# Cross-Lingual Singing Voice Synthesis Using Multilingual Acoustic Modeling and Deep Learning

Tymon Szyca
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
tymon.work@binghamton.edu

Luciano Gotthard
Department of Computer Science, University of Houston, Houston, TX, USA.
lucianogotthard@uh.edu

Tola Basa
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
basa1983@buffalo.edu

Tonia Flaiano
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
flaiano1992@colostate.edu

**Abstract**

Cross-lingual singing voice synthesis aims to generate a sung performance in a target language from a musical score and lyrics while preserving melody, rhythm, singer identity, and expressive detail. Most neural singing systems are trained within a single language and entangle language-specific phonotactics with acoustic modeling, so pronunciation transfer can cause phoneme substitution errors, duration drift, pitch instability, and timbre degradation. This paper develops a methodological framework based on multilingual acoustic modeling and deep learning. A score encoder represents pitch, duration, and expression; a language-conditioned phoneme encoder maps lyrics into a shared articulatory space; a transformer backbone predicts acoustic frames; and an adversarial language classifier encourages language-invariant hidden representations while preserving singer-dependent timbre. Cross-lingual alignment is handled by monotonic attention constrained by score timing, and duration and pitch predictors provide explicit prosodic conditioning. Objective functions combine spectral reconstruction, duration, pitch, alignment, and language-invariance terms. The analysis explains why shared acoustic representations can reduce negative transfer across languages but also why they risk collapsing phonemic contrasts that are meaningful in only one language. It discusses boundary conditions involving unseen phonemes, tone languages, low-resource languages, singer leakage, and vocoder artifacts. The contribution is a coherent design rationale and analytical account rather than a claim of empirical superiority. The framework offers a technically grounded basis for building multilingual singing systems, organizing future corpora, and evaluating cross-lingual transfer with metrics that separate pronunciation accuracy from musical and timbral fidelity.

**Keywords**

cross-lingual singing voice synthesis; multilingual acoustic modeling; score-conditioned attention; language-adversarial training; neural vocoding; singing corpus

## 1. Introduction

Singing voice synthesis converts a symbolic score and lyrics into a sung acoustic signal. The task differs from ordinary text-to-speech because the acoustic realization must satisfy note-level pitch, duration, phrasing, and expressive timing, while also rendering intelligible lyrics. Deep neural approaches changed the field by learning mappings from score and phoneme features to acoustic parameters with less hand-engineered parametric control than earlier statistical systems [1]. Subsequent end-to-end and attention-based speech synthesis architectures showed that sequence-to-sequence learning can align text and acoustics without an external duration model, and transformer and feed-forward designs improved stability and controllability [2][3][4]. Neural vocoders then made it possible to render high-fidelity waveforms from intermediate acoustic representations [5]. These advances provide the technical vocabulary for singing synthesis, but they do not by themselves solve the cross-lingual problem.

The central difficulty is that a singer's voice and a lyric language are not independent. A model trained on one language learns phoneme categories, syllable structure, vowel space, consonantal timing, and coarticulation patterns specific to that language. When the same model is asked to sing lyrics in another language, the acoustic decoder may substitute familiar phonemes for unseen ones, stretch or compress vowels according to the wrong metrical habits, or distort pitch accents because the training language did not use the same tonal or stress system. Multilingual acoustic modeling offers a partial solution by sharing hidden representations across languages, but sharing creates a second risk: the model may erase contrasts that are phonemic in one language while irrelevant in another. Cross-lingual singing synthesis therefore requires a deliberate balance between language-invariant acoustic structure and language-specific phonological detail.

This paper develops a unified methodological framework for cross-lingual singing voice synthesis using multilingual acoustic modeling and deep learning. The contribution is not an empirical benchmark claim but a technically substantive design account: it specifies how a score-conditioned acoustic model, a shared multilingual phoneme space, monotonic alignment, prosodic conditioning, adversarial language regularization, and neural waveform generation can be organized into one coherent system. The framework identifies where language transfer should occur, where it must be constrained, and what failure modes follow from each design choice. It also clarifies the evaluation logic implied by the architecture, separating pronunciation accuracy from musical fidelity and singer timbre.

The remainder of the paper is organized as follows. Section 2 defines the synthesis problem and system overview. Section 3 develops the multilingual acoustic representation. Section 4 analyzes cross-lingual alignment and score-conditioned prosody. Section 5 discusses neural waveform generation and adversarial learning. Section 6 formalizes the training objectives. Section 7 examines boundary conditions and analytical trade-offs, and Section 8 concludes.

## 2. Problem Formulation and System Overview

A singing score can be represented as a sequence of note and phoneme events. For each lyric unit t, let p_t denote the phoneme identity, d_t the notated duration, f_t the fundamental frequency or pitch target, and v_t an expression descriptor such as legato, accent, or vibrato propensity. Let L denote the lyric sequence, s a singer identity embedding, and l the language of the lyrics. The singer may have a different native language from l; cross-lingual synthesis

occurs when the lyric language is not the language that dominates the singer's training data or the model's phoneme inventory. The system produces a waveform x by composing an acoustic model A_θ and a waveform generator G_φ:

$$x = G_{\varphi}\big(A_{\theta}(S, L, s, l)\big)$$

Here S denotes the score representation containing pitch, duration, and expression features; θ and φ are learned parameters; A_θ predicts an intermediate acoustic representation such as a mel-spectrogram or a vocoder-oriented latent sequence; and G_φ converts that representation into a waveform. The score S supplies timing and pitch constraints that are stronger than those in ordinary speech synthesis. The lyrics supply phoneme identities and syllable structure, while l conditions the linguistic front end and the speaker embedding s conditions timbre and singing style.

The acoustic model follows a sequence-to-sequence design with explicit score conditioning. A lyric encoder converts phoneme symbols into contextual embeddings, a score encoder projects note-level features into a continuous representation, and a transformer backbone fuses the two sequences. The transformer is appropriate because attention can relate distant lyric units to local acoustic frames while retaining parallel computation during training [4]. A duration predictor converts phoneme-level timing into frame-level boundaries, and a pitch predictor models the fundamental-frequency contour. The decoder predicts acoustic frames, and the neural vocoder renders the waveform. This division is consistent with end-to-end speech synthesis systems that separate acoustic prediction from waveform generation [2], with feed-forward duration-modeled architectures that improve controllability [3], and with vocoder designs that recover phase and high-frequency detail from acoustic features [5].

The cross-lingual requirement changes the role of each component. The lyric encoder must be multilingual rather than a language-specific lookup table. The score encoder remains largely language-neutral, although expressive timing may vary by musical tradition. The acoustic decoder must share parameters across languages to benefit from common articulatory and phonatory structure, but it must not force every language into the same phonetic contrasts. The speaker embedding must represent singer identity without becoming a proxy for the training language. These requirements motivate the multilingual acoustic model and the regularization strategy developed in the following sections.

## 3. Multilingual Acoustic Modeling

A multilingual singing system needs a phoneme representation that is shared enough to transfer learning but granular enough to preserve language-specific distinctions. One common strategy is to map each language's phoneme inventory into a union inventory, with language embeddings indicating the source system. Another strategy uses articulatory features such as place, manner, voicing, vowel height, and backness, which can generalize to unseen phonemes. Singing-specific systems have shown that acoustic models can learn expressive timbre and pitch behavior from parallel score and audio data [6], that web-mined singing data can support synthesis when carefully filtered [7], and that diffusion-based acoustic generators can model complex singing distributions [8]. These systems are mostly trained within one language, however, and their front ends assume a fixed phoneme set. Multilingual representation learning from speech offers a useful analogy because shared subword and acoustic representations can transfer across languages when the training signal is large and diverse [9]. For singing, transfer must additionally preserve pitch and timing, so the shared representation cannot be purely lexical.

The proposed phoneme encoder combines a language embedding, a phoneme embedding, and a tone or stress embedding:

$$e_t = E_lang(l) \oplus E_ph(p_t) \oplus E_tone(\tau_t)$$

In this expression, e_t is the input embedding for lyric unit t; E_lang(l) is a learned vector for lyric language l; E_ph(p_t) is the phoneme embedding for p_t; E_tone(τ_t) encodes tone, stress, or lexical accent category; and $\oplus$ denotes concatenation followed by a projection, or a gated fusion when dimensions differ. The language embedding allows the model to adjust pronunciation without duplicating the entire phoneme table. The tone variable is important because languages differ in whether pitch is lexical, morphological, or purely expressive. For a non-tone language, τ_t can represent stress or be set to a neutral category.

This formulation has three consequences. First, shared phoneme embeddings encourage acoustically similar phones across languages to occupy nearby regions, which helps low-resource languages benefit from high-resource languages. Second, the language embedding can shift the interpretation of a shared symbol when the same phone has different allophonic behavior across languages. Third, the tone embedding separates linguistic pitch from musical pitch; the score supplies the sung fundamental frequency, while τ_t supplies phonological pitch information that may affect vowel quality, duration, or onset timing. A global multi-technique singing corpus with realistic music scores can provide the kind of multilingual, score-aligned material needed to instantiate such a representation across languages and singing techniques [10]. The corpus is relevant as a resource design reference, not as evidence that the present architecture has been trained or validated.

The multilingual acoustic model should not be interpreted as a single averaged voice. Singer identity is represented by s, which conditions normalization layers, decoder affine transforms, or a learned embedding added to hidden states. Language identity is represented by l, which conditions the phoneme encoder and may also modulate the duration and pitch predictors. The model therefore has two conditioning paths: one for who is singing and one for what language is being sung. Keeping these paths separate is essential because a singer's timbre should remain stable when the lyric language changes, while pronunciation should change according to l. If singer and language are entangled during training, the model may learn to associate a particular language with a particular voice and fail when the two are recombined.

## 4. Cross-Lingual Alignment and Score-Conditioned Prosody

Singing alignment is more constrained than speech alignment because note boundaries and lyric syllables impose a partial temporal ordering. Attention-based sequence models can learn soft alignments between input units and acoustic frames [11], and recurrent encoder-decoder models established the basic conditional generation framework on which many later systems build [12]. In singing, however, free attention can produce skipped phonemes, repeated syllables, or drift across long notes. The score provides an external timing prior that should be used to restrict attention. Contextual language representations can also improve the lyric encoder by disambiguating polyphonic characters, morphophonological alternations, and language-specific grapheme-to-phoneme mappings [13]. Multilingual speech corpora demonstrate that language diversity improves shared acoustic representations when data are carefully balanced [14], but singing requires additional score-level supervision.

The alignment mechanism can be written as a monotonic attention operation. For acoustic frame t and lyric unit u, let q_t be the frame query, k_u and v_u the lyric key and value, and d the attention dimension. The context vector c_t is:

$$c_t = \sum_u \alpha_t,u \, v_u, \alpha_t,u = softmax_u(q_t \, ^{T}k_u/\sqrt{d})$$

Here α_t,u is the attention weight of lyric unit u at frame t, and softmax_u normalizes over lyric positions. The attention is masked so that α_t,u is zero when u lies outside a score-derived monotonic window around the current frame. This constraint reflects the fact that singers may hold a vowel, ornament a note, or insert breath, but they do not normally reorder syllables arbitrarily. The mask should be wide enough to permit expressive micro-timing and narrow enough to prevent global misalignment. The score-derived window is computed from notated durations, tempo, and a learned tempo-scaling factor.

After alignment, the acoustic hidden state is conditioned on score and singer variables. Let $\tilde{h}$_t be the pre-prosodic hidden state at frame t, d_t the duration feature, f_t the pitch feature, v_t the expression feature, and s the singer embedding. The conditioned state is:

$$h_t = LayerNorm(\tilde{h}_t + MLP([d_t, f_t, v_t, s]))$$

The MLP is a small feed-forward network, and the brackets denote concatenation. Layer normalization stabilizes the fusion of heterogeneous features. The duration feature tells the model whether the current phoneme is sustained, the pitch feature supplies the target fundamental frequency, and v_t carries expressive annotations such as accent or vibrato tendency. The singer embedding s allows the same score to be rendered with different vocal timbres. This equation is not a claim that prosody is fully determined by these variables; rather, it makes explicit the conditioning pathway through which musical and speaker information enters the acoustic decoder.

Frame-level timing is obtained by expanding phoneme durations at the vocoder frame rate. Let r be the frame rate in frames per second and d_j the duration of phoneme j in seconds. The cumulative boundary n_t for phoneme t is:

$$n_t = round(r \sum_\{j \le t\} d_j)$$

The rounding operator converts cumulative time into an integer frame index, and d_j is constrained to be nonnegative. This equation connects score timing to the frame sequence used by the acoustic model. In cross-lingual settings, the duration predictor must be regularized so that it does not simply transfer the duration statistics of the singer's dominant language to the target language. Vowel lengthening, consonant cluster timing, and phrase-final lengthening differ across languages and musical styles; the model should therefore condition duration on l as well as on local phoneme context.

**5. Neural Waveform Generation and Adversarial Learning**

The acoustic model produces an intermediate representation, but the final perceptual quality depends on waveform generation. Autoregressive neural waveform models such as WaveNet demonstrated that raw audio can be generated with high fidelity by modeling the conditional distribution of each sample given previous samples [15]. Autoregressive generation is computationally expensive for long singing phrases, so parallel and adversarial vocoders are often preferred. Multi-singer singing vocoders aim to preserve singer identity while synthesizing high-fidelity audio from acoustic features [16]. Generative adversarial networks provide a general training principle in which a generator and discriminator compete, allowing the generator to produce sharper and more realistic outputs than a purely likelihood-

based objective may yield at practical model sizes [17]. In singing synthesis, adversarial vocoding can improve the rendering of breathiness, vibrato micro-modulation, and high-frequency harmonics, but it can also introduce instability, metallic artifacts, or uneven singer timbre if the discriminator learns language- or singer-specific shortcuts.

A technically defensible vocoder design for cross-lingual singing should therefore separate acoustic content from singer style. The vocoder receives the predicted acoustic representation, a singer embedding, and possibly a language-invariant prosody code. The discriminator should be trained to distinguish real from generated audio while being discouraged from classifying the underlying language. If the vocoder can infer language from singer identity, it may learn to distort target-language phonemes toward the training language of that singer. Conversely, if the vocoder is forced to ignore singer identity completely, it may lose timbre detail. The appropriate constraint is partial: preserve singer-dependent spectral envelope and vocal tract characteristics, but avoid language-dependent pronunciation shortcuts. This is why the adversarial principle is useful but must be paired with the language-invariance objective described next.

**6. Objective Functions and Training Strategy**

The acoustic model is trained with a composite objective. Let L_mel be the spectral reconstruction loss between predicted and reference acoustic frames, L_dur the duration prediction loss, L_pitch the pitch contour loss, L_align a monotonic alignment loss, and $||\theta||^2$ a parameter regularization term. The total loss is:

$$L = L_mel + \lambda 1 L_dur + \lambda 2 L_pitch + \lambda 3 L_align + \lambda 4 ||\theta||^2$$

The coefficients λ1 through λ4 control the relative influence of each term. L_mel ensures acoustic fidelity; L_dur prevents the duration predictor from collapsing to average timing; L_pitch constrains the fundamental-frequency contour; L_align penalizes attention weights outside the score-derived monotonic window; and the final term limits overfitting. The weights are training hyperparameters, not empirically observed values in this analysis. A practical training schedule would increase the alignment and pitch terms after an initial acoustic-only warm-up, because premature hard alignment can prevent the model from learning useful contextual representations.

Language invariance is encouraged by an auxiliary classifier. Let T be the number of acoustic frames, K the number of languages, z_t the true language labels, and D_k(h_t) the predicted probability that frame t belongs to language k under a classifier D. The classifier loss is:

$$L_lang = -\left(\frac{1}{T}\right)\sum_{t=1}^{T}\sum_{k=1}^{K} 1[z_t = k]\log D_k(h_t)$$

The indicator function 1[z_t=k] selects the true language. During training, a gradient reversal layer is placed between the acoustic hidden state and the language classifier, so that the classifier learns to predict language while the acoustic model learns to make language identity less discriminable. The strength of this reversal must be moderate. If the reversal is too weak, the model continues to rely on language-specific cues and transfers poorly. If it is too strong, the model erases phonemic contrasts, causing target-language vowels to merge with dominant-language categories. The objective should therefore be applied selectively to the upper acoustic layers rather than to the phoneme embedding itself, because the lower layers must preserve phonetic detail.

The composite objective also implies a data balancing requirement. If one language dominates the training corpus, the shared acoustic parameters will be optimized primarily for that language, and the language-adversarial loss may simply suppress minority-language cues rather than learn a genuinely shared representation. Balanced sampling, language-specific batch normalization statistics, and per-language validation are therefore not optional refinements but structural conditions for the framework. The model should be evaluated on intelligibility in the target language, pitch and duration accuracy on held-out scores, and singer similarity across languages. These three dimensions answer different questions and should not be collapsed into a single aggregate score.

## 7. Boundary Conditions and Analytical Assessment

Several boundary conditions limit the proposed framework. The first is unseen phonemes. A shared phoneme inventory can represent many cross-lingual correspondences, but languages contain phones and allophones absent from the training inventory. Articulatory feature decomposition and grapheme-to-phoneme fallback can reduce errors, but they cannot guarantee correct pronunciation for an unseen contrast. The second is tone. In tone languages, lexical pitch and musical pitch interact. The model must distinguish score-supplied fundamental frequency from phonological tone; otherwise, a rising musical phrase may be misread as a rising lexical tone or vice versa. The tone embedding in Equation 2 is designed for this separation, but the training data must contain enough tonal variation for the embedding to be meaningful.

The third boundary condition is low-resource transfer. Shareable parameters help only when the target language has enough data to prevent the shared representation from becoming a proxy for the dominant language. Adversarial invariance can worsen this problem if it removes cues that are rare but phonemically essential. A safer strategy is to increase the capacity of language-specific adapters for low-resource languages while keeping the core acoustic model shared. The fourth condition is singer leakage. If the same singer appears only in one language, the model may learn an association between timbre and language. Singer-balanced multilingual data and speaker-conditioned normalization reduce this risk, but evaluation must include cross-combinations of singer and language that were not observed during training.

The fifth condition is vocoder artifact. A vocoder trained mostly on one language may reproduce that language's vowel spectra and consonant noise patterns even when the acoustic model predicts target-language features. Perceptual evaluation should therefore include native-speaker intelligibility judgments, not only spectral distance or mean opinion scores from listeners unfamiliar with the target language. Neural speech synthesis surveys emphasize that objective and subjective measures capture different aspects of quality [18], and singing adds musical dimensions that require separate assessment. Pronunciation accuracy should be measured with phoneme error rate or word error rate from a target-language recognizer where available; musical fidelity should be measured with pitch and duration error on aligned notes; and timbre should be measured with singer similarity or discrimination tests. No single metric can establish cross-lingual success.

The framework also has an important analytical trade-off. A fully shared acoustic model maximizes transfer but risks phonetic collapse. A fully language-specific model avoids collapse but loses the benefit of multilingual data. The proposed combination of shared phoneme embeddings, language embeddings, selective adversarial regularization, and language-conditioned duration and pitch predictors occupies an intermediate position. It

assumes that articulatory and phonatory structure is partly universal while phonological categorization is partly language-specific. That assumption is plausible for singing because sustained vowels and pitch production share physiological constraints across languages, whereas syllable structure and tone assignment do not. The design is therefore not a generic multilingual model applied to singing; it is a singing-specific organization of shared and language-specific factors.

## 8. Conclusion

Cross-lingual singing voice synthesis requires more than translating lyrics before synthesis. The acoustic model must learn a shared articulatory and phonatory space, preserve language-specific phonological contrasts, respect score-imposed pitch and timing, and maintain singer identity across languages. This paper has developed a methodological framework that addresses these requirements through multilingual acoustic modeling, score-conditioned transformer alignment, explicit duration and pitch conditioning, language-adversarial regularization, and neural waveform generation. The equations formalize the synthesis mapping, phoneme embedding composition, monotonic attention, prosodic conditioning, frame expansion, composite training loss, and language classifier loss. The analysis explains why sharing across languages can reduce negative transfer but also why unconstrained invariance can erase meaningful phonetic detail.

The practical implication is that cross-lingual singing systems should be designed around separable factors. Singer identity, lyric language, score timing, and phonological tone should enter the model through distinct conditioning paths, and evaluation should report intelligibility, musical accuracy, and timbre preservation separately. Future corpus design should include balanced singer-language combinations, realistic score alignments, and explicit metadata for tone, technique, and expression. The framework presented here provides a technically grounded basis for such systems and for subsequent empirical work that tests these design choices under controlled conditions.

## References


1. Nishimura, M., Hashimoto, K., Oura, K., Nankaku, Y., & Tokuda, K. (2016). Singing voice synthesis based on deep neural networks. In Proceedings of Interspeech (pp. 2478–2482). ISCA.

2. Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., ... & Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. In Proceedings of Interspeech (pp. 4006–4010). ISCA.

3. Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T. Y. (2019). FastSpeech: Fast, robust and controllable text to speech. In Advances in Neural Information Processing Systems (Vol. 32). Curran Associates.

4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30, pp. 5998–6008). Curran Associates.

5. 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). Curran Associates.

6. Lu, P., Wu, Y., & Lian, J. (2020). XiaoiceSing: A high-quality and integrated singing voice synthesis system. In Proceedings of Interspeech (pp. 1306–1310). ISCA.

7. Ren, Y., Tan, X., Qin, T., Luan, J., Zhao, S., Zhao, Z., & Liu, T. Y. (2020). DeepSinger: Singing voice synthesis with data mined from the web. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 1979–1989). ACM.

8. Liu, J., Li, C., Ren, Y., Chen, F., & Zhao, Z. (2022). DiffSinger: Singing voice synthesis via shallow diffusion mechanism. Proceedings of the AAAI Conference on Artificial Intelligence, 36(10), 11020–11028.

9. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 8440–8451). ACL.

10. Zhang, Y., Pan, C., Guo, W., Li, R., Zhu, Z., Wang, J., ... & Zhao, Z. (2024). Gtsinger: A global multi-technique singing corpus with realistic music scores for all singing tasks. Advances in Neural Information Processing Systems, 37, 1117-1140.

11. Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Proceedings of the 3rd International Conference on Learning Representations. ICLR.

12. Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (pp. 1724–1734). ACL.

13. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (pp. 4171–4186). ACL.

14. Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., & Collobert, R. (2020). MLS: A large-scale multilingual dataset for speech research. In Proceedings of Interspeech (pp. 2757–2761). ISCA.

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

16. Huang, R., Chen, F., Ren, Y., Liu, J., Cui, C., & Zhao, Z. (2021). Multi-Singer: Fast multi-singer singing voice vocoder with a large-scale corpus. In Proceedings of the 29th ACM International Conference on Multimedia (pp. 3945–3954). ACM.

17. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems (Vol. 27, pp. 2672–2680). Curran Associates.

18. Tan, X., Qin, T., Soong, F., & Liu, T. Y. (2021). A survey on neural speech synthesis. arXiv preprint arXiv:2106.15561.