A DESIGN FRAMEWORK FOR EXPRESSIVE VOICE SYNTHESIS

Camille Noufi and Lloyd May
Center for Computer Research in Music and Acoustics, Stanford University
{cnoufi,lloyd}@ccrma.stanford.edu

ABSTRACT

This extended abstract proposes a design framework for interactive, real-time control of expression within a synthesized voice. Within the framework, we propose two concepts that would enable a user to flexibly control their personalized sound. The *voice persona* that determines the "tone of voice" is defined as a point existing within a continuous probability space. This point defines the parameters that determine the distribution space of the latent features required for synthesis, allowing for flexible modification and fine-tuning. Secondly, expression within a persona can be achieved through modification of meaningful high-level abstractions, which we call *macros*, that subsequently modify the distribution space of corresponding latent features of the synthetic speech signal.

1. INTRODUCTION

The human voice is an important component of identity, affecting how we communicate and express complex ideas and nuanced emotions. A controllable, expressive synthetic voice provides a form of communication to those who seek to augment what their voice or their abilities might allow them to do otherwise.

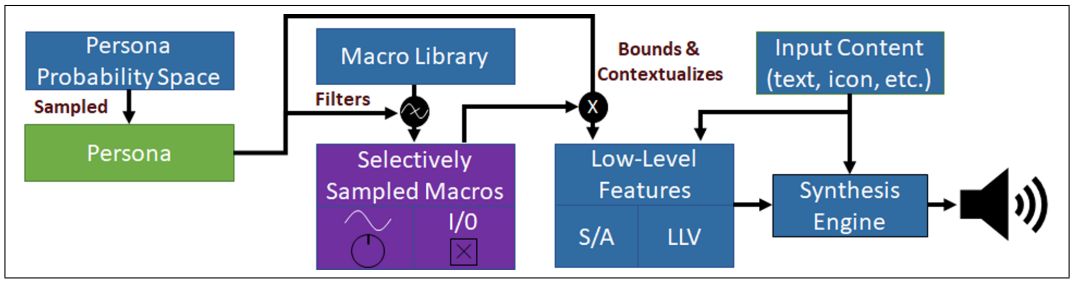
Modern text-to-speech (TTS) and voice cloning systems can synthesize human-like speech very similar to that of a human voice. Depending on the implementation, these systems can generate speech that emulates the voice of a user or the voice of a desired personality or character. Similarly, singing synthesizers are often modeled after a performer or virtual celebrity. Modern state-of-the-art TTS and voice cloning systems leverage generative models as their main synthesis engine. At a high level, both classical parametric synthesis models and modern generative neural networks synthesize the voice via a probability maximization over a set of low-level parameters [1, 2]. Classical parametric synthesis has leveraged well-studied parameters, such as spectral and acoustical properties of the human voice, alongside linguistic, prosodic and phonetic contexts [1–3], while many generative neural networks seek to learn these parameters as latent features of an observed speech audio dataset [4–7]. Although their parameter sets may differ, these two models are both capable of generating intelligible speech while having a perceptually-continuous "tone of voice" given the "correct" choice of low-level features.

However, an important capability currently underdeveloped in neural network-based voice synthesis is that of temporally-dynamic vocal expressivity that is 1) perceptually meaningful and 2) user-determined. Vocal attributes leveraged for expressivity such as pitch variability, voice quality, pronunciation/stress, and speech rate/cadence are determined by the parameter choices when a voice is generated–either by example [4, 6] or by direct control [4, 5, 7]). These attributes are often inaccessible to the user of the voice once the voice is built. In this extended abstract, we propose a framework for expressive vocal synthesis that leverages 1) a voice persona, drawn from an underlying continuous probability space, that bounds and contextualizes low-level/latent synthesis features and 2) user-defined perceptual abstraction that allow for real-time performativity and expressivity within the chosen voice persona.

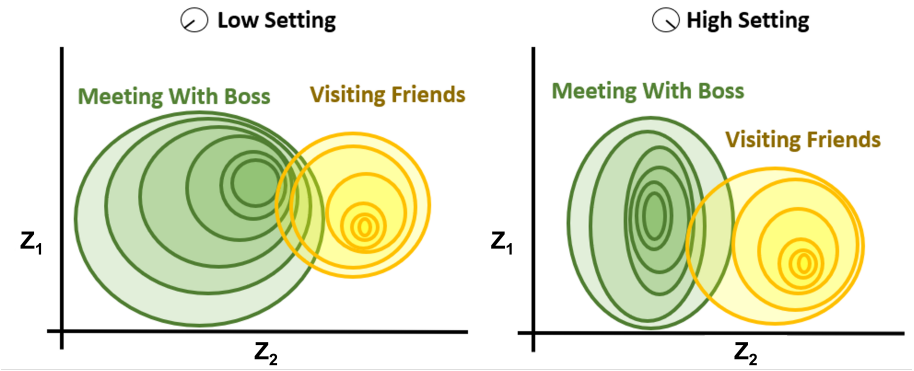
2. FRAMEWORK

Drawing from theories of gender performativity [8], vocal code-switching [9], and digital instrument design [10], we propose that a vocal persona is sampled from a fluid *persona probability space* that contextualizes the voice one may use in a certain setting or to embody a certain personality. For example, one vocal persona adopted for "meeting with boss" and another adopted for "visiting friends" perhaps share a similar distribution space, as visualized by Figure 1b. In contrast, a user's chosen voice for "visiting friends" may have much less overlap with a "Dolly Parton" vocal persona.

We present our proposed framework through a visualization of the proposed *persona probability space* $\mathbf{P}$ and its relationship to latent variables $\mathbf{Z}$ used to synthesis a voice (Figure 1a). We present this framework generally, allowing for the parameter space to consists of classical synthesis parameters such as a necessary and sufficient spectral/acoustic set or a set of learned latent variables (denoted in Figure 1a as S/A and LLV, respectively). We characterize this *persona probability space* $\mathbf{P}$ as being a distribution of parameters describing another N-dimensional probability mixture model describing low-level synthesis features $\mathbf{Z} = \{Z_1, Z_2, Z_n, ..., Z_N\}$. Sampling persona P_a defines the set of N probability density functions (PDF) $f_a(z_n|\theta_{n_a})$ for each synthesis parameter Z_n, where θ_n are



(a) A sampled vocal persona parameterizes the sample space of either spectral/acoustic (S/A) or learned/latent variables (LLV) required by a speech synthesis engine. A set of K user-selected expression *macros* allows for modification of this parameterization.



(b) An example of the effect of adjusting an expression *macro* x to "low" and "high" on the underlying distributions of synthesis variables Z_1 and Z_2 for two different vocal personas.

Figure 1: Proposed human-in-the-loop expressive vocal synthesis framework.

the parameters describing the PDF. Sampling a different persona P_b defines another set of N PDFs $f_b(x_n|\theta_{n_b})$ for each latent feature Z_n. These distribution spaces could be as overlapping or as separated as is perceptually meaningful for the user.

The bottom row of the flowchart in Figure 1a shows how perceptually-meaningful expressivity attributes affect the low-level features utilized by a speech synthesis engine. We propose the concept of a *macro* as a perceptually-informed abstraction that modifies the low-level features such that the modification yields a vocal tone aligned with the intended expressivity. For example, a user may want to modify how "excited" the voice sounds within the bounds of current persona P_a. An "excitement" control gives the user the ability to modify the "amount" of excitement in their current voice on a scale from 0 to 100. Given control variable $x \in X \sim \mathit{Uniform}[0, 100]$, a function $m_n(x) = w_n y_n(x)$ maps x to a corresponding modification value applied to PDF parameters θ_n. Here, w_n is a scalar weight corresponding to the involvement of synthesis feature Z_n in the high-level "excitement" macro M. $y_n(\cdot)$ is a transformation or warping function that allows for the weighting of each macro to be configurable and could be learned or selected by the user. Macro M is the set of functions $m_n(\cdot)\forall n \in [1...N]$. Within a persona, a set of K macros $\{M_1, ..., M_k, ..., M_K\}$ can be created by or presented to the user that allow for modification that is useful or meaningful. Within our current proposed design, these macros multiplicatively combine to determine the modification to θ_n. More explicitly, a user-determined set of K macros can modify the parameters θ_{n_a} describing a PDF $f_a(z_n|\theta_{n_a})$ within the current persona mixture model P_a such that:

$$\theta_{n_a} = (\prod_{k=1}^{K} m_{n_k}(x_k))\theta_{n_a}, \forall n \in [1...N]. \quad (1)$$

The proposed framework aims to maximize the agency a user has over the speech synthesis while actively accommodating users with differing levels of desired control. Therefore the framework allows multiple levels of interaction through the use of "default" personas and macros which work directly upon implementation but are highly configurable. Users are able to customize macro controls and persona attributes to create a user-experience that is catered to, and by, each user. This enables power users wanting detailed control over expressive nuance to create and modify hyper-specific macros by also accessing the low-level parameters weights $\mathbf{w}$ and warping functions $\mathbf{y}(\cdot)$. Paramount to this framework is that the user dictates the level of interaction and complexity that they desire.

2.1 Future Work

Future work includes user experience studies examining the efficacy of different gesture and interaction paradigms for both macro and persona-level control as well as the design and implementation of a proof-of-concept prototype for expressive control within an existing synthesized voice. We also plan to augmenting existing speech synthesis models with the proposed "macro" abstraction.

3. REFERENCES

[1] T. Yoshimura, K. Tokuda, T. Kobayashi, T. Masuko, and T. Kitamura, "Simultaneous Modeling Of Spectrum, Pitch And Duration In HMM-Based Speech Synthesis," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[2] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, nov 2009.

[3] X. Serra, "A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition," Ph.D. dissertation, Stanford University, 1989.

[4] Y. Wang, D. Stanton, Y. Zhang, R. J. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *35th International Conference on Machine Learning, ICML 2018*, vol. 12, 2018, pp. 8229–8238.

[5] W. N. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen, P. Nguyen, and R. Pang, "Hierarchical generative modeling for controllable speech synthesis," in *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[6] R. Valle, K. Shih, R. Prenger, and B. Catanzaro, "Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis," 2020.

[7] R. Valle, J. Li, R. Prenger, and B. Catanzaro, "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Institute of Electrical and Electronics Engineers Inc., 2020, pp. 6189–6193.

[8] J. Butler, *Gender trouble*. Routledge Publishing, 2002.

[9] B. E. Bullock and A. J. Toribio, *The Cambridge Handbook of Linguistic Code-switching*. Cambridge University Press, 2009.

[10] P. Cook, "Remutualizing the musical instrument: Co-design of synthesis algorithms and controllers," *Journal of New Music Research*, vol. 33, pp. 315–320, 09 2004.