Emotion-Controllable Speech Synthesis Through Hierarchical Prosody Modeling and Multi-Level Acoustic Control

Triya Mishra
Department of Computer Science, University of Houston, Houston, TX, USA.
triyamishra72@uh.edu

Giacinto Tiepolo
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
giacinto.work@missouri.edu

Alessandra Almagia
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
alessandramail@uc.edu

Vadim Hanel
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
vadim.hanel334@unr.edu

Abstract

Emotion-controllable text-to-speech requires more than selecting a categorical speaking style; it requires coordinated control of prosodic phenomena that operate at different temporal scales. Utterance-level affect, phrase-level intonation, and syllable- or phone-level microprosody interact, and collapsing them into a single global embedding limits interpretability, controllability, and disentanglement from speaker identity. This manuscript develops a methodological framework for emotion-controllable speech synthesis that combines hierarchical prosody modeling with multi-level acoustic control. The framework factorizes prosody into utterance, phrase, and segment representations, estimates each representation under an explicit emotion condition, and fuses them at the acoustic decoder through level-specific gating. A conditional variational formulation captures utterance-level affective intent, a basis-function model represents phrase-level fundamental-frequency contours and energy, and a microprosody predictor controls local duration and segmental timing. Multi-level acoustic control is realized through latent variables that can be inspected, interpolated, and selectively strengthened during inference. Training objectives combine acoustic reconstruction, prosodic supervision, emotion prediction, disentanglement regularization, and stability constraints. The paper specifies inference procedures, control-strength interpolation, evaluation variables, and boundary conditions concerning data labeling, prosodic parsing, speaker leakage, and naturalness. The contribution is a coherent architecture and analytical account of how hierarchical prosody factorization can make emotional speech synthesis more controllable without treating emotion as an undifferentiated global attribute.

Keywords

emotional speech synthesis; hierarchical prosody; multi-level acoustic control; disentangled representation; expressive text-to-speech; prosodic modeling

1. Introduction

Text-to-speech synthesis has shifted from concatenative and statistical parametric approaches [1, 2] toward neural sequence-to-sequence systems that predict acoustic representations from text [3–5] and neural vocoders that reconstruct waveforms from those representations [6]. This shift improved naturalness and reduced the need for hand-engineered acoustic inventories. It also created a representation-learning problem: prosodic variation is no longer specified entirely by explicit parameters, and control must be imposed through conditioning, latent variables, or inference-time manipulation. Expressive synthesis therefore depends on whether the model can separate what is said, who says it, and how it is said.

Prosody transfer and controllable synthesis have addressed part of this problem by conditioning acoustic generation on reference prosody or latent style variables [7, 8]. Emotion is a particularly demanding form of control because it is not a single acoustic dimension. A perceived emotional state is realized through global tendencies in pitch range, speaking rate, and energy, through phrase-level intonation contours, and through local timing and voice-quality adjustments. Emotional speech synthesis has long been studied through rule-based and parametric approaches [9], and acoustic parameter sets provide compact summaries of relevant variation [10]. Emotional voice conversion research further shows that affect, speaker identity, and linguistic content are entangled in observable speech [11]. Surveys of affective speech synthesis and conversion emphasize the difficulty of obtaining labeled emotional data, the sensitivity of systems to recording conditions, and the gap between acted and spontaneous expression [12]. Unsupervised controllable synthesis attempts to reduce dependence on explicit labels by learning latent attributes [13], while multi-level style control in singing synthesis demonstrates that global and local style factors can be separated through hierarchical representations [14].

The unresolved design problem is how to expose emotion control at the same temporal granularity at which prosody is produced and perceived. A single utterance-level emotion embedding can shift the overall tone of a sentence, but it cannot independently specify a falling phrase-final contour, a locally stressed syllable, and a sustained increase in vocal effort without also altering speaker identity or unintentionally changing segmental duration. Conversely, purely frame-level conditioning can overfit to local acoustic details and produce unstable or diffuse control. The central research question is therefore: how can emotional speech synthesis be structured so that emotion conditions are factorized across prosodic timescales and converted into acoustic control signals that remain interpretable, stable, and partially disentangled from speaker and content?

This manuscript develops a framework that treats the question as a hierarchical representation and control problem. The proposed formulation factorizes prosody into utterance-level affective intent, phrase-level contour structure, and segment-level microprosody. Each level is conditioned on emotion and speaker variables, and each contributes to acoustic generation through level-specific gating. The multi-level acoustic controller allows control strength to be adjusted during inference without retraining, supports phrase-level emphasis and duration scaling, and provides an explicit mechanism for limiting speaker leakage. Training combines reconstruction, prosodic supervision, emotion

classification, adversarial disentanglement, and stability regularization. The framework is methodological and analytical: it specifies variables, objectives, inference rules, and evaluation logic, and it derives design consequences from the stated factorization rather than reporting experimental outcomes.

2. Related Work and Design Rationale

Neural TTS systems typically map text or phoneme sequences to mel spectrograms [3, 4] or to explicit duration and acoustic features [5]. Vocoders then generate waveforms from predicted acoustic representations [6]. These systems provide strong acoustic fidelity but weak direct control over prosodic structure unless prosody is represented in the conditioning signal. Prosody transfer work conditions generation on reference utterances and demonstrates that prosodic information can be transferred across speakers, but it also reveals leakage of speaker and recording characteristics [7]. End-to-end models with latent variables, such as conditional variational architectures, improve expressiveness by modeling uncertainty in acoustic generation [8]. The present framework builds on these ideas but does not treat style as a single latent code; it introduces explicit prosodic levels with distinct predictive roles.

Emotional speech synthesis has a longer history than neural TTS. Early reviews describe the use of prosodic rules for pitch, duration, and intensity, together with limitations in naturalness and emotional differentiation [9]. Acoustic parameter studies identify fundamental frequency, energy, spectral balance, and timing as core correlates of affect [10]. Emotional voice conversion research formalizes the problem as a transformation among affective states while preserving linguistic content and speaker identity, and it highlights the difficulty of separating emotion from speaker and channel factors [11]. More recent overviews of affective synthesis and conversion emphasize data scarcity, cultural variation, and evaluation validity [12]. These observations motivate a representation that does not force all emotional information through a single utterance embedding and that provides separate paths for global and local prosodic effects.

Controllable synthesis with unsupervised disentanglement seeks to discover interpretable attributes without exhaustive labels [13]. Such approaches are valuable when attribute annotations are unavailable, but they do not guarantee that discovered factors align with linguistically meaningful prosodic units. Multi-level style control in singing synthesis shows that style can be factorized into global and local components and that hierarchical conditioning can support transfer [14]. Spoken emotional speech differs in that prosody is tightly coupled to syntax, phrasing, and segmental timing; emotional expression is distributed across intonational phrases and stressed syllables rather than carried by musical structure alone. The proposed design therefore uses a linguistically motivated hierarchy rather than only a data-driven global-local split.

3. Hierarchical Prosody Modeling

3.1 Prosodic factorization

Let T denote the text or phoneme sequence, e an emotion condition that may be categorical or continuous, s a speaker or timbre condition, and X the acoustic observation sequence, such as mel-spectrogram frames with derived pitch and energy tracks. The hierarchical model introduces three prosodic latent variables: z_u for utterance-level affective intent, z_p for phrase-level contour structure, and z_s for segment-level microprosody. The acoustic likelihood is factorized so that each level contributes through its own conditional distribution and the decoder combines them through a shared acoustic observation model.

$$p(X|T,e,s) = \int p(X|z) \prod_l \, p(z_l|T,e,s) \, dz$$

Here z=(z_u,z_p,z_s), l ranges over the three prosodic levels, and the integral denotes marginalization over latent prosody. The factorization does not assert that the levels are statistically independent in natural speech; it defines a structured prior in which each level has a distinct conditioning role. Utterance-level variables capture broad affect and speaking-rate tendencies, phrase-level variables capture intonational movement across syntactic or breath-group boundaries, and segment-level variables capture local duration and stress realization. This structure supports inspection and intervention at the level where a control request is expressed.

3.2 Utterance-level emotion intent

The utterance level is modeled with a conditional variational autoencoder. A posterior encoder infers z_u from acoustic observations and the emotion condition, while a prior conditioned on emotion and speaker generates z_u at inference. The objective for this level follows an evidence lower bound with a capacity-controlling coefficient.

$$L_u = E_q[log \, p(X|z_u, e)] - \beta \, KL(q(z_u|X,e) \| p(z_u|e))$$

In this expression, q(z_u|X,e) is the approximate posterior, p(z_u|e) is the emotion-conditioned prior, β controls the pressure toward prior matching, and KL denotes Kullback-Leibler divergence. A larger β encourages a smoother utterance-level space that is easier to interpolate, but excessive pressure can remove affective distinctions that are acoustically subtle. The prior mean can be interpreted as the expected global realization of an emotion for a given speaker condition, while the posterior captures utterance-specific deviation. This level should not encode fine timing detail; otherwise it becomes a shortcut that absorbs phrase and segment variation that the controller is meant to expose.

3.3 Phrase-level contour modeling

Phrase-level prosody is represented as a contour over normalized time within an intonational or syntactic phrase. For phrase b, a basis-function expansion models the fundamental-frequency contour as a combination of smooth basis functions, an emotion-dependent offset, and a residual. The same structure can be applied to energy or to a normalized intensity track.

$$\hat{y}_b = \mu_b + \Sigma_{k=1}^{K} a_{b,k} \varphi_k(\tau_b) + c_b e + \varepsilon_b$$

Here ŷ_b is the predicted contour for phrase b, μ_b is a phrase baseline, φ_k are basis functions such as polynomials or discrete cosine components, a_{b,k} are predicted coefficients, τ_b is normalized position within the phrase, c_b is an emotion-conditioned coefficient vector, and ε_b is a residual. The basis expansion enforces smoothness while allowing phrase-level shape to vary with emotion. A control interface can modify selected coefficients, for example increasing final rising movement or flattening a phrase contour, without directly editing frame-level pitch values. Phrase boundaries must be supplied by a parser or predicted by a separate model; boundary errors are a principal failure mode because they shift the meaning of the contour coefficients.

3.4 Segment-level microprosody

The segment level controls local duration, short-range energy changes, and stress-related timing. Its inputs include phoneme identity, syllable position, lexical stress, phrase-level contour context, and the emotion condition. Unlike the phrase level, this level should not

carry long-range intonation trends; its role is to realize local prosodic adjustments that make an utterance sound emotionally coherent at the syllable and phone scale. A microprosody predictor can output duration multipliers and local energy offsets, with variance regularization to prevent unstable predictions around unvoiced segments or short function words. This level is where over-control is most audible: an excessive local pitch or duration perturbation can sound mannered even when the phrase contour is appropriate.

4. Multi-Level Acoustic Control

4.1 Gated fusion of prosodic levels

The acoustic decoder receives a sequence of hidden states derived from text and speaker conditions. At each frame t, a gating network assigns weights to the utterance, phrase, and segment representations. The gate is conditioned on the decoder state, the speaker or timbre variable, and the emotion embedding so that the contribution of each level can vary over time.

$$\alpha_\{t,m\} = softmax_m(w\text{^}T\ tanh(Wh_t + U_m\ z_m + Ve))$$

In this equation, h_t is the decoder hidden state, z_m is the representation for level m, e is the emotion embedding, W, U_m, and V are learned projections, w is a projection vector, and α_{t,m} is the normalized gate weight for level m at frame t. The softmax is taken over m∈{u,p,s}. The gate does not replace the latent variables; it controls how strongly each level influences the current acoustic prediction. This is important because phrase-final segments may require stronger phrase-level influence, while stressed vowels may require more segment-level control.

The fused control vector is then formed as a weighted combination of the level representations.

$$c_t = \Sigma_\{m \in \{u,p,s\}\}\ \alpha_\{t,m\}\ z_m$$

The vector c_t is supplied to the decoder through conditional modulation, for example by predicting scale and bias parameters for normalization layers or by concatenation with the decoder state. The acoustic decoder predicts mel-spectrogram frames and auxiliary prosodic targets, and a waveform generator reconstructs the signal. Neural vocoders provide one suitable waveform-generation path [15], while vocoder-based analyzers can supply pitch and aperiodicity features during data preparation [16]. The design keeps control at the latent and gate level rather than editing the waveform directly, which reduces the risk of discontinuities but requires the decoder to preserve the intended prosodic distinctions.

4.2 Control strength and interpolation

Inference-time control can be expressed as interpolation between a reference latent state and a target conditioned state. For each level m, a control coefficient κ determines how far the generated representation moves toward the requested emotion or prosodic target.

$$z_m\text{^}\{out\} = (1-\kappa)z_m\text{^}\{ref\} + \kappa\ z_m\text{^}\{cond\}$$

Here z_m^{ref} is a reference or neutral representation, z_m^{cond} is the representation produced under the target emotion condition, and κ∈[0,1] controls strength. A value of zero recovers the reference condition, and a value of one uses the target condition directly. The interpolation is most interpretable when the latent space is approximately smooth along the emotion direction; otherwise the perceptual effect of increasing κ may not be monotonic. Level-specific coefficients can be used, so that utterance-level affect is strengthened while

phrase-level contour changes remain moderate. This separation is useful when a user wants a clearly emotional tone without exaggerated intonation.

4.3 Disentanglement and speaker preservation

Emotion control becomes unreliable if the emotion latent also carries speaker identity, recording channel, or lexical content. The framework therefore includes an adversarial or orthogonality-based disentanglement term. A discriminator attempts to predict emotion from the prosodic representation while a speaker classifier is trained to predict speaker identity from a separate timbre representation. The intended effect is that emotion-conditioned prosody remains informative about affect but less predictive of speaker identity. This objective is a design pressure, not a guarantee; adversarial training can remove useful affective cues if the discriminator becomes too strong. The balance between emotion discriminability and speaker preservation must therefore be treated as a model-selection variable rather than a fixed architectural property.

5. Training, Inference, and Evaluation Protocol

5.1 Composite objective

Training combines several terms. Acoustic reconstruction encourages fidelity to the target mel spectrogram and auxiliary acoustic features. Prosodic supervision matches predicted contour coefficients, duration multipliers, and energy offsets to extracted targets. Emotion classification encourages the emotion-conditioned representation to retain affective information. Disentanglement reduces leakage of speaker and content factors. Stability regularization penalizes excessive variance in short segments and discourages abrupt gate changes.

$$L = L_rec + \lambda_p L_pros + \lambda_e L_emo + \lambda_d L_dis + \lambda_s L_stab$$

In this expression, L_rec is the acoustic reconstruction loss, L_pros is the prosodic supervision loss, L_emo is the emotion classification loss, L_dis is the disentanglement loss, L_stab is the stability regularizer, and λ_p, λ_e, λ_d, and λ_s set relative weights. The weights are not interchangeable: increasing λ_e can improve emotion classification while reducing naturalness if the classifier exploits channel artifacts, and increasing λ_d can reduce speaker leakage while also removing prosodic variation that is legitimately correlated with speaker anatomy.

The disentanglement term can be written with a discriminator D that receives a prosodic representation and an emotion condition.

$$L_dis = \mathrm{E}\left[\log D(z_u, e) + \log\left(1 - D(z_p, e)\right)\right]$$

Here D is trained to distinguish the utterance-level emotion representation from the phrase-level representation under the same emotion condition, and the expectation is over training examples. The precise form of D is less important than its role: it forces the model to distribute affective information across levels rather than collapsing all emotion into one latent variable. The loss must be combined with reconstruction and prosodic supervision; otherwise the model can satisfy the discriminator by discarding information that is needed for intelligible and natural speech.

5.2 Inference procedure

At inference, text is encoded into a phoneme or character sequence, and phrase boundaries are predicted or supplied. The emotion condition e and speaker condition s are used to sample or compute the utterance-level prior. Phrase-level coefficients are predicted for each phrase, and segment-level duration and energy controls are generated for each phoneme or syllable. The gating network combines the three representations at each decoder step. The decoder produces acoustic frames, and the vocoder generates the waveform. The user-facing controls are the emotion label or continuous affect vector, the level-specific interpolation coefficients κ_m, a phrase-level contour scaling factor, and a duration scaling factor. These controls are not independent in perception: a large duration scaling can change perceived emotion even when the emotion latent is fixed, so the interface should expose them as coordinated controls rather than as unrelated knobs.

5.3 Evaluation logic

A rigorous evaluation must separate naturalness, emotion accuracy, controllability, and disentanglement. Naturalness can be assessed with mean opinion scores and preference tests. Emotion accuracy can be measured with a held-out emotion classifier and with human listeners, but classifier accuracy alone is insufficient because a system may produce stereotyped or exaggerated cues that are easy to classify but poorly calibrated. Controllability should be tested by varying κ_m and measuring monotonic changes in intended acoustic correlates, such as pitch range, phrase-final slope, and duration. Disentanglement should be tested with probe classifiers for speaker identity and content from the prosodic representation, together with speaker-similarity judgments. Because emotional speech corpora differ in spontaneity, language, and label reliability, evaluation should include multiple datasets, such as dyadic interactive speech [17], multimodal acted speech [18], and German emotional speech [19]. Work on emotion recognition in human-computer interaction cautions that perceived emotion depends on context, culture, and task [20], so listener instructions and normalization procedures must be reported explicitly. Ablation conditions should remove the phrase level, the segment level, and the disentanglement term separately to identify which component supports which type of control.

6. Technical Discussion and Boundary Conditions

The hierarchical factorization offers interpretability at the cost of inductive bias. If phrase boundaries are incorrect, the phrase-level contour model can assign a rising contour to the wrong syntactic unit, and the decoder may compensate by producing unstable local pitch movement. If the utterance-level latent is too expressive, it can absorb phrase-level variation and reduce the effect of the lower-level controls. The gating network is intended to mediate this competition, but gating can also become degenerate if one level dominates across all frames. A useful diagnostic is to inspect gate entropy across phrases and phoneme classes; uniformly saturated gates indicate that the hierarchy has collapsed in practice.

The disentanglement objective introduces a second trade-off. Speaker identity and emotion are correlated in many corpora because speakers differ in expressive style, vocal range, and recording conditions. An adversarial term that is too aggressive can remove prosodic cues that listeners use to infer emotion, particularly breathiness, tension, and pitch range. A term that is too weak allows speaker leakage and makes emotion transfer sound like speaker conversion. The appropriate balance depends on whether the target application prioritizes preserving a specific speaker's voice or producing a clearly identifiable emotional style. The framework therefore treats disentanglement strength as a calibration parameter tied to the deployment context.

Control strength is also not indefinitely extensible. Interpolation in Eq. 6 assumes that the latent path between reference and target is approximately smooth. In practice, increasing κ can move the representation outside the region covered by the training distribution, producing exaggerated pitch excursions or unstable durations. Per-level coefficients and bounded κ values reduce this risk, but they do not eliminate it. A practical controller should expose a small number of coordinated presets, such as mild, moderate, and strong emotion, together with fine-grained phrase-level adjustments for expert users. This design keeps the control space interpretable while limiting the chance of pathological combinations.

Data conditions remain a fundamental boundary. Acted emotional corpora provide controlled labels but may exaggerate prototypical expression, while spontaneous corpora provide ecological validity but have noisy or ambiguous labels. Multilingual and cross-cultural evaluation is necessary because pitch range, speech rate, and intensity norms vary across languages and communities. The proposed framework does not resolve these data limitations; it provides a representation that makes them visible, because utterance-level, phrase-level, and segment-level effects can be evaluated separately rather than being confounded in a single style score. The multi-level style control perspective from singing synthesis [14] is relevant here, but spoken emotion requires additional attention to syntax, phrasing, and segmental timing. The resulting methodology is best understood as a controlled representation design for emotional TTS, not as a claim that any particular acoustic realization is universally perceived as a given emotion.

7. Conclusion

This manuscript presented a methodological framework for emotion-controllable speech synthesis through hierarchical prosody modeling and multi-level acoustic control. The framework factorizes prosody into utterance-level affective intent, phrase-level contour structure, and segment-level microprosody, and it combines these representations through conditioned gating at the acoustic decoder. A conditional variational formulation supports global emotion modeling, a basis-function representation supports phrase-level intonation control, and a microprosody predictor supports local timing and stress adjustments. The training objective integrates reconstruction, prosodic supervision, emotion classification, disentanglement, and stability regularization, while inference exposes control strength, phrase-level contour scaling, and duration scaling. The analytical contribution lies in making the control problem explicit at multiple prosodic timescales and in identifying the trade-offs that arise between controllability, naturalness, speaker preservation, and data coverage. The framework provides a coherent basis for evaluating whether hierarchical prosody factorization improves emotional TTS control and for diagnosing where leakage, gate collapse, or over-control occurs.

References

1. Taylor, P. (2009). Text-to-speech synthesis. Cambridge University Press.

2. Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical parametric speech synthesis. Speech Communication, 51(11), 1039–1064.

3. Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., & Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. In Proceedings of Interspeech 2017 (pp. 4006–4010).

4. Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R. J., Saurous, R. A., Agiomyrgiannakis, Y., & Wu, Y. (2018). Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4779–4783).

5. Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T. Y. (2021). FastSpeech 2: Fast and high-quality end-to-end text to speech. In International Conference on Learning Representations.

6. Kong, J., Kim, J., & Bae, J. (2020). HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In Advances in Neural Information Processing Systems (Vol. 33, pp. 17022–17033).

7. Skerry-Ryan, R. J., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., Weiss, R. J., Clark, R., & Saurous, R. A. (2018). Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron. In Proceedings of the 35th International Conference on Machine Learning (pp. 4693–4702).

8. Kim, J., Kong, J., & Son, J. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In Proceedings of the 38th International Conference on Machine Learning (pp. 5530–5540).

9. Schröder, M. (2001). Emotional speech synthesis: A review. In Proceedings of Eurospeech 2001 (pp. 561–564).

10. Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., & Truong, K. P. (2016). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing, 7(2), 190–202.

11. Zhou, K., Sisman, B., Liu, R., & Li, H. (2022). Emotional voice conversion: Theory, databases and ESD. Speech Communication, 137, 1–18.

12. Triantafyllopoulos, A., Schuller, B. W., Iymen, G., Sezgin, M., He, X., Yang, Z., Tzirakis, P., & Tao, J. (2023). An overview of affective speech synthesis and conversion in the deep learning era. Proceedings of the IEEE, 111(10), 1355–1381.

13. Henter, G. E., Lorenzo-Trueba, J., Wang, X., & Yamagishi, J. (2018). Deep encoder-decoder models for unsupervised learning of controllable speech synthesis. arXiv preprint arXiv:1807.11470.

14. Zhang, Y., Jiang, Z., Li, R., Pan, C., He, J., Huang, R., ... & Zhao, Z. (2024, November). Tcsinger: Zero-shot singing voice synthesis with style transfer and multi-level style control. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 1960-1975).

15. van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.

16. Morise, M., Yokomori, F., & Ozawa, K. (2016). WORLD: A vocoder-based high-quality speech synthesis system for real-time applications. IEICE Transactions on Information and Systems, 99(7), 1877–1884.

17. Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42(4), 335–359.

18. Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE, 13(5), e0196391.

19. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. In Proceedings of Interspeech 2005 (pp. 1517–1520).

20. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. (2001). Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine, 18(1), 32–80.