Robust Speech Reconstruction Using Probabilistic Acoustic Modeling and Reference-Driven Feature Adaptation

Nora Madigan
School of Computing, Clemson University, Clemson, SC, USA.
madigan879@clemson.edu

Leif Schinke
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
leifschinke618@uab.edu

Cornelis Scardino
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
cornelismail@ku.edu

Abstract

Speech reconstruction from degraded observations remains difficult because acoustic evidence is sparse, channel distortions are nonstationary, and deterministic regressors provide point estimates without reliable uncertainty. This manuscript develops a probabilistic acoustic modeling framework with reference-driven feature adaptation for robust speech reconstruction. The observed signal is modeled as a noisy realization of a latent acoustic state under a conditional Gaussian likelihood, while a temporal prior captures continuity across frames. Amortized variational inference estimates the posterior over latent states, separating aleatoric observation uncertainty from epistemic uncertainty in the acoustic state. A reference encoder converts an enrolled utterance or domain prototype into a low-rank adaptation of latent features, and a covariance-weighted fusion rule combines reference-conditioned and evidence-driven estimates. The resulting reconstruction operator is uncertainty aware: it trusts the reference when acoustic evidence is weak and trusts the observed signal when the reference is mismatched. We derive the variational objective, adaptation operator, fusion rule, and training loss, and analyze failure modes under model mismatch, reference contamination, and nonstationary noise. The framework provides a principled route to robust speech reconstruction without assuming that clean targets are available at inference. Its significance lies in linking probabilistic uncertainty, temporal acoustic structure, and reference-conditioned adaptation within a single inferential model, thereby giving reconstruction decisions an explicit statistical interpretation rather than relying on an uncalibrated regression output.

Keywords

probabilistic acoustic modeling; speech reconstruction; reference-driven adaptation; variational inference; uncertainty calibration; robust feature fusion

1. Introduction

Speech reconstruction seeks to recover intelligible and acoustically faithful speech from observations corrupted by noise, reverberation, clipping, or channel filtering. The task differs from generic denoising because reconstruction must preserve phonetic identity, temporal fine structure, and speaker-dependent spectral detail while suppressing distortion. Classical speech enhancement methods provide a mature statistical foundation, including spectral subtraction and minimum-mean-square-error amplitude estimation [1], [3]. Supervised separation and enhancement systems later replaced hand-designed priors with learned mappings from degraded to clean acoustic features, and large-scale training made such mappings effective across many noise conditions [2]. Nevertheless, point-estimate regression remains vulnerable when the test distribution differs from the training distribution, because a single output cannot express ambiguity about whether a spectral component is speech, noise, or an artifact of channel filtering.

Probabilistic acoustic modeling addresses this limitation by representing the clean acoustic state as a latent variable and treating the observed signal as a stochastic realization conditioned on that state. Variational autoencoders provide a practical inference machinery for such models [4], while autoregressive and diffusion-based generative models demonstrate the value of expressive likelihoods for raw audio and spectral structure [5], [6]. These approaches can represent multimodal uncertainty, but their reconstruction behavior depends critically on the prior and on the inference network. In degraded conditions, the posterior can become diffuse or biased, and a generative model may hallucinate plausible acoustic detail that is not supported by the observation. A robust reconstruction framework therefore needs both a structured likelihood and an explicit mechanism for incorporating auxiliary acoustic evidence without overriding the observed signal.

Self-supervised speech representations offer one source of such evidence. Models trained by contrastive or masked prediction learn features that are relatively stable across recording conditions and useful for downstream recognition [7], [8]. Large-scale weakly supervised recognition systems further show that broad acoustic coverage can improve robustness to domain shift [9]. These representations are not reconstruction models by themselves, but they can supply reference statistics that describe speaker, channel, or domain characteristics. The central problem is how to adapt a probabilistic acoustic model with such reference statistics while preserving calibrated uncertainty and avoiding negative transfer when the reference is mismatched.

This manuscript develops a unified framework for robust speech reconstruction that couples probabilistic acoustic modeling with reference-driven feature adaptation. The contribution is a latent-variable reconstruction model in which the observation likelihood is conditionally Gaussian, the latent state evolves under a temporal prior, and an amortized variational encoder estimates a posterior over clean acoustic states. A reference encoder maps an enrolled utterance or domain prototype to a low-rank transformation of latent features. The transformed estimate is then fused with the evidence-driven estimate by covariance-weighted inference, so that the influence of the reference is governed by estimated reliability rather than by a fixed interpolation weight. We derive the observation model, variational objective, temporal prior, adaptation operator, fusion rule, and training loss, and we analyze the conditions under which reference adaptation improves robustness or introduces systematic bias. The framework is analytical and design-oriented: it specifies a coherent inferential

model and its consequences without asserting empirical performance that has not been established.

2. Problem Formulation and Acoustic Modeling Assumptions

Let x_t denote an observed acoustic feature vector at frame t, and let z_t denote a latent clean acoustic state that describes spectral envelope, excitation characteristics, and short-time phonetic context. The conditioning variable s_t encodes available side information such as phonetic context, speaker identity, or an estimated noise class. A reference variable r_t represents auxiliary evidence derived from an enrolled utterance, a domain prototype, or a reference feature bank associated with the target condition. The reconstruction target is the clean acoustic state or a deterministic acoustic function of it, not the observed degraded feature itself.

The observation process is modeled as a conditional transformation of the latent state with additive residual uncertainty:

$$x_t \;=\; g_\theta(z_t, s_t) \;+\; \varepsilon_t, \varepsilon_t \sim N(0, \Sigma_t)$$

Here g_θ is a parameterized degradation mapping, θ collects its parameters, and ε_t is a zero-mean residual with covariance Σ_t. The covariance is not a nuisance term; it represents aleatoric uncertainty caused by noise, reverberation, quantization, and unmodeled channel effects. A large Σ_t indicates that the observed frame provides weak evidence about z_t, whereas a small covariance indicates that the observation is informative. This assumption is appropriate for short-time spectral features because additive acoustic interference and channel perturbations can often be approximated as conditionally Gaussian residual terms within a local time-frequency region, although the approximation is weaker for impulsive or highly nonstationary distortions.

The model makes three explicit assumptions. First, given the latent state and conditioning variables, observation residuals are conditionally independent across frames. Second, the latent state evolves smoothly enough that a first-order temporal prior is informative, but the prior is not so rigid that it suppresses genuine phonetic transitions. Third, the reference variable is informative about the target condition but is not assumed to be a clean copy of the target utterance. The third assumption is essential: reference-driven adaptation must remain valid when the reference shares speaker or channel characteristics with the target but differs in phonetic content.

3. Probabilistic Acoustic Modeling Framework

3.1 Latent Acoustic State and Generative Model

The generative model factorizes the joint distribution of observations and latent states across time. For a sequence of length T, the factorization is

$$p(x_{1:T}, z_{1:T}|s_{1:T}) = \prod_{t=1}^{T} p(x_t|z_t, s_t)p(z_t|z_{t-1})$$

where x_{1:T} and z_{1:T} denote observation and latent-state sequences, s_{1:T} denotes conditioning variables, and p(z_1 | z_0) is interpreted as the initial-state prior. The emission term p(x_t | z_t, s_t) is the Gaussian likelihood implied by the observation model, and p(z_t | z_{t-1}) is a temporal transition prior. The latent state is intended to capture acoustic factors that persist across frames, such as vocal-tract configuration and speaker-

dependent spectral bias, while the conditioning variable captures factors that change rapidly, such as phonetic identity.

This factorization is useful because it separates two sources of uncertainty. Observation uncertainty enters through the emission covariance Σ_t, while state uncertainty enters through the posterior over z_t. A reconstruction system that ignores the latter cannot distinguish between a frame that is confidently observed and a frame whose acoustic content is ambiguous. The latent-state formulation also permits reference adaptation to act on z_t rather than directly on the observed features, which reduces the risk that a reference transformation will overwrite evidence contained in x_t.

3.2 Variational Inference and Uncertainty

Exact posterior inference over z_{1:T} is intractable for nonlinear degradation mappings. We therefore use an amortized variational distribution q_φ(z_{1:T} | x_{1:T}, s_{1:T}) parameterized by φ and optimize the evidence lower bound. For a single sequence, the bound has the form

$$L(\theta, \varphi) = E_{q_\varphi(z|x,s)}[\log p_\theta(x|z,s)] - KL\big(q_\varphi(z|x,s)||p(z|s)\big)$$

where E_{q_φ} denotes expectation under the variational posterior, log p_θ(x|z,s) is the reconstruction log-likelihood, and KL denotes the Kullback–Leibler divergence between the variational posterior and the prior. The first term rewards latent states that explain the observed features; the second term regularizes the posterior toward the temporal prior. Standard variational inference theory justifies this objective as a lower bound on the marginal likelihood and clarifies the approximation introduced by the chosen posterior family [10].

The posterior covariance provides a frame-level measure of epistemic uncertainty. When the observed frame is weakly informative, the likelihood term is flat with respect to z_t, and the posterior remains close to the prior. When the observed frame is strongly informative, the likelihood concentrates the posterior around states that reproduce the observation. This behavior is desirable for robust reconstruction because it allows later adaptation and fusion stages to treat uncertain frames differently from confident frames. The variational posterior also supports sampling-based reconstruction, in which multiple plausible clean states can be generated and summarized by their posterior mean and covariance.

3.3 Temporal Smoothing and Robustness

The temporal prior supplies continuity across frames and prevents isolated noisy observations from dominating the reconstruction. A tractable linear Gaussian transition model is

$$z_t \;=\; A\, z_\{t-1\} \;+\; B\, s_t \;+\; \eta_t, \eta_t \sim N(0, Q)$$

where A is the state-transition matrix, B maps conditioning variables into the latent space, and η_t is a zero-mean Gaussian process noise with covariance Q. This model is a controlled approximation rather than a claim that speech follows a globally linear dynamical system. Its role is to encode the expectation that acoustic states change smoothly unless supported by strong evidence. The matrix Q controls the allowable rate of change: larger process noise permits rapid transitions but weakens smoothing, whereas smaller process noise enforces continuity but can blur onsets and transients.

The linear Gaussian form connects the proposed model to classical state-space filtering and smoothing [11], and to broader linear Gaussian modeling frameworks that clarify the

relationship between latent-variable models and temporal priors [12]. In the proposed setting, the transition model is used as a structured prior inside variational inference rather than as a closed-form Kalman filter. This choice preserves the ability to use nonlinear emission and adaptation networks while retaining an interpretable mechanism for temporal regularization. Robustness arises because the prior reduces the variance of the posterior when observations are unreliable, but the likelihood can still override the prior when the observed evidence is strong.

4. Reference-Driven Feature Adaptation

4.1 Reference Representation

A reference encoder e_ψ maps the reference variable r_t to a compact embedding $u_t = e_\psi(r_t)$. The reference may be an enrolled clean utterance from the target speaker, a short adaptation utterance recorded in the target channel, or a domain prototype constructed from previously observed conditions. The embedding is not assumed to contain phonetic content aligned with the target; it is intended to capture condition-level attributes such as speaker timbre, channel coloration, and noise statistics. This distinction is important because direct frame-wise copying from a reference utterance would introduce phonetic mismatch and could corrupt reconstruction.

Reference-conditioned feature transformation has precedents in adaptive normalization and feature-wise conditioning. Adaptive instance normalization shows that global style statistics can modulate deep feature representations without changing their spatial content [13]. Feature-wise linear modulation generalizes this idea by learning affine transformations conditioned on auxiliary inputs [14]. Layer normalization provides a stable normalization mechanism that is compatible with such conditioning [15]. The proposed adaptation operator borrows the principle of conditioned modulation but applies it in a probabilistic latent space, where the strength of the transformation can be weighed against posterior uncertainty.

4.2 Low-Rank Adaptation Operator

To limit the risk of negative transfer and to keep the adaptation interpretable, the reference embedding modulates a low-rank transformation of latent features. The adaptation is

$$\Delta_t = U\, diag(\lambda(r_t))\, V^T\, h(z_t)$$

where $h(z_t)$ is a nonlinear feature map of the latent state, U and V are learned basis matrices, $\lambda(r_t)$ is a vector of reference-dependent gains, and $diag(\lambda(r_t))$ forms a diagonal scaling matrix. The matrices U and V define a shared low-rank subspace for adaptation, while $\lambda(r_t)$ determines how strongly each direction in that subspace is activated by the reference. The low-rank constraint restricts the adaptation to a small number of condition-sensitive directions, which reduces the capacity of the reference to overwrite phonetic content encoded in z_t.

The adapted latent feature is $z'_t = z_t + \Delta_t$. This additive form is analytically convenient because it separates the evidence-driven state from the reference-driven correction. It also makes the failure mode transparent: if the reference is mismatched, Δ_t becomes an additive bias whose effect can be controlled by the fusion stage. The operator is not a substitute for the observation likelihood. It is a structured prior-like correction that shifts the latent representation toward the reference condition while leaving the evidence term available to correct the shift when the observation contradicts it.

4.3 Uncertainty-Aware Fusion

The evidence-driven posterior and the reference-conditioned estimate are combined by precision-weighted fusion. Let μ_{0,t} and Σ_{0,t} denote the mean and covariance of the evidence-driven posterior, and let μ_{r,t} and Σ_{r,t} denote the mean and covariance of the reference-conditioned estimate. The fused mean is

$$\mu *_t = \Sigma *_t (\Sigma^{-1}_{0,t} \mu_{0,t} + \Sigma^{-1}_{r,t} \mu_{r,t})$$

and the fused covariance is

$$\Sigma *_t = (\Sigma^{-1}_{0,t} + \Sigma^{-1}_{r,t})^{-1}$$

where Σ^{-1} denotes the inverse covariance, or precision matrix. The fused estimate is a precision-weighted compromise between the two sources. When the evidence covariance Σ_{0,t} is large, the evidence is uncertain and the reference estimate receives greater weight. When the reference covariance Σ_{r,t} is large, the reference is unreliable and the evidence-driven posterior dominates. The fused covariance is smaller than either individual covariance in the directions where both sources are informative, which reflects the statistical benefit of combining independent evidence.

This fusion rule is preferable to a fixed interpolation coefficient because it adapts frame by frame and direction by direction. A reference may be informative for low-frequency spectral coloration but uninformative for high-frequency phonetic detail. The covariance matrices can express this anisotropy, and the precision-weighted rule automatically allocates influence accordingly. Calibration is nevertheless important: if either covariance is systematically overconfident, the fusion will overweight that source. Modern neural networks are often poorly calibrated in this sense [16], so the covariance estimates should be regularized during training and interpreted as relative rather than absolute measures of reliability.

4.4 Relation to Reference-Conditioned Synthesis

Reference-conditioned style transfer in out-of-domain singing voice synthesis addresses a related but distinct problem: it uses a reference to control acoustic style while preserving content [17]. The present framework shares the principle that a reference can provide condition-level information not present in the degraded observation, but it differs in two respects. First, the adaptation is applied to a probabilistic latent state whose uncertainty is explicitly modeled. Second, the reference is fused with the observation through precision weighting rather than used as a direct style target. These differences matter because speech reconstruction must preserve phonetic identity under noise, and an unconstrained reference transformation can introduce content artifacts even when it improves perceptual style matching.

5. Reconstruction and Inference Procedure

5.1 Training Objective

The model is trained by maximizing the variational lower bound while regularizing the adaptation operator. A practical objective to minimize is

$$J = E[||x_t - g_\theta(z_t, s_t)||^2_{\Sigma_t^{-1}}] + \beta KL(q_\varphi \,||\, p) + \gamma ||\Delta_t||^2_*$$

where ‖·‖^2_{Σ_t^{-1}} denotes the squared Mahalanobis distance under the observation precision, β and γ are nonnegative hyperparameters, and ‖Δ_t‖^2_ is a norm on the adaptation term. The first term encourages the latent state to explain the observed features under the

assumed noise covariance. The second term regularizes the variational posterior toward the temporal prior. The third term penalizes large reference-driven corrections, which limits negative transfer when the reference is mismatched.

The hyperparameters β and γ control the trade-off between data fidelity, temporal regularization, and reference conservatism. They are not assigned numerical values here because the appropriate trade-off depends on the feature representation, the expected noise covariance, and the reliability of the reference source. The norm on Δ_t can be chosen as a weighted Frobenius norm or as a precision-weighted norm that reflects the uncertainty of the reference subspace. The objective is stated as a design specification rather than as a fitted model; its components follow directly from the probabilistic model and the adaptation operator.

5.2 Inference-Time Adaptation

At inference time, the variational encoder produces $\mu_{0,t}$ and $\Sigma_{0,t}$ from the observed sequence. The reference encoder produces u_t, which parameterizes $\lambda(r_t)$ and the reference-conditioned estimate $\mu_{r,t}$, $\Sigma_{r,t}$. The fused estimate is computed by the precision-weighted rule, and the reconstructed acoustic features are obtained by applying a deterministic decoder to μ_t. If no reference is available, the reference precision is set to zero and the reconstruction reduces to the evidence-driven posterior. If the reference is available but its reliability is unknown, a conservative reference covariance should be used so that the observation dominates in informative frames.

The inference procedure is modular. The probabilistic acoustic model, the reference encoder, and the fusion rule can be trained jointly or in stages. Joint training is conceptually preferable because it allows the fusion rule to learn how reference errors and observation errors interact. Staged training may be more stable in practice, but it requires care to avoid a mismatch between the covariance estimates used during training and those used at inference. In either case, the reconstruction output should be accompanied by the fused covariance Σ_t, which provides a frame-level uncertainty summary for downstream decision making.

6. Analytical Assessment and Technical Discussion

6.1 Robustness to Nonstationary Noise

The proposed model improves robustness to nonstationary noise by making observation uncertainty explicit. When a noise burst produces a large residual, the likelihood term becomes flat, and the posterior relies more heavily on the temporal prior and on the reference-conditioned estimate. When the noise subsides, the likelihood term sharpens, and the posterior returns toward the observed evidence. This behavior is qualitatively different from a deterministic regressor, which must produce a single output regardless of how ambiguous the input is. The covariance-weighted fusion rule also prevents a noisy frame from overwhelming a stable reference estimate, provided that the reference covariance is not underestimated.

The main failure mode is covariance misspecification. If Σ_t is too small, the model treats corrupted observations as reliable and suppresses the reference and temporal prior. If Σ_t is too large, the model over-smooths and may fail to reconstruct transient phonetic events. The model therefore requires a noise covariance estimator or a learned covariance function that is sensitive to local signal-to-noise conditions. This requirement is a design burden, but it is also a source of interpretability: the covariance field reveals which time-frequency regions the model considers reliable.

6.2 Reference Mismatch and Negative Transfer

Reference-driven adaptation is beneficial only when the reference shares condition-level attributes with the target. If the reference comes from a different speaker, channel, or speaking style, the adaptation operator may shift the latent state in an unhelpful direction. The low-rank form of Δ_t mitigates this risk by restricting the adaptation to a small subspace, and the precision-weighted fusion rule provides a second safeguard by down-weighting the reference when its covariance is large. Nevertheless, the model cannot detect all forms of mismatch from covariance alone. A reference that is confidently wrong can still bias the reconstruction if its covariance is underestimated.

This issue connects the proposed framework to reference-conditioned voice conversion and one-shot adaptation. Instance-normalization-based voice conversion separates content and speaker representations to support one-shot conversion [18], and autoencoder-based voice style transfer demonstrates zero-shot adaptation with a reference [19]. Benchmarking protocols for voice conversion provide useful evaluation logic for reference fidelity and speaker similarity [20]. The present framework differs in that its primary target is reconstruction under degradation, not conversion between speakers. Consequently, reference fidelity must be balanced against phonetic preservation, and the fusion rule must be conservative when the reference is not known to match the target condition.

6.3 Temporal Modeling and Transient Preservation

The temporal prior improves robustness by smoothing across frames, but excessive smoothing can blur plosives, onsets, and rapid formant transitions. The process noise covariance Q controls this trade-off. A small Q enforces continuity and suppresses isolated noise, while a large Q permits rapid changes at the cost of weaker smoothing. The variational posterior can adapt Q implicitly if it is parameterized as a function of the conditioning variables, but this flexibility must be regularized to avoid overfitting to the training conditions. A practical design should therefore expose Q as a structured parameter and interpret it as a prior on the expected rate of acoustic change.

The interaction between temporal smoothing and reference adaptation is also important. A reference estimate that is stable across time can dominate the posterior when the observation is weak, but it may also introduce a persistent bias that the temporal prior reinforces. The precision-weighted fusion rule counteracts this risk only if the reference covariance reflects the temporal stability of the reference. Reference statistics that are reliable for stationary channel coloration may be unreliable for dynamic phonetic content, and the covariance model should distinguish these aspects rather than treating the entire reference embedding as equally trustworthy.

6.4 Comparison with Alternative Approaches

Deterministic enhancement systems are computationally efficient and can achieve strong performance when training and test conditions are well matched [1], [2]. Their weakness is not a lack of capacity but a lack of explicit uncertainty: they cannot represent the possibility that a reconstructed frame is ambiguous. Generative models such as autoregressive and diffusion models can represent rich distributions [5], [6], but they may require many sampling steps and can hallucinate detail under weak conditioning. The proposed model occupies a middle ground. It uses a latent Gaussian posterior for tractable uncertainty estimation and a low-rank reference adaptation for condition-level control, while leaving the emission and adaptation mappings nonlinear.

Self-supervised features provide robust acoustic representations [7], [8], and large-scale weakly supervised systems show the value of broad acoustic coverage [9]. These methods are complementary to the proposed framework. A self-supervised encoder could supply the reference embedding $e_\psi(r_t)$ or initialize the variational encoder q_φ. The probabilistic model would then add an explicit uncertainty layer and a fusion mechanism that are not inherent in a deterministic feature extractor. The combination is promising because it separates representation learning from the inferential logic that governs reconstruction.

7. Conclusion

This manuscript has developed a probabilistic acoustic modeling framework with reference-driven feature adaptation for robust speech reconstruction. The observation process is represented as a conditional Gaussian transformation of a latent acoustic state, and amortized variational inference estimates a posterior that separates observation uncertainty from state uncertainty. A temporal prior supplies continuity across frames, while a low-rank adaptation operator modulates latent features using a reference embedding. Precision-weighted fusion combines evidence-driven and reference-conditioned estimates, allowing the reconstruction to shift toward the reference when acoustic evidence is weak and toward the observation when the reference is unreliable.

The framework's main contribution is the integration of three ideas that are often treated separately: probabilistic acoustic modeling, temporal state estimation, and reference-conditioned adaptation. The equations make the trade-offs explicit. The observation covariance governs trust in the degraded signal, the process noise covariance governs temporal smoothing, the low-rank adaptation matrices govern the capacity of the reference to alter the latent state, and the fusion rule governs how the two estimates are combined. These quantities have clear interpretations and can be regularized, inspected, and calibrated.

The analysis also identifies boundary conditions. Covariance misspecification can cause overconfidence in corrupted observations or excessive smoothing. Reference mismatch can produce negative transfer even when the adaptation is low rank. Temporal priors that are too strong can blur transients. These limitations do not undermine the framework; they define the design requirements for a reliable implementation. A well-designed system should estimate observation uncertainty locally, regularize reference adaptation conservatively, and expose temporal process noise as a structured parameter rather than a hidden constant. Under these conditions, probabilistic acoustic modeling with reference-driven feature adaptation offers a coherent and technically defensible route to robust speech reconstruction.

References

1. Loizou, P. C. (2013). Speech enhancement: Theory and practice (2nd ed.). CRC Press.
2. Wang, D., & Chen, J. (2018). Supervised speech separation based on deep learning: An overview. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(10), 1702–1726.
3. Ephraim, Y., & Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Transactions on Acoustics, Speech, and Signal Processing, 32(6), 1109–1121.
4. Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. In Proceedings of the 2nd International Conference on Learning Representations.

5. 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.

6. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems (Vol. 33, pp. 6840–6851).

7. Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In Advances in Neural Information Processing Systems (Vol. 33, pp. 12449–12460).

8. Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 3451–3460.

9. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In Proceedings of the 40th International Conference on Machine Learning (pp. 28492–28518).

10. Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

11. Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. Journal of Basic Engineering, 82(1), 35–45.

12. Roweis, S., & Ghahramani, Z. (1999). A unifying review of linear Gaussian models. Neural Computation, 11(2), 305–345.

13. Huang, X., & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1501–1510).

14. Perez, E., Strub, F., de Vries, H., Dumoulin, V., & Courville, A. (2018). FiLM: Visual reasoning with a general conditioning layer. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).

15. Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. arXiv preprint arXiv:1607.06450.

16. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (pp. 1321–1330).

17. Zhang, Y., Huang, R., Li, R., He, J., Xia, Y., Chen, F., ... & Zhao, Z. (2024, March). Stylesinger: Style transfer for out-of-domain singing voice synthesis. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 17, pp. 19597-19605).

18. Chou, J.-C., & Lee, H.-Y. (2019). One-shot voice conversion by separating speaker and content representations with instance normalization. In Proceedings of Interspeech 2019 (pp. 664–668).

19. Qian, K., Zhang, Y., Chang, S., Yang, M., & Hasegawa-Johnson, M. (2020). AutoVC: Zero-shot voice style transfer with only autoencoder loss. In Proceedings of the 37th International Conference on Machine Learning (pp. 7630–7639).

20. Zhao, Y., Huang, W.-C., Tian, X., Yamagishi, J., Das, R. K., Kinnunen, T., Ling, Z., & Toda, T. (2020). Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion. arXiv preprint arXiv:2008.12527.