Transferência de Estilo para Síntese de Fala Expressiva

Leonardo B. de M. M. Marques , Lucas H. Ueda , Paula D. P. Costa
{l218479@dac.unicamp.br, l156368@dac.unicamp.br, paulad@unicamp.br}

Departamento de Engenharia de Computação e Automação (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Campinas, SP, Brasil

Resumo – Embora pesquisas recentes em sistemas de conversão de texto em fala tenham mostrado melhorias significativas na naturalidade e inteligibilidade, transmitir aspectos expressivos da fala por meio da síntese ainda é um problema em aberto. Esta é uma característica crucial para permitir que agentes artificiais socialmente interativos exibam comportamentos típicos da comunicação humana. A maneira mais comum de abordar a expressividade é considerar os estilos de fala, uma descrição de alto nível das maneiras de falar, como "narrativa", "amigável" ou "sussurrando". Neste contexto, este trabalho aborda a seguinte questão: como modelar estilos de fala de maneira realista? Para resolver esse problema, exploramos o uso de modelos generativos de difusão, e visamos usar recursos prosódicos de baixo nível da fala, frequência fundamental, duração e energia, a fim de obter representações de estilo que melhor condicionam o modelo texto-fala.

Palavras-chave – síntese de fala expressiva, estilos de fala, transferência neural

1. Introdução

Devido ao rápido desenvolvimento das técnicas neurais de modelagem acústica e geração de formas de onda, as tecnologias de texto-fala estão reduzindo progressivamente a lacuna entre a fala natural e a sintética [9]. No entanto, um problema ainda em aberto é a síntese de fala expressiva realista [8].

A síntese de fala expressiva pode ser caracterizada como um problema de mapeamento do tipo *one-to-many*, uma vez que um mesmo fonema pode apresentar diferentes produções acústicas e prosódicas, observáveis, por exemplo, em diferentes entonações, sotaques, ritmos e velocidade de produção [12].

Os estilos de fala são definidos como os atributos globais que descrevem a emoção, afeto e/ou atitude social transmitida através da fala por um falante em um domínio particular. Leitura, locutor, conversação ou emoção (feliz, zangado, triste, etc...) são alguns exemplos [9].

Nas conversas cotidianas face-a-face, as interações contém sinais sociais como humor, empatia e compaixão por meio tanto do conteúdo linguístico quanto do estilo da fala. Assim, para alcançar um meio de comunicação mais afetivo e humano, é de grande importância que os sistemas de conversão de texto em fala sintetizem falas com estilos de fala adequados que estejam de acordo com o contexto conversacional [7].

As principais abordagens que visam introduzir expressividade tentam modelar o estilo usando uma rede neural para gerar um vetor latente único e global através do aprendizado não supervisionado que represente o estilo [14]. A grande maioria desses trabalhos usa como entrada o mel-espectrograma (uma representação do sinal de áudio no domínio do tempo e da frequência) e, através de uma rede recorrente e convolucional (chamada de codificador de referência [10]), constrói-se um vetor de referência, que é utilizado por outros módulos na obtenção do vetor de estilo. À composição desses módulos junto com o codificador de referência, dá-se o nome de codificador de estilo. A arquitetura típica de um sistema desse tipo é apresentada na Figura 1, na qual os blocos verdes são as entradas e saída, os cinzas são redes neurais, e em amarelo destaca-se o vetor de estilo.

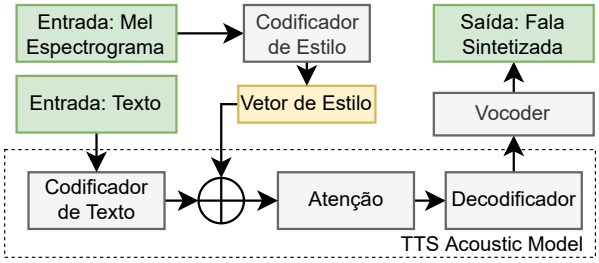

Figura 1. Arquitetura de um TTS neural incorporando estilo.

Existem várias tentativas de melhorar a capacidade de modelagem de estilo do vetor de referência, como a partir desse realizar o aprendizado de um banco de vetores chamado "Global Style Tokens" (GST) [11], esperando que cada um capture um aspecto global aleatório da distribuição de áudio, como por exemplo velocidade de fala, ruído de fundo, timbre, etc. Outra abordagem consiste no uso de modelos generativos, como o auto-encoder variacional (VAE) [15] e fluxos normalizadores

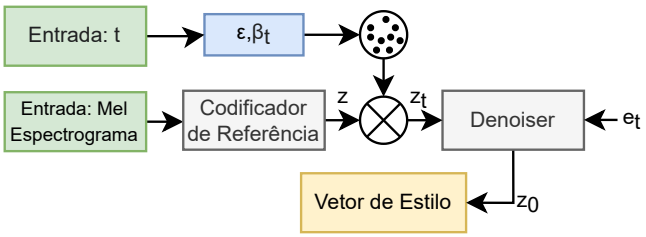

Figura 2. Codificador de estilo baseado em modelos de difusão.

(*flows*) [1]. Um problema recorrente desses modelos é a questão do vazamento de informações: não se sabe se o vetor de estilo está modelando apenas o estilo isoladamente: não há interpretabilidade nos módulos citados acima, e informações do falante, ruído de fundo, ambiência, ou quais outras características podem estar sendo modeladas. Alguns trabalhos já propuseram métodos para desembaraçar os atributos do falante e do estilo obtidas através do mel-espectrograma de referência [2].

Com base no exposto, recentemente, o desafio de incorporar estilo em sistemas TTS expressivos foi subdividido em duas questões principais [13]: como obter um vetor de estilo significativo, dado o rótulo de estilo e como injetar adequadamente o vetor de estilo em um modelo acústico texto-fala. Nesse contexto, o presente trabalho se concentra no primeiro problema: como modelar o estilo de fala de maneira realista.

2. Proposta

Dado o recente sucesso de modelos de difusão [3] na tarefa de síntese de imagem a partir do texto, obtendo melhor desempenho que as GANs, inicialmente exploramos a modelagem de estilo através desses. Modelos de difusão [5] são um tipo de modelo generativo consistindo em uma cadeia de Markov que gradativamente remove a informação presente nos dados através da adição de ruído sequencial. Dessa maneira, leva-se a distribuição original dos dados à uma distribuição gaussiana. Após esse procedimento, o processo reverso gradativo de reconstrução é aprendido através de redes neurais, criando-se assim a capacidade de sintetizar um dado partindo de uma amostra de ruído gaussiano.

O processo de geração do vetor de estilo, mostrado na Figura 2, consiste então na entrada com o mel-espectrograma de referência, o qual se deseja capturar o estilo, que é transformado num vetor de referência, denominado z, através do codificador de referência. Entra-se também com o número t de passos de ruído gaussiano que serão adicionados ao vetor de referência z, levando esse o vetor de referência para o nível de ruído z_t, a partir da seguinte equação:

$$\boldsymbol{z}_t(\boldsymbol{z}, \boldsymbol{\epsilon}) = \sqrt{\bar{\alpha}_t}\boldsymbol{z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \tag{1}$$

na qual $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ e $\bar{\alpha}_t$ é um hiper parâmetro do processo de difusão.

Esse vetor ruidoso de referência é então utilizado como conhecimento a priori para iniciar o processo reverso de difusão. Em cada passo, uma única rede neural, o "denoiser", recebe o vetor de referencia ruidoso no nível de ruído t, z_t, juntamente com o vetor que sinaliza o passo que está sendo executado, e_t e retorna o vetor reconstruído z_{t-1}. A modelagem do processo reversa também é feita através de distribuições gaussianas, tendo a média e variância modelada por redes neurais a partir de z_t, de acordo com a equação a seguir:

$$p_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t) := \mathcal{N}(\boldsymbol{z}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{z}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{z}_t, t)), \tag{2}$$

na qual $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ representam a média e a matriz de covariância respectivamente.

Esse processo é então repetido t vezes para reconstruir o vetor o qual foi aplicado t passos de ruído, sendo o resultado final, z_0, o denominado vetor de estilo, que será usado para condicionar o modelo acústico de texto-fala a gerar fala no estilo da referência dada. O vetor de estilo é então concatenado à saída do codificador de texto e ambos servem de entrada para os módulos de atenção e o decodificador. Construiu-se a hipótese de que, a partir processo de remoção de ruído iterativo, o modelo de difusão era capaz de selecionar os atributos mais importantes do vetor ruidoso e reconstruir o vetor de estilo da maneira que melhor condicionasse o modelo acústico.

Seguindo a tendência de modelos TTS de começarem a usar atributos prosódicos de baixo nível para realizar a síntese de fala [6], e também com o intuito de controlar as informações que o codificar de estilo recebe/processa, a abordagem de utilizar esses atributos ao invés do mel-espectrograma será investigada. Especificamente, ao entrarmos com um mel-espectrograma para referência de estilo, não há garantia de que a rede explicitamente esteja modelando o estilo, podendo capturar outros atributos indesejados presentes, como o timbre do falante, ruído de fundo, ambiência, entre outros.

Hipotetizamos que um número finito de características prosódicas de um sinal de áudio seja suficiente para conseguir capturar seu estilo. Dessa maneira, ao introduzir essas características no codificador de estilo, pretende-se mitigar problemas de vazamento de informações não desejadas (como timbre do falante, ruído de fundo), que ocorrem atualmente ao se entrar com o mel-espectrograma, avançando num caminho da interpretabilidade para obter representações de estilo mais significativas.

Assim, considera-se utilizar um subconjunto de parâmetros do GeMAPS [4], um grupo de parâmetros acústicos selecionados com base no potencial de indicar características afetivas fisiológicas, utilização em trabalhos passados e significância teórica; a fim de ser um padrão para pesquisas futuras. Visa-se realizar um estudo baseado na importância de atributos para avaliar quais são suficientes para capturar bem o estilo em diferentes bancos de dados contendo diferentes estilos.

Adicionalmente, um objetivo também é o de mudar do aprendizado não-supervisionado para o supervisionado, a fim de obter uma melhora no desempenho. Com a introdução dos rótulos de estilo, é possível adicionar módulos classificadores de emoção após o codificador de estilo para fazer com que os gradientes do classificador torne os vetores de estilo mais discriminativos, tendo esses a informação que distingue os estilos. Também, a fim de fazer com que o codificador de estilo não aprenda informações de falante, é possível introduzir um classificador de locutor com uma camada de reversão do gradiente, com o objetivo de se afastar dos mínimos.

3. Resultados

Um banco de dados de uma única falante em Português do Brasil foi usado. Ele contém 15 horas de fala, sendo 6 de conteúdo expressivo, falado por uma atriz de voz profissional. Os estilos presentes no banco de dados são categorizados como “animado”, “acolhedor” e “ríspido”, e foram projetados para aplicações baseadas em serviços com foco em consumidores. Existe um total de 12400 enunciados neutros, 1307 animados, 1308 acolhedores, e 1256 ríspidos. Pra cada categoria, 90% das sentenças foram usadas para treinamento e 10% para validação e teste.

Para avaliar o desempenho do modelo, foi realizado um experimento perceptual no qual 30 participantes foram solicitados para ouvir e atribuir valores de naturalidade e expressividade de cada síntese. Especificamente, comparou-se o modelo proposto baseado em difusão, com aqueles que eram estado-da-arte na literatura: o VAE e o VAE+Flow. Para avaliar a naturalidade, uma frase do conjunto de teste foi sintetizada por cada modelo, que recebiam o mel-espectrograma correspondente como entrada de estilo, e então solicitou-se o julgamento de 0 a 100 o quão natural cada áudio soava. As médias do resultado são mostrado na Figura 3. Nele, observa-se que, enquanto que nos estilos acolhedor e neutro os desempenho são bastante similares, o modelo de difusão obtém melhor desempenho nos estilos ríspido e animado.

Para avaliar a expressividade, um experimento de preferência ABX foi conduzido, no qual cada um dos modelos recebia uma frase e uma referência de estilo também do conjunto de teste não correspondentes com intuito de realizar a transferência de estilos. Os participantes então escolhiam qual síntese tinha o estilo mais parecido com o da referência, para avaliar o quão bem o estilo foi transferido para o conteúdo textual. Nessa, comparou-se o de difusão com o VAE+Flow, com a opção de “ambos são igualmente parecidas” também inclusa. Os resultados são mostrado na Figura 4. No caso de transferência paralela, na qual o mel-espectrograma de estilo de referência tem o mesmo conteúdo que o texto de entrada, o modelo de difusão obteve melhor desempenho, enquanto no caso não paralelo (texto da referência de estilo diferente do texto de entrada) foi um pouco pior. Considerando o caso dos dois agregados, o modelo de difusão obteve 14.41% de preferência a mais que o VAE+Flow.

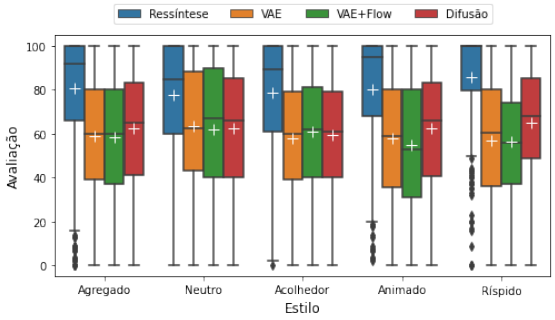

Figura 3. Experimento de naturalidade

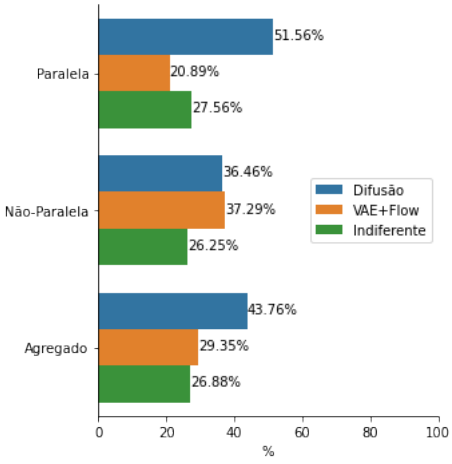

Figura 4. Experimento ABX de preferência de estilo

4. Conclusões

Foram investigados as principais técnicas para a modelagem de estilo baseado em modelos generativos não-supervisionados. Experimentando os modelos de difusão para gerar os vetores de estilos, observou-se uma melhoria na expressividade e naturalidades nos estilos mais energéticos: ríspido e animado. Nos estilos neutro e aco-

lhedor, os modelos obtiveram desempenhos similares. Para trabalhos futuros, busca-se fazer uso de técnicas supervisionadas e utilizar atributos prosódicos a fim de obter representações de estilo mais significativas para melhor condicionar o modelo acústico de texto-fala.

Agradecimentos

Os autores agradecem ao Centro de Pesquisa e Desenvolvimento (CPQD), em especial ao Flávio O. Simões, Mário Uliani Neto, Edson J. Nagle, Fernando O. Runstein, e Bianca Dal Bó, pelo apoio, disponibilização dos recursos e banco de dados; e ao Ministério da Ciência, Tecnologia e Inovações pelo apoio e financiamento deste projeto. Este trabalho é apoiado pelo BI0S - Instituto Brasileiro de Ciência de Dados, bolsa #2020/09838-0, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

Referências

[1] Vatsal Aggarwal, Marius Cotescu, Nishant Prateek, Jaime Lorenzo-Trueba, and Roberto Barra-Chicote. Using vaes and normalizing flows for one-shot text-to-speech synthesis of expressive speech. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6179–6183. IEEE, 2020.

[2] Xiaochun An, Frank K. Soong, and Lei Xie. Disentangling style and speaker attributes for tts style transfer. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 30:646–658, jan 2022.

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[4] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 2015.

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[6] Adrian Łańcucki. Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6588–6592. IEEE, 2021.

[7] Jingbei Li, Yi Meng, Chenyi Li, Zhiyong Wu, Helen Meng, Chao Weng, and Dan Su. Enhancing speaking styles in conversational text-to-speech synthesis with graph-based multi-modal context modeling. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7917–7921. IEEE, 2022.

[8] Yi Ren, Ming Lei, Zhiying Huang, Shiliang Zhang, Qian Chen, Zhijie Yan, and Zhou Zhao. Prosospeech: Enhancing prosody with quantized vector pre-training in text-to-speech. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7577–7581. IEEE, 2022.

[9] Manuel Sam Ribeiro, Julian Roth, Giulia Comini, Goeric Huybrechts, Adam Gabryś, and Jaime Lorenzo-Trueba. Cross-speaker style transfer for text-to-speech using data augmentation. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6797–6801, 2022.

[10] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J. Weiss, Rob Clark, and Rif A. Saurous. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron, 2018.

[11] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A. Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis, 2018.

[12] Ning-Qian Wu, Zhao-Ci Liu, and Zhen-Hua Ling. Discourse-level prosody modeling with a variational autoencoder for non-autoregressive expressive speech synthesis. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7592–7596, 2022.

[13] Fengyu Yang, Jian Luan, and Yujun Wang. Improving emotional speech synthesis by using sus-constrained vae and text encoder aggregation. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8302–8306, 2022.

[14] Yuanhao Yi, Lei He, Shifeng Pan, Xi Wang, and Yujia Xiao. Prosodyspeech: Towards advanced prosody model for neural text-to-speech. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7582–7586, 2022.

[15] Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua Ling. Learning latent representations for style control and transfer in end-to-end speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6945–6949. IEEE, 2019.