Separando atributos de fala: conversão texto-fala expressiva entre locutores cruzados baseada na aprendizagem de representações

Lucas H. Ueda, Leonardo B.M.M. Marques , Paula Dornhofer Paro Costa
{l1156368@dac.unicamp.br, l218479@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 – A conversão texto-fala expressiva entre locutores cruzados consiste na transferência de um estilo de fala de um locutor para outro que nunca gravou falas com tal estilo. A efetividade nessa tarefa permite que seja possível transferir uma fala expressiva para locutores que só temos falas neutras em posse, amenizando assim a necessidade de gravação de dados expressivos de um novo locutor. O aprendizado de representações consiste em construir espaços onde os atributos de interesse são modelados, em particular, se os atributos são modelados independentemente torna-se possível condicioná-los de forma independente. O presente trabalho busca, através do aprendizado de representações, gerar espaços onde os atributos expressivos da fala (prosódia) e o timbre da voz sejam independentes, permitindo que um locutor neutro fale de forma expressiva sem nunca tê-las gravado.

Palavras-chave – síntese de fala, fala expressiva, modelagem de sequência, aprendizado de representações, prosódia, transferência de estilo.

1. Introdução

A fala sintetizada está presente cada vez mais no nosso cotidiano, das vozes gravadas na secretária eletrônica alguns anos atrás até os recentes assistentes virtuais controlados por voz. Os recentes avanços na área de aprendizado de máquina, em particular as redes neurais artificiais, possibilitaram que esses sistemas gerem voz artificial com qualidade próxima à fala humana [7, 8, 13].

No entanto, a comunicação oral humana não se baseia somente no conteúdo da mensagem a ser transmitida, mas também na forma como essa mensagem é realizada. Uma mesma frase pode ser lida de diferentes formas, alterando-se a entonação, o ritmo ou mesmo a emoção da fala. Essas diferentes formas de se enunciar uma mesma frase é o que chamamos de prosódia [1]. A prosódia está relacionada ao "como se fala" e não ao "o que se fala", e ela é responsável por não somente tornar uma frase mais interessante, como também auxilia na compreensão de seu conteúdo [2].

Diversos trabalhos propuseram formas de se incorporar a expressividade em sistemas de conversão texto-fala. Em [9], é proposto o *Reference Encoder*, um codificador de representação de estilos onde a prosódia de uma fala de referência é transferida para a fala sintética. Já em [10], uma base com diferentes estilos é utilizada, e o espaço gerado pelo *Reference Encoder* é analisado e utilizado para condicionar o estilo desejado na fala sintética. Mais recentemente, abordagens que modelam explicitamente componentes prosódicos vem sendo propostos, como o *FastPitch* [13], que propõe a modelagem da duração dos fonemas e da curva de frequência fundamental como parte da incorporação de expressividade no modelo.

Neste trabalho apresentamos resultados iniciais com foco na modelagem de representações de estilos. Utilizamos a arquitetura *FastPitch* como base do nosso modelo proposto, além disso, incorporamos o codificador de estilo *Reference Encoder* para gerar o espaço de estilos e assim condicionar o estilo alvo à fala sintetizada final. Adicionamos também componentes que buscam isolar os atributos modelados por cada componente do modelo final.

2. Base de dados

A base de dados utilizada consiste em cerca de 15 horas de áudios gravados por uma locutora brasileira, disponibilizada pela Fundação CPQD. Quatro estilos de fala foram gravados: animado, acolhedor, ríspido e neutro. O estilo animado se caracteriza por uma fala alegre, que transmita energia positiva. O acolhedor como uma fala calma, tranquila e compreensiva. O ríspido como alguém irritado e que cobra o interlocutor por algo. E por fim o neutro, somente com a leitura do texto desejado. A volumetria da base disponível é apresentada na Tabela 1.

As falas gravadas consistem em amostras de áudio em arquivos *wav* à 22KhZ, conjuntamente com a transcrição fonética falada na amostra.

Estilo	Horas aproximadas
Neutro	11
Animado	2
Acolhedor	2
Ríspido	2

Tabela 1. Volumetria aproximada da base de dados por estilos (em horas).

3. Modelo

Modelos de conversão texto-fala frequentemente são divididos em dois módulos principais, o modelo acústico e o vocoder. O modelo acústico é responsável por mapear o texto de entrada (ou sua transcrição fonética) no mel-espectrograma da fala sintetizada. Mel é uma escala perceptiva que divide o espectro em bandas para representar os tons como se fossem iguais em distância um do outro, levando em consideração que o ouvido humano não percebe frequências em uma escala linear [11]. Essa escala mapeia o espectro para que as variações tonais sejam percebidas linearmente pelos humanos. Para gerar de fato o sinal de fala é necessário mapear este mel-espectrograma em amostras de áudio e o vocoder é responsável por isso.

Nosso modelo acústico proposto (Figura 1) tem como base a arquitetura *FastPitch* proposto por [13].

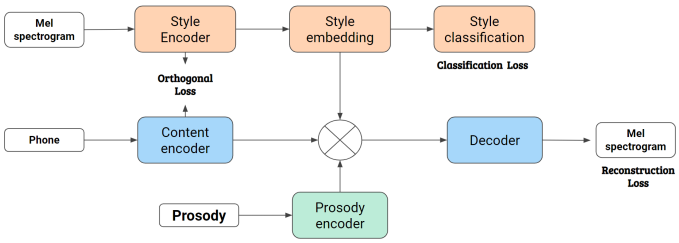

Figura 1. Arquitetura do modelo proposto.

Essa arquitetura consiste em uma camada de *encoder* que codifica os fonemas de entrada em representações densas de dimensão 384. Essas representações são então utilizadas por 3 módulos distintos, pelo preditor de duração, responsável por predizer a duração que cada fonema terá na fala sintetizada, pelo preditor de pitch, responsável por prever a curva de frequência fundamental da fala sintetizada, e por fim pelo *decoder*, que recebe a soma residual dessas representações e as predições dos demais módulos para predizer o mel-espectrograma da fala sintetizada. Adicionalmente, utilizamos o *Reference Encoder* proposto por [9], para gerar representações de estilos dos mel-espectrogramas de referência, onde tais representações são somadas a saída do *encoder* e utilizadas para condicionar os estilos na fala sintetizada. Como forma de gerar representações separáveis entre os diferentes estilos, uma camada de classificação é adicionada após o codificador de estilo, responsável por classificar os estilos a partir das representações geradas [12]. Além disso, como uma forma de induzir o modelo a não codificar o conteúdo da frase de entrada conjuntamente com o estilo, utilizamos uma função de perda que mede a ortogonalidade entre as representações originadas do codificador de estilo e do codificador de fonemas, similar ao proposto por [5].

Para condicionar um estilo específico na fala sintetizada condicionamos o *decoder* ao centroide das regiões de cada estilo no espaço gerado pelas representações, similar ao proposto por [4]. Por fim, para gerar as amostras de fala, utilizamos um vocoder capaz de mapear o mel-espectrograma em amostras de áudio, baseado na arquitetura *HiFi-GAN* [3].

3.1. Experimentos

Realizamos ao todo cinco experimentos (Tabela 2). As duas primeiras baseadas em técnicas mais simples, em uma criamos modelos especialistas para cada estilo baseado no ajuste fino das arquiteturas nos dados expressivos (Vanilla 1), e em outra utilizamos um mapeamento simples, conhecido como *look-up table*, para condicionar cada estilo a partir de camadas de embedding simples (Vanilla 2).

Nome	Descrição
Vanilla 1	Ajuste fino
Vanilla 2	Mapeamento de estilo por tabela
Baseline	Não supervisionado
Experimento 1	Proposto não balanceado
Experimento 2	Proposto balanceado

Tabela 2. Descrição dos experimentos realizados.

Para os demais experimentos, inicializamos as arquiteturas com pesos de um *FastPitch* pré-treinado, e realizamos os ajustes nas camadas adicionais que o modelo proposto possui por 200k iterações. Para uma primeira abordagem, adicionamos o *Reference Encoder* de forma simples, e sem supervisão no treinamento (Baseline). Já os Experimentos 1 e 2, consistem na arquitetura proposta apresentada, com a função de perda de ortogonalidade e classificação, treinadas em duas partições de dados diferentes, uma com os dados totais disponíveis, e outra com uma partição balanceada (mesma quantidade de horas para cada estilo).

4. Resultados

Para analisar as representações de estilos utilizamos a técnica *UMAP* [6] e projetamos a representação 2D delas. Para todos os experimentos a fala sintetizada final é inteligível e sem presença de ruídos.

O espaço gerado pela Baseline pode ser observado na Figura 2. Nota-se que para os diferentes estilos as representações se concentram em diferentes regiões do espaço. No entanto, é possível observar alguns pontos sobrepostos entre os estilos. É interessante observar que houve uma separação mais clara entre o estilo neutro (em verde) e os demais.

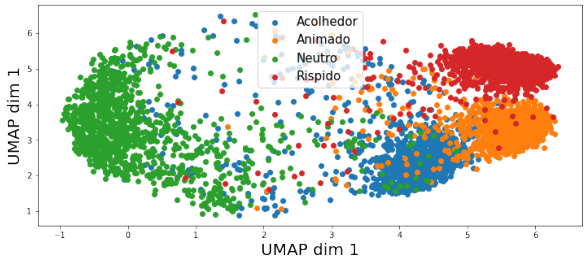

Figura 2. Espaço de estilos da Baseline.

Já para o Experimento 1 (Figura 3), nota-se que os diferentes estilos apresentam clusteres bem separados, com exceção de poucos pontos sobrepostos ao estilo neutro na região central da figura.

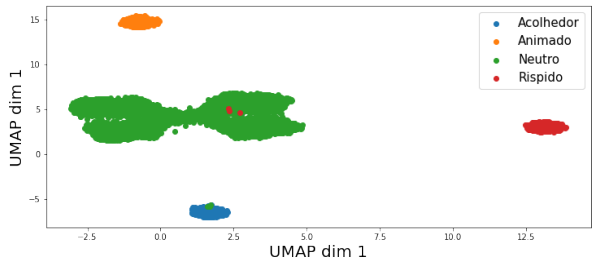

Figura 3. Espaço de estilos do Experimento 1.

Por fim, o uso de dados balanceados no Experimento 2 (Figura 4) modelou claramente quatro clusteres distintos para cada um dos estilos presentes na base.

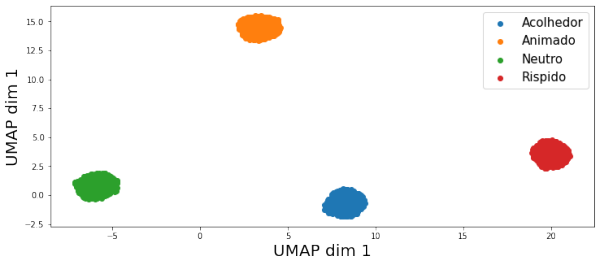

Figura 4. Espaço de estilos do Experimento 2.

Para todos os experimentos, a fala sintetizada final continua inteligível e com qualidade próxima a das falas gravadas. As abordagens baseadas em representações (Baseline, Experimentos 1 e 2), conseguem gerar modulações diferentes na fala final ao se condicionar a diferentes regiões do espaço de estilos, enquanto que nos dois experimentos Vanilla o condicionamento é único para cada estilo.

5. Conclusões

O problema de se modelar fala expressiva em arquiteturas de conversão texto-fala é complexo e possui diferentes abordagens na literatura. Neste trabalho propomos o uso de duas delas, a modelagem explícita de componentes prosódicos utilizando a arquitetura *FastPitch*, e o uso de um codificador de estilos (*Reference Encoder*). Além disso, propomos o uso de duas funções de perda para auxiliar as representações geradas pelo *Reference Encoder*, a de classificação e a de ortogonalidade entre o conteúdo e o estilo. Os resultados mostraram que de fato, o uso desses dois componentes adicionais auxiliam a arquitetura em gerar representações mais separadas entre os diferentes estilos. No entanto, para avaliar a capacidade de cada modelo gerar os estilos alvos desejados ainda é necessário uma avaliação perceptual.

5.1. Próximos passos

A importância das representações geradas serem desembaraçadas de outras informações é, particularmente, importante quando se deseja transferir as representações para outros locutores (transferência de estilo), dito isso, um próximo passo é realizar experimentos com uma base multi-locutor e testar se é possível transferir as representações de um estilo para um locutor que nunca gravou dados naquele estilo.

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] Plínio A. Barbosa. *Prosódia*. Parábola, May 2019.
[2] F. Eyben, S. Buchholz, N. Braunschweiler, J. Latorre, V. Wan, M. J. F. Gales, and K. Knill. Unsu-

pervised clustering of emotion and voice styles for expressive TTS. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4009–4012, March 2012. ISSN: 2379-190X.

[3] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17022–17033. Curran Associates, Inc., 2020.

[4] O. Kwon, I. Jang, C. Ahn, and H. Kang. An Effective Style Token Weight Control Technique for End-to-End Emotional Speech Synthesis. *IEEE Signal Processing Letters*, 26(9):1383–1387, September 2019. Conference Name: IEEE Signal Processing Letters.

[5] Tao Li, Xinsheng Wang, Qicong Xie, Zhichao Wang, and Lei Xie. Cross-speaker emotion disentangling and transfer for end-to-end speech synthesis, 2021.

[6] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426 [cs, stat]*, September 2020. arXiv: 1802.03426.

[7] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning. *arXiv:1710.07654 [cs, eess]*, February 2018. arXiv: 1710.07654.

[8] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783, 2018.

[9] R. J. 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. *arXiv:1803.09047 [cs, eess]*, March 2018. arXiv: 1803.09047.

[10] Alexander Sorin, Slava Shechtman, and Ron Hoory. Principal Style Components: Expressive Style Control and Cross-Speaker Transfer in Neural TTS. In *Interspeech 2020*, pages 3411–3415. ISCA, October 2020.

[11] S. S. Stevens, J. Volkmann, and E. B. Newman. A Scale for the Measurement of the Psychological Magnitude Pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190, January 1937. Publisher: Acoustical Society of America.

[12] Lucas H. Ueda, Paula D. P. Costa, Flavio O. Simoes, and Mário U. Neto. Are we truly modeling expressiveness? A study on expressive TTS in Brazilian Portuguese for real-life application styles. In *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, pages 84–89, 2021.

[13] 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, 2021.