD, MOS	LibriSpeech 2015 (public), LJSpeech (public), LibriTTS (public)	English	Y	FastSpeech 2, StyleSpeech, Glow-TTS, Grad-TTS, YourTTS, Prosody-TTS
82	Jiang et al. [2024b]	F0	Goodness of fit, MOS, AB test	BIAOBEL (public)	Mandarin Chinese	Y	FastSpeech2, Tacotron, BertSpeech
83	Chen et al. [2021]	F0, energy, duration	MSE, MOS	An internal single-speaker dataset (internal), LJSpeech (public), An internal multi-speaker dataset (internal), VCTK (public), LibriTTS (public)	English	Y	Transformer TTS
84	Liu et al. [2023]	F0, energy, duration	F0 RMSE, F0 correlation, Duration correlation, MOS	An internal dataset (internal), An internal dataset (internal)	Chinese	Y	FastSpeech 2
85	Li et al. [2024b]	Pitch (F0), duration	MCD, CMOS	Pre-training: WenetSpeech corpus (public), High-quality internal dataset (internal)	Mandarin	Y	FastSpeech, VALL-E
86	Li et al. [2024a]	Pitch, duration	Deviations (STD) of F0 and duration by Harvest, MOS, N-MOS, S-MOS	An internal Mandarin (internal), Multi-speaker reading-style internal Mandarin dataset (internal)	Mandarin	Y	FastSpeech
87	Zhao et al. [2021]	Duration	MOS	Baker (public)	Mandarin Chinese	Y	Tacotron 2, FastSpeech, AlignTTS
88	Ogun et al. [2023]	Pitch, duration	Distribution of log-F0 values, MOS	Common Voice dataset (public)	English	Y	Glow-TTS
89	Li et al. [2023]	Pitch, duration, energy	F0 RMSE, Mean absolute deviation of phoneme duration (DUR MAD), Mean value of the F0 and energy curves, MOS	LJSpeech (public), VCTK (public), LibriTTS (public)	English	Y	NaturalSpeech, VITS, VALL-E, StyleTTS 2
90	Hodari [2022]	F0, duration, energy	F0 RMSE, MOS	Blizzard Challenge 2016 (public), An expressive single-speaker proprietary Amazon dataset (internal)	English	Y	Tacotron 2
91	O'Mahony et al. [2023]	F0, intensity, duration	Legendre Polynomial (LP) decomposition, MOS	CANDOR Corpus (public), LJSpeech (public)	English	Y	FastPitch
92	Zhang et al. [2023a]	F0	F0 RMSE, Correlation, DTW, MOS	VCTK (public), L2-ARCTIC (public)	Arabic, English, Hindi, Korean, Spanish, Vietnamese	Y	Tacotron 2
93	Kaiki et al. [2021]	F0, pause	F0 pauses and resets in alignment with syntactic boundaries, AB test, MOS	Speech data of a single speaker (internal)	Japanese	Y	Tacotron 2
94	The Nguyen et al. [2023]	F0	MCD, F0 RMSE, MOS	LJSpeech (public), News data (-)	Vietnamese, English	Y	VITS
95	Kumar et al. [2022]	F0, energy	GPE, FFE, RMSE, MOS	VCTK (public), LibriTTS (public)	English	Y	NVS, TFSV-TTS, DeepVoice3
96	Xu et al. [2024]	F0	RMSE	Corpus spoken by an adult male (internal)	Mandarin Chinese	N	-
97*	Raitio et al. [2020]	Pitch, pitch range, phone duration, energy e spectral tilt	Means and standard deviations of the measured prosodic features, MOS	An internal dataset (internal), A Conversational expressive dataset (internal)	English	Y	Tacotron 2
98*	Ju et al. [2022]	Pitch	Mean value of pitch, MOS	LJSpeech (public), VCTK (public)	English	Y	FastPitch, Glow-TTS
99*	Łańcucki [2021]	Pitch	MOS	LJSpeech (public)	English	Y	FastPitch, Tacotron 2
100*	Ren et al. [2021a]	Pitch, duration, energy	MAE, MOS	LJSpeech (public)	English	Y	Tacotron 2, TransformerTTS, FastSpeech, FastSpeech 2