Computer Science and Application 计算机科学与应用, 2020, 10(12), 2159-2167
Published Online December 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.1012227

Hans汉斯

基于变分自编码的语气语音合成模型

王　研，吴怡之

东华大学，信息科学与技术学院，上海
Email: 1559679340@qq.com, yz_wu@dhu.edu.cn

收稿日期：2020年11月20日；录用日期：2020年12月3日；发布日期：2020年12月10日

摘　要

语气作为一种重要情感表达信息，对说话人内容的表达起着重要作用。目前语音合成系统缺乏对语气的良好支持，合成语音也表现出乏味、单一的缺点。为了解决上述问题，提高合成语音的自然度，本文将统计参数语音合成(Statistical Parameter Speech Synthesis, SSPS)与具有强学习能力的变分自编码(Variational Autoencoder, VAE)模型相结合，以无监督的方式学习说话人潜在的语气信息，再通过加入分类器提高模型语气学习的准确率。我们提出了语气语音合成的系统框架，分为三部分：声学模型、语气模型以及合成模型。待合成的目标文本和语气分别利用声学模型与语气模型重构出的包括基频F0的声学特征。最后，将声学特征输入到WORLD声码器合成出带有目标语气的语音信号。本篇文章使用Blizzard Challenge 2018作为模型训练的语料库，最后通过实验结果表明，所提出的模型具有良好的语气生成性能。

关键词

语气，Variational Autoencoders，语音合成，WORLD声码器

A Speech Synthesis Model with Mood Based on Variational Autoencoder

Yan Wang, Yizhi Wu

College of Information Science and Technology, Donghua University, Shanghai
Email: 1559679340@qq.com, yz_wu@dhu.edu.cn

Received: Nov. 20th, 2020; accepted: Dec. 3rd, 2020; published: Dec. 10th, 2020

Abstract

Mood as the important emotional expression information plays an important role in the expression of the speaker's content. The current speech synthesis system lacks good support for mood and synthetic speech also shows the shortcomings of monotonous and boring. In order to solve the above problems and improve the naturalness of the synthesized speech, we use Statistical Para-

文章引用: 王研, 吴怡之. 基于变分自编码的语气语音合成模型[J]. 计算机科学与应用, 2020, 10(12): 2159-2167.
DOI: 10.12677/csa.2020.1012227

meter Speech Synthesis (SSPS) and Variational Autoencoder (VAE) model with strong learning ability to learn the speaker's potential mood information in an unsupervised manner, and then improve the accuracy of model mood learning by adding classifiers. We propose a systematic framework for speech synthesis with mood, which is divided into three parts: an acoustic model, a speech mood model, and a synthetic model. The target text and mood to be synthesized are reconstructed using the acoustic features including the fundamental frequency F0 using the acoustic model and the mood model, respectively. Finally, the acoustic features are input into the WORLD vocoder to synthesize speech signals with target mood. This article uses Blizzard Challenge 2018 as a corpus for model training, and finally, the experimental results show that the proposed model has a good performance for mood generation.

Keywords

Mood, Variational Autoencoders, Speech Synthesis, WORLD Vocoder

1. 引言

语音合成又称为文语转换技术(Text-to-Speech Synthesis, TTS) [1]，是一项广泛应用于文本阅读器与人机交互的技术。因而语音合成技术愈发受到研究学者的关注，语音合成的方法从最早的拼接法语音合成[2]和统计参数语音合成[3]，到近年来以直接输入文本或注音字符指导语音合成的端到端语音合成系统不断发展，使得合成语音的质量不断提高，几乎接近于人类的语音质量[4]。但尽管如此，目前合成的语音普遍缺乏情感色彩以及语气信息。在日常对话中，语气(Mood)是表示说话人对某一行为或事物现象的看法和态度[5]。在不改变句子所陈述内容事实的同时可以给话语信息带来广泛的情感色彩。中英语气大致可以分为陈述句、疑问句、感叹句、祈使句，当然每种语气可以细分到更小的粒度，如疑问句可以分为是非疑问句，特指疑问句等[5]。而本文主要讨论陈述语气、感叹语气与疑问语气。

为了提高合成语音的自然度，论文[6]从情感方面，提出一种迁移学习与自学习情感表征的情感语音合成方法，构建出端到端情感语音合成器，实现个性化语音合成。同时论文[7]中采用 VAE 模型分别用来对噪声、基频、说话能量和音素时长进行学习，实现说话人的语音控制。目前对语音合成的研究大多在情感以及韵律、音调等进行研究，但是合成语音在语气上的表达没有考虑在内。所以本文主要研究目标是在语音合成系统中将语气信息考虑在内，对输入文本，实现特定的语气输出，达到语气控制。

相比论文[8]中将 VAE 引入端到端语音合成，虽然可以达到有效合成带有风格的语音，提高自然度。但是对于端到端语音合成的方法存在需要大量语料库训练、模型复杂度高、合成语音速度缓慢等缺点。所以在本文中，我们主要采用统计参数语音合成方法，其系统主要由文本分析、声学建模、声码器三个模块组成。在声学建模的训练过程，我们加入语气模型的训练并引入变分自编码模型(Variational Autoencoder, VAE) [9]以无监督学习的方式捕捉声学特征 F0 的语气信息，更好的模拟说话人个性的整体语言特性。提出了一个称为语气语音合成的系统，包括模型训练与合成两部分(详细介绍见 2.2 节)，尝试在不改变语义信息表达的情况下，实现带有语气的语音合成。最后通过主观与客观的评测方法对语气语音合成的性能进行评估。

2. 模型

在这一部分，我们首先介绍与回顾一下 VAE 的概念以及模型的相关细节，并且给出关于损失函数的理解。最后对于我们提出的语气语音合成系统进行论述。

2.1. Variational Autoencoder (VAE)

与生成式对抗网络(Generative Adversarial Networks, GAN)一样，VAE 是生成模型[10]的一种，其主要目标是以无监督学习方式[11]对输入高维数据信息的特征进行表征学习[12]，并且最终让生成数据和原始数据分布的分布尽可能接近。VAE 是由 Kingma 等人于 2014 年提出[9]，如图 1 为经典的隐变量概率模型示意图。

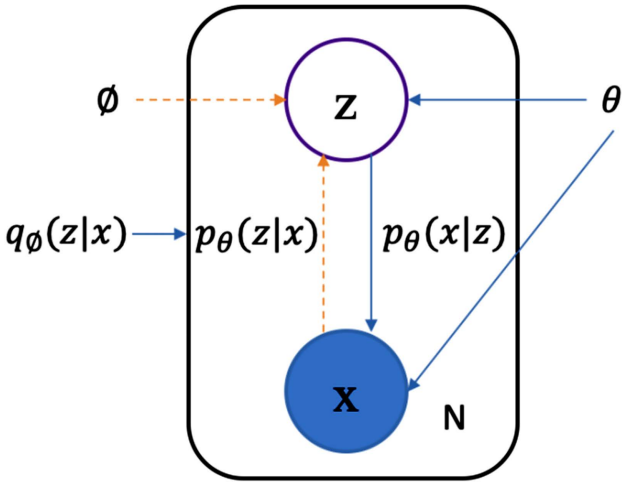


Figure 1. The graphical representation of VAE: where X is the observed data, N is the number of samples, Z is the latent variable, $\varnothing$ is the variable fraction parameter, and θ is the model parameter
图 1. VAE 的概率图模型表示：其中 X 为观测数据，N 为样本数，Z 为隐变量，$\varnothing$为变分参数，θ 为模型参数

橙色线部分表示推断过程，通过编码器(Encoder)实现。是对观测数据 X 与隐变量 Z 进行建模，目的是确定关于隐变量 Z 的后验分布概率分布 $p_\theta(z|x)$：

$$p_\theta(z|x)=\frac{p_\theta(Z)p_\theta(x|z)}{p_\theta(x)} \tag{1}$$

但是对于真实的后验概率分布 $p_\theta(z|x)$难以显式计算[13]，而变分推断(Variational Inference) [14]提供了此类问题的解决方法。通过引入识别模型 $q_\varnothing(z|x)$ 去逼近无法确定的后验分布 $p_\theta(z|x)$。对于识别模型与真实后验分布的匹配程度，可以通过 *Kullback-Leibler* (*KL*)散度[15]来衡量优劣。那么，对于概率分布 $p(x)$与 $q(x)$的 *KL* 散度定义为：

$$D_{KL}(p\|q)=-\int p\log\frac{q}{p}\mathrm{d}y \tag{2}$$

对于蓝色线部分表示生成过程，可以通过解码器(Decoder)实现。VAE 框架中对于隐变量采用了重参数技巧(Reparameterization Techniques)，作为神经网络的编码器根据样本 x_i 学习计算出均值 $\mu(x_i)$与方差 $\mu(x_i)$，所以我们通过从噪声分布采样得到隐变量：

$$z_i=\mu(x_i)+\delta\cdot\sigma(x_i),\ \delta\sim\mathcal{N}(\delta;0,1) \tag{3}$$

而生成模型的目标是希望由推断产生的隐变量变分概率分布拟合出观测数据 x 的近似概率分布 $p_\theta(x|z)$，即：

$$p_\theta(x|z)=\prod_{i=1}^{N}\mathcal{N}\left(x_i;\mu(z_i),\sigma(z_i)\right) \tag{4}$$

所以编码器与解码器的模型结构实现了推断与生成过程，通过联合训练学习，只要优化变分下界 ELBO (Evidence lower bound Objective)：

$$\mathcal{L}(\theta,\varnothing;x)=-KL\left[q_\varnothing(z|x)\|p_\theta(z)\right]+\mathbb{E}_{q_\varnothing(z|x)}\left[\log p_\theta(x|z)\right] \tag{5}$$

ELBO 中第一项是编码器推断近似后验分布的误差称为 *KL* 损失，而第二项起着解码器的作用，是对隐变量 z 进行重构出观测数据 x 的重构误差。

2.2. 基于 VAE 的语气语音合成模型

本文为了实现有效表征学习说话人的语气特征，我们将 VAE 模型引入语气语音合成系统。论文[16]提出了一种基于 VAE 的语音合成基频离散化表征方法，但无法有效进行特定语气的合成。为此，本文进一步提出基于 VAE 的带有分类器辅助的基频离散化学习方法，进行语气语音合成。如图 2 所示，该系统分为三个部分。

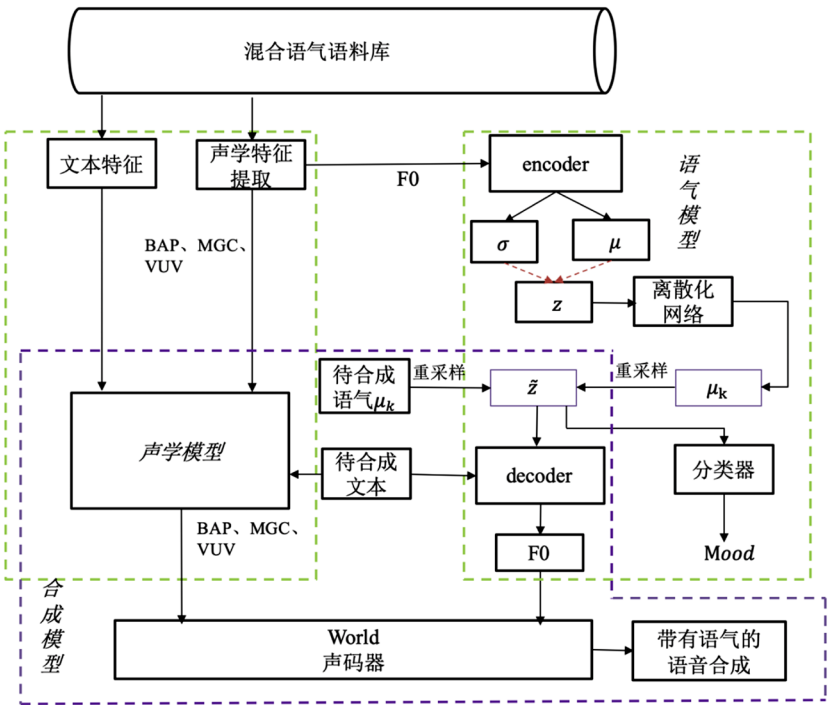


Figure 2. VAE-based speech synthesis framework with mood
图 2. 基于 VAE 的语气语音合成框架

本模型包括模型训练和目标语音的合成两个阶段。其一在模型训练部分，如图 2 绿色框图部分所示，包括声学模型的训练和语气模型的训练。声学模型是以待合成文本特征作为输入，输出是合成语音所必须的声学特征，如谱参数 MGC (Mel-Generalized Cepstral)、BAP (Band Aperiodicity, BAP)、VUV (Voice, Un Voice)及基频参数 F0 (Fundamental Frequency)等[17]。而声学模型的训练，是以语料库提取的声学特征[16]作为训练目标，对声学参数分布进行建模。声学模型算法从传统的基于 HMM 到基于

神经网络的声学模型，发展为近年来的基于端到端的声学模型，使得统计参数语音合成逐渐发展，而本文主要采取神经网络声学模型[18]。通常声学特征 F0 反映出说话人的音调高低，其曲线的包络在语气控制中起着尤为重要的作用，所以提出的语气模型是以 F0 对语气进行建模，达到合成语音带有语气控制。

语气模型中使用了典型的编－解码结构，编码器(Encoder)可以学习输入输入数据的潜在特征，把数据压缩至潜在空间表示，而解码器(Decoder)则根据学习到的低维特征，从潜在空间中重构出原始数据。将提取到的声学特征 F0 作为编码器的输入，通过编码器学习并计算出隐变量 z 的均值 μ 和方差 σ。对于 k 个不同种类的语气，预期学习到 k 个不同组合的均值与方差，即 k 个高斯分布代表不同语气。如前所述，根据重参数技巧，可以从噪声分布中进行采样得到表征语气的隐变量，得到的 Z 的分布应该近似逼近混合高斯模型，计算两者 KL 散度为：

$$D_{kL} = KL\left[p_\theta(z) \| \sum_{k=1}^{k} \pi_k \mathcal{N}(x \mid \mu_k, \sigma_k) \right] \tag{6}$$

我们通过训练离散化的神经网络得到表征不同语气的均值 μ_k，最终利用重采样技巧得到隐变量 $\tilde{z}$。将隐变量$\tilde{z}$经过解码器重构得到声学特征 F0，即所谓的生成模型。与此同时为了提高模型学习到语气信息的准确度，我们将采样得到的隐变量 $\tilde{z}$ 作为已经训练好的分类器输入，计算其输出目标语气的概率：

$$S(m_i) = \frac{e^{m_i}}{\sum_{j=1}^{n} e^{m_j}} \tag{7}$$

其中 n 为语气标签类别个数，因此整个语气模型的损失函数为：

$$Loss = KL\left[q_\varnothing(z \mid x) \| p_\theta(z) \right] - \mathbb{E}_{q_\varnothing(z|x)} \left[\log p_\theta(x \mid z, t, m) \right] - S(x_i) + D_{kL} \tag{8}$$

损失函数中的第一项为编码器推断近似后验分布的误差，第二项是通过隐变量 z、文本 t 以及语气 m 重构出观测数据的误差，当然 $S(x_i)$ 为分类器的损失函数，D_{kL} 为隐变量 z 与高斯混合模型的 KL 散度。

其二紫色框图为语音合成部分，将待合成文本输入到训练好的声学模型生成合成所需要的声学特征，同时待合成文本与语气模型中重采样得到表征语气信息的隐变量 $\tilde{z}$ 一同输入到解码器重构 F0。最后利用声码器，例如 WORLD [19]，实现从声学特征中重构出语音波形，产生带有语气的语音合成。

3. 实验与结果分析

3.1. 数据库

基于语气语音合成系统训练的语料库要求语音数据在语气种类上尽量丰富，有着丰富情感色彩，并存在与语音数据相对应的文本。我们采取 Usborne 公司出版发行的语料库 Blizzard Challenge 2018 [20]，该语音共包含 6.5 小时，约 7250 个句子的专业录音。演讲者的名字叫 Lesley Sims，她是一位以英语为母语的女性，通过以故事表达的方式讲述给 4~6 岁的观众。这种带有丰富变化语气的语音更加有利于 VAE 在其潜在空间中捕捉到更多的语气变化，学习各种语气种类的特征。考虑到语气系统的复杂性，加之时间有限，故而本文主要考虑对陈述语气、疑问语气以及感叹语气在语音合成中实现语气语音合成。为了有效地生成隐变量混合高斯模型，提高模型学习语气的确率，模型中引入了分类器进行语气判断。因此，数据预处理阶段，我们需要对语音文本自动进行标签，区分出语气种类即陈述语气、感叹语气以及疑问语气。如表 1 为所举样例文本：

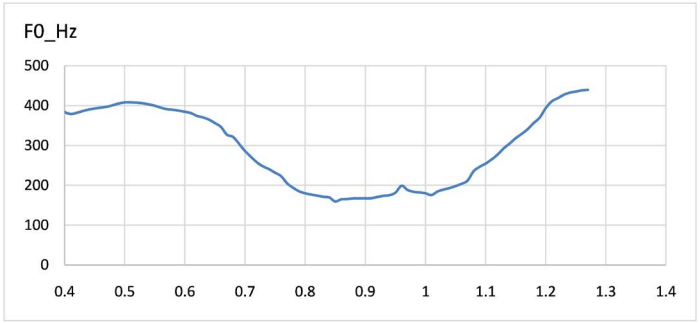
Table 1. Text pre-processing
表 1. 语音文本预处理

序号	语气种类	文本示例
S1	疑问语气	Do you remember me?
S2	疑问语气	Do you want to be eaten?
S3	陈述语气	A man in a blue tunic nodded to the Duke.
S4	陈述语气	He followed Titania as she walked through the forest, nagging her until the Queen's head ached.
S5	感叹语气	I love you!
S6	感叹语气	Oh, no!

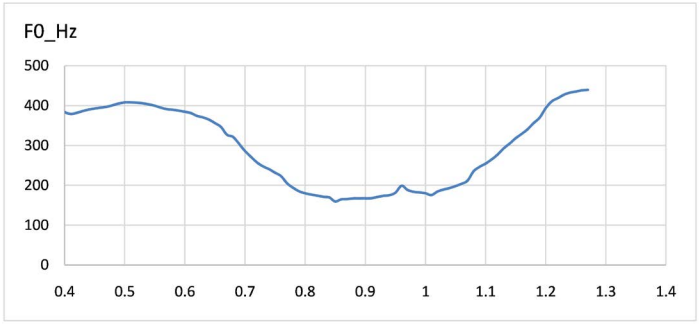
3.2. 实验评估

实验在 Ubuntu18.04 系统下，基于 PyTorch 平台实现对模型的训练。对于提出的语气模型具有自动编码器结构，通过对 F0 特征进行有效信息编码学习和重建出目标语气 F0。MLPG [21]使用全局标准差来生成 F0 轮廓，最后使用 WORLD 声码器从声学特征中恢复语音波形。

我们从语料库中选取了三种不同语气的 F0 轮廓图，如图 3~5 可以看出不同种语气的 F0 轮廓图呈现出不同的趋势。图 3 中图 3(a)，图 3(b)分别对应表 1 中 S1 与 S2，疑问语气句尾呈现语调上升趋势，侧重表达疑问的内容。对于图 4 中图 4(a)，图 4(b)分别对应表 1 中 S3 与 S4，除去异常时间数据下 F0 值，陈述句往往表现出语气舒缓，平稳的特征。图 5 所示图 5(a)，图 5(b)分别对应表 1 中 S5 与 S6，感叹语气 F0 轮廓与陈述语气大有不同，表现出起起伏伏，感情色彩强烈。

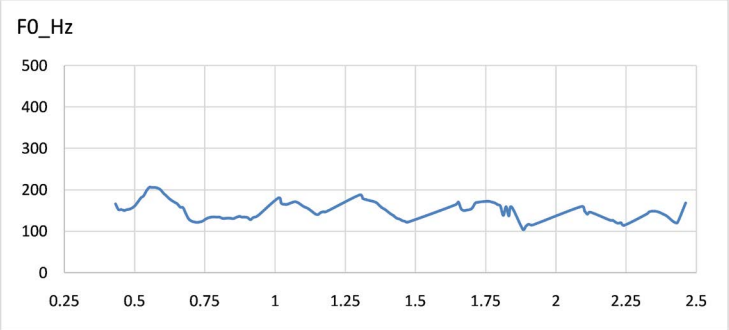


(a)

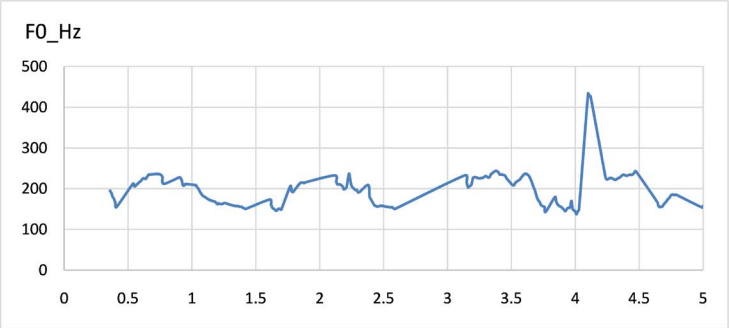


(b)

Figure 3. F0 contour diagram of the interrogative mood
图 3. 疑问语气 F0 轮廓图

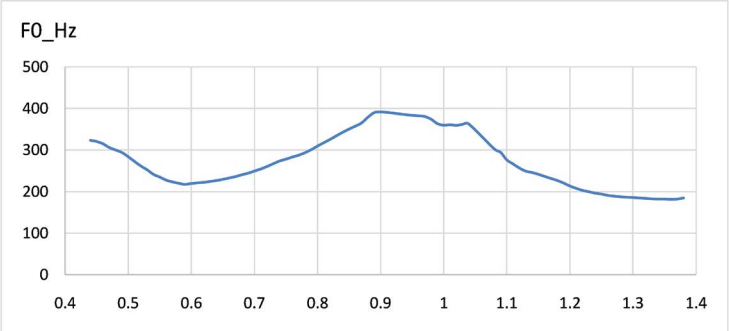


(a)

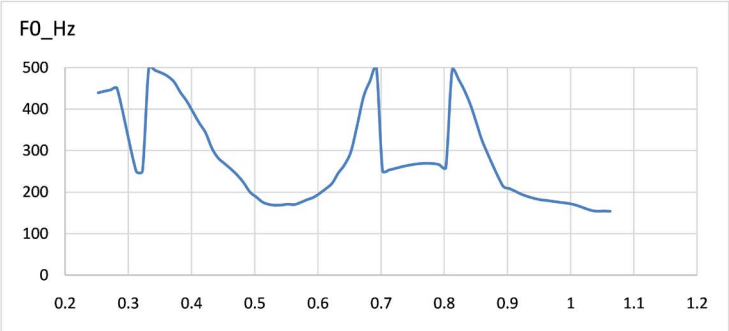


(b)

Figure 4. F0 contour diagram of the indicative mood
图 4. 陈述语气 F0 轮廓图



(a)



(b)

Figure 5. F0 contour diagram of the exclamatory mood
图 5. 感叹语气 F0 轮廓图

本文从主观与客观两个方面对语气语音合成性能做出评估。如图 6 所示为待合成文本在提出的系统下合成的目标语音的声学特征 F0 曲线即合成语音的客观评价。其中，易知语句 1 为陈述句：“This is an Usborne audio production.”；语句 2 为疑问句：“Is this part of the show?”；语句 3 为感叹句：“Oh dear!”。可以看出不同的语气具有明显不同的基频特性，反映了说话者语调的变化。并且其基频轮廓特性和图 3 至图 5 相应的语气呈现一致的轮廓。实验结果表明，提出的语气语音合成系统可以合成带有语气的语音，并且在不同语气上有较好的区分度。

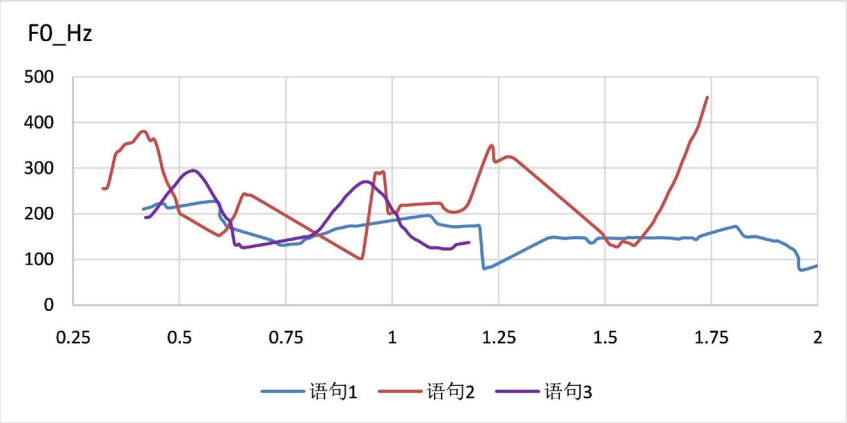


Figure 6. F0 contours for different mood
图 6. 合成不同语气的 F0 图

最后，我们同时也采用了主观意见评分(Mean Opinion Score, MOS)对本语气语音合成系统进行进一步的性能评估。MOS 是语音合成领域测评中主流使用的评分方法之一，且因为语音合成的目的与 MOS 方法都是基于评测人的主观感受，所以具有说服性。我们随机选取 15 句进行语音合成实验，4 位评测人进行主观打分(分值为 1~5，保留一位小数)，最后计算加权平均数。如图 7 所示，4 位评价者对合成的语音 MOS 打分较为接近，评价较为一致。表明合成的语音带有明显语气节奏，可懂度高。

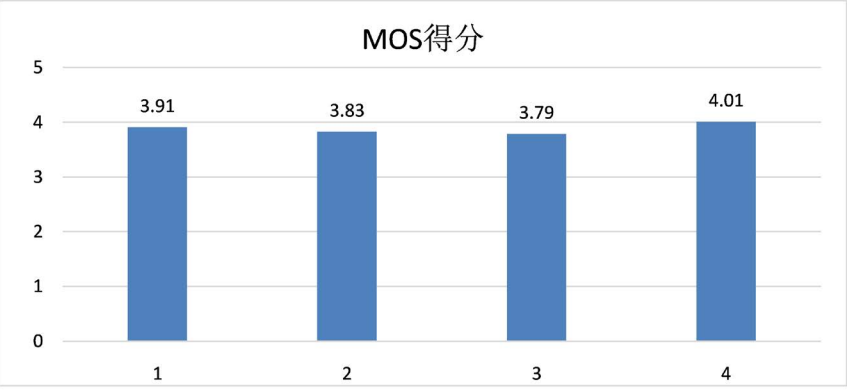


Figure 7. Subjective evaluation of results
图 7. 主观评测结果

4. 总结

在本篇文章中，我提出一个语气语音合成系统模型，主要包括语气模型、声学模型、合成模块。我们主要对于表征语气信息 F0 进行建模，希望解决当前语音合成系统在语气输出上的短板。我们通过 VAE 网络

学习句子潜在的语气信息，通过离散化神经网络与分类器提高模型学习的准确率。最后通过实验表明，对于待合成的文本输入，借助 WORLD 声码器重构语音波形，可以输出预期目标语气的合成语音，更具表现力。整个系统可以达到很好的性能。未来的下一步工作会增加研究语气的种类与中英文语气语音合成的实现。

参考文献

[1] Dutoit, T. (2001) An Introduction to Text-to-Speech Synthesis. Kluwer Academic Publishers, Dordrecht.

[2] Gonzalvo, X., Tazari, S., Chan, C.A., *et al*. (2016) Recent Advances in Google Real-Time HMM-Driven Unit Selection Synthesizer. *Interspeech* 2016, San Francisco, 8-12 September 2016, 2238-2242. https://doi.org/10.21437/Interspeech.2016-264

[3] Zen, H., Agiomyrgiannakis, Y., Egberts, N., *et al*. (2016) Fast, Compact, and High Quality LSTM-RNN Based Statistical Parametric Speech Synthesizers for Mobile Devices. Interspeech 2016, San Francisco, 8-12 September 2016, 2273-2277. https://doi.org/10.21437/Interspeech.2016-522

[4] Li, N., Liu, S., Liu, Y., *et al*. (2018) Close to Human Quality TTS with Transformer.

[5] 王飞华. 汉英语气系统对比研究[D]: [博士学位论文]. 上海: 复旦大学出版社, 2005.

[6] 张亚强. 基于迁移学习和自学习情感表征的情感语音合成[D]: [硕士学位论文]. 北京: 北京邮电大学, 2019.

[7] Sun, G., Zhang, Y., Weiss, R.J., *et al*. (2020) Fully-Hierarchical Fine-Grained Prosody Modeling for Interpretable Speech Synthesis. *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), Barcelona, 4-8 May 2020, 6264-6268. https://doi.org/10.1109/ICASSP40776.2020.9053520

[8] Zhang, Y.J., Pan, S., He, L., *et al*. (2019) Learning Latent Representations for Style Control and Transfer in End-to-End Speech Synthesis. 2019 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), Brighton, 12-17 May 2019, 6945-6949. https://doi.org/10.1109/ICASSP.2019.8683623

[9] Kingma, D.P. and Welling, M. (2014) Auto-Encoding Variational Bayes. 2*nd International Conference on Learning Representations*, Banff, 14-16 April 2014.

[10] Wittrock, M.C. (2010) Learning as a Generative Process. *Educational Psychologist*, **45**, 40-45. https://doi.org/10.1080/00461520903433554

[11] Bishop, C.M. (2006) Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, New York.

[12] Bengio, Yoshua, Courville, *et al*. (2013) Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **35**, 1798-1828. https://doi.org/10.1109/TPAMI.2013.50

[13] Khurana, S., Joty, S.R., Ali, A., *et al*. (2019) A Factorial Deep Markov Model for Unsupervised Disentangled Representation Learning from Speech. 2019 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), Brighton, 12-17 May 2019, 6540-6544. https://doi.org/10.1109/ICASSP.2019.8683131

[14] Wainwright, M.J. and Jordan, M.I. (2008) Graphical Models, Exponential Families, and Variational Inference. *Foundations & Trends® in Machine Learning*, **1**, 1-305. https://doi.org/10.1561/2200000001

[15] Joyce, J.M. (2011) Kullback-Leibler Divergence. In: Lovric, M., Ed., *International Encyclopedia of Statistical Science*, Springer, Berlin, 720-722. https://doi.org/10.1007/978-3-642-04898-2_327

[16] Hodari, Z., Lai, C. and King, S. (2020) Perception of Prosodic Variation for Speech Synthesis Using an Unsupervised Discrete Representation of F0. 10*th International Conference on Speech Prosody*, Tokyo, 25-28 May 2020, 965. https://doi.org/10.21437/SpeechProsody.2020-197

[17] He, M., Deng, Y. and He, L. (2019) Robust Sequence-to-Sequence Acoustic Modeling with Stepwise Monotonic Attention for Neural TTS. *Interspeech* 2019, Graz, 15-19 September 2019, 1293-1297. https://doi.org/10.21437/Interspeech.2019-1972

[18] Xue, S. and Yan, Z. (2017) Improving Latency-Controlled BLSTM Acoustic Models for Online Speech Recognition. *IEEE International Conference on Acoustics*, *Speech and Signal Processing*, New Orleans, 5-9 March 2017, 5340-5344. https://doi.org/10.1109/ICASSP.2017.7953176

[19] Morise, M., Yokomori, F. and Ozawa, K. (2016) WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. *Ice Transactions on Information & Systems*, **99**, 1877-1884. https://doi.org/10.1587/transinf.2015EDP7457

[20] King, S., Crumlish, J., Martin, A. and Wihlborg, L. (2017) The Blizzard Challenge 2018. *Proc. Blizzard Challenge Workshop*, Hyderabad.

[21] Tokuda, K., Yoshimura, T., Masuko, T., *et al*. (2002) Speech Parameter Generation Algorithms for HMM-Based Speech Synthesis. *IEEE International Conference on Acoustics*, Orlando, 13-17 May 2002, 1315-1318.