Let There Be Sound: Reconstructing High Quality Speech from Silent Videos

Ji-Hoon Kim, Jaehun Kim, Joon Son Chung
Korea Advanced Institute of Science and Technology, Daejeon, South Korea
{jh.kim, kjaehun, joonson}@kaist.ac.kr

Abstract

The goal of this work is to reconstruct high quality speech from lip motions alone, a task also known as lip-to-speech. A key challenge of lip-to-speech is the one-to-many mapping caused by (1) the existence of homophenes and (2) multiple speech variations, resulting in a mispronounced and over-smoothed speech. In this paper, we propose a novel lip-to-speech system that significantly improves the generation quality by alleviating the one-to-many mapping problem. Specifically, we incorporate (1) self-supervised speech representations to disambiguate homophenes, and (2) acoustic variance information to model diverse speech styles. Additionally, we employ a flow based post-net which captures and refines the details of the generated speech. We perform extensive experiments and demonstrate that our method achieves the generation quality close to that of real human utterance. Synthesised samples are available at the demo page: https://mm.kaist.ac.kr/projects/LTBS.

1. Introduction

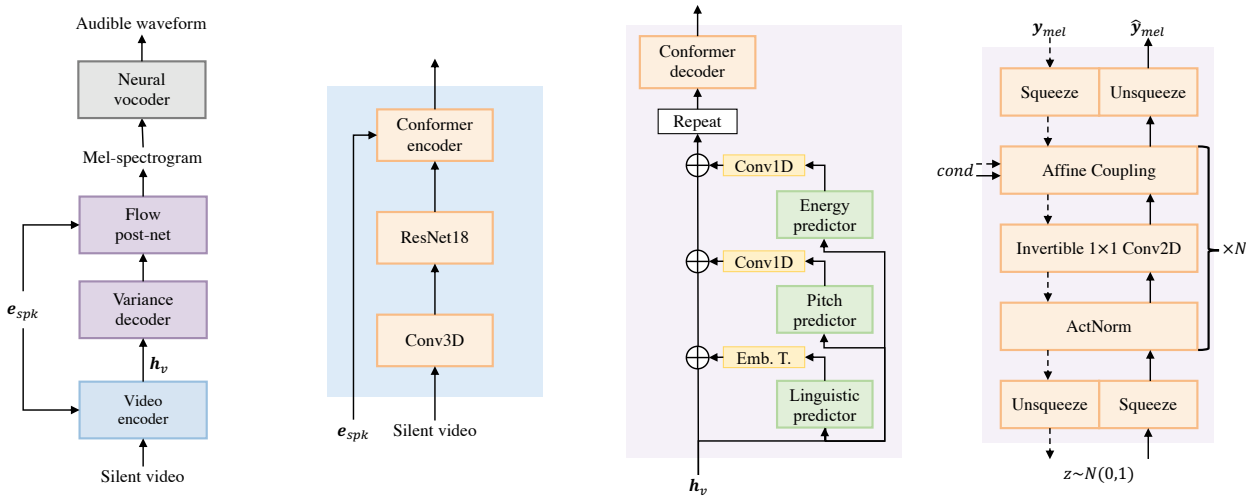
Have you ever wondered what Charlie Chaplin's movies would sound like if they weren't silent? Indeed, there have been many discussions about what is said in archival silent movies [32, 24]. The ability to reconstruct speech from silent videos opens up interesting applications, such as redubbing silent movies, simulating natural utterance for those who suffer from aphonia. As a result, the research in lip-to-speech has attracted an increasing amount of attention in recent years [8, 25]. This line of research has also benefited from the advances in deep-learning, in particular, self-supervised learning, since the training leverages natural occurrence of audio and video as a mode of supervision.

A lip-to-speech (LTS) system aims to learn a mapping from silent lip movements to the corresponding speech. This is a challenging one-to-many mapping caused by the two major obstacles. One is the existence of homophenes, words that have almost identical lip movements but distinct phonemes (e.g. 'bit' and 'pit'). The ambiguity of homophenes brings the one-to-many relationship between lip motions and phonemes [5, 18]. Another obstacle is the multiple variations in speech; same phonemes can be mapped to diverse speech styles based on individual characteristics such as timbre, intonation, and accents [7, 14].

Numerous attempts have been made to improve the quality of LTS systems. Early deep-learning based methods [9, 20] estimate linear predictive coding (LPC) features within a short video clip. Recently, many works [26, 16, 25, 17] adopt mel-spectrogram as a regression target because it contains more sufficient acoustic information than LPC. Despite the efforts, the previous methods do not fully address the one-to-many mapping issue, suffering from a mispronounced and over-smoothed synthetic speech.

In this work, we propose a novel LTS system that highly improves the synthetic quality by alleviating the intrinsic one-to-many mapping problem. To disambiguate homophenes, we employ self-supervised speech representations as a condition for linguistic information. Previous studies [1, 12] have proved that self-supervised learning (SSL) speech models can acquire rich speech representations without manually labeled text. In particular, it has been demonstrated that the representations from specific layers of the SSL model contain elaborate linguistic information independent of paralinguistic features [3]. Motivated by this, we explore the intermediate layers of SSL model and utilise the hidden representations to produce accurate content without using text labels.

Moreover, we adopt acoustic variance information in order to model diverse speech variations. With the help of the acoustic variations, the model can not only ease the one-to-many mapping but also learn prosody of speech which is a key factor for realistic speech synthesis [31, 33]. To further address the one-to-many mapping, we use a flow based post-net [29] which refines acoustic representations with enhanced modelling capability of capturing fine-grained details [30]. Combined with the variance information, the post-net helps to learn the complex one-to-many-mapping between phonemes and speech, thereby improving the naturalness of the synthesised speech. In the experiments, the effectiveness of our method is proved on various metrics.



(a) Overall architecture (b) Video encoder (c) Variance decoder (d) Flow post-net

Figure 1: In subfigure (a) and (b), $\boldsymbol{e}_{spk}$ is a speaker embedding. In subfigure (a) and (c), $\boldsymbol{h}_v$ denotes the encoded video feature. In (c), Emb.T. refers to an embedding table. In subfigure (d), the modules with dotted lines are operated only in a training stage. $\boldsymbol{y}_{mel}$ and $\hat{\boldsymbol{y}}_{mel}$ refer to the ground truth and predicted mel-spectrogram, respectively. $cond$ means the post-net conditions which contain the input and output of the conformer decoder, and $\boldsymbol{e}_{spk}$. In our experiment, we set $N = 8$.

2. Approach

Given a silent talking face video, our goal is to synthesise the corresponding mel-spectrogram. As shown in Figure 1(a), the proposed model mainly consists of three components: video encoder, variance decoder, and flow based post-net. The video encoder extracts distinct visual features from input videos, and the variance decoder successively produces coarse mel-spectrogram conditioned on linguistic features and acoustic speech variations. The flow based post-net elaborates the coarse mel-spectrogram into the fine-grained one, and the result is finally converted to an audible waveform by a pre-trained neural vocoder [15].

2.1. Video encoder

From the input video with T_v frames, the video encoder extracts distinct hidden representations $\boldsymbol{h}_v \in \mathbb{R}^{T_v \times d}$ where d denotes the hidden embedding dimension. As depicted in Figure 1(b), the video encoder comprises 3D convolution [13], ResNet18 [11], and a conformer encoder [10]. Since each individual has different visual characteristics, we inject speaker identity to the conformer encoder through an embedding table.

2.2. Variance decoder

To ease the one-to-many mapping problem in LTS, the variance decoder aims to generate acoustic representation with rich variance information. As shown in Figure 1(c), the variance decoder consists of variance predictors and a conformer decoder. The variance predictors are composed of linguistic, pitch, and energy predictor, each of which aims to condition the corresponding variance information into the $\boldsymbol{h}_v$. During training, we take the ground truth variance information to the hidden sequence, and use predicted value during inference. The following conformer decoder then converts the empowered hidden visual features to intermediate acoustic representations.

2.2.1 Linguistic predictor

The presence of homophenes hinders the synthesis of intelligible speech with accurate pronunciation [9]. Although the previous work [17] attempts to address the ambiguity of homophenes by leveraging text supervision, it requires manually annotated text labels and fails to enjoy the benefits of the LTS system. To generate intelligible speech while preserving the self-supervised nature, we propose a linguistic predictor that disambiguates homophenes without the need for text labels.

To this end, we adopt quantised self-supervised speech representations. We extract continuous linguistic representations from raw waveforms using pre-trained SSL speech model, namely HuBERT [12], and quantise the continuous representations for robust training[1]. Previous studies [21, 19] report that the quantised speech representations from the specific layer of HuBERT contain elaborate linguistic features relevant to accurate pronunciation. We in-

[1]Before quantisation, the continuous features are downsampled to match the length of video by nearest-neighbor interpolation.

vestigate the effects of different configurations of linguistic feature extraction, and empirically find that the representations from the 12^{th} layer of HuBERT-LARGE[2], quantised by K-means algorithm with 200 clusters, exhibits the highest correlation with linguistic information. The linguistic predictor is trained to estimate the cluster indices of each frame by cross-entropy loss between the target and predicted indices ($\mathcal{L}_l$).

2.2.2 Pitch predictor

Pitch plays an important role in synthesising realistic speech with natural prosody [22]. However, the pitch exhibits multiple variations across gender, age, and emotions, exacerbating the one-to-many problem in LTS. To accurately capture pitch information from lip motions, we construct a pitch predictor [22] that estimates the pitch sequence based on the hidden visual features.

Following [22], we extract the ground truth pitch values from the ground truth audio through `pYIN` algorithm, and standardise them to have zero mean and unit variance for better sampling. The extracted pitch values are successively downsampled to match the temporal dimension of the visual features. The pitch predictor is optimised with L1 loss between the target and predicted pitch sequence ($\mathcal{L}_p$).

2.2.3 Energy predictor

Energy represents the intensity of speech, which affects the volume and prosody of speech [2]. We obtain the target energy sequence by taking the L2-norm of the mel-spectrogram along the frequency-axis [4]. To estimate the energy sequence from $\boldsymbol{h}_v$, we construct energy predictor [28], which is optimised by L1 loss between the ground truth and predicted energy sequence ($\mathcal{L}_e$).

Each variance information is encoded into variance embeddings either through an embedding table (linguistic), or a single 1D convolution layer (pitch and energy). The variance embeddings are added to the visual representations $\boldsymbol{h}_v$, and the adapted representation is upsampled to match the time resolution of target mel-spectrogram. Lastly, the conformer decoder converts the adapted representations to a coarse mel-sepctrogram. We apply L1 loss between the ground truth mel-spectrogram and the predicted mel-spectrogram ($\mathcal{L}_{mel}$).

Note that all the variance predictors simplify the acoustic target distribution by providing conditional information, thereby mitigating the one-to-many mapping issue as proved in [30].

[2] https://huggingface.co/facebook/hubert-large-ls960-ft

2.3. Post-net

Natural human speech comes with dynamic variations. However, simple reconstruction loss (L1 or L2 loss) is limited to capture such details, resulting in a blurry and over-smoothed synthetic speech [23]. To further improve the sample quality, we apply a flow based post-net [29] which elaborates the coarse-grained mel-spectrogram into a fine-grained one.

The architecture of the post-net is depicted in Figure 1(d). In training stage, the post-net transforms mel-spectrogram training data $\boldsymbol{x}$ into a tractable prior distribution through a series of invertible functions, conditioned with the input and output of the conformer decoder, and the speaker embedding. The post-net is optimised with minimizing the negative likelihood of data $\boldsymbol{x}$ ($\mathcal{L}_{post}$). During inference stage, we take samples $\boldsymbol{z}$ from the prior distribution and feed them into the post-net reversely to generate the final mel-spectrogram. As proved in [30], this flow-based module enhances the capability of modelling complex data distributions, which helps to address one-to-many mapping.

To summarise, the final loss ($\mathcal{L}_{final}$) is given by:

$$\mathcal{L}_{final} = \mathcal{L}_{mel} + \lambda_{var}\mathcal{L}_{var} + \lambda_{post}\mathcal{L}_{post}, \qquad (1)$$

where $\mathcal{L}_{var} = \mathcal{L}_l + \mathcal{L}_p + \mathcal{L}_e$. In our experiments, we set $\lambda_{var} = \lambda_{post} = 0.1$.

3. Experimental Settting

We perform experiments on GRID [6] and Lip2Wav [26] video datasets. Audio data are resampled to 16kHz and transformed to mel-spectrograms with 40ms window length, 10ms hop length, and 80 mel filterbanks. To achieve the temporal synchronisation with the audio, the video data are resampled to 25 frames per second, resulting in a fixed ratio of 1 to 4 between the lengths of video and audio. We then extract 68 face landmarks for each video frame using FaceAlignment[3]. Based on the landmarks, the lip regions are aligned to a fixed position, and cropped to their centers with a dimension of 112×112. The cropped images are converted to grayscale.

We construct the identical model architecture for each dataset, with the exception of the conformer encoder. For the GRID dataset, the conformer encoder is designed with 6 attention heads and a hidden dimension of 384, and for Lip2Wav, the encoder is designed with 8 heads and a hidden dimension of 512. We follow the recent works [25, 17] for the configuration of 3D convolution and ResNet18. The pitch and energy predictors are composed of two 1D convolution layers [22], while the linguistic predictor consisted of four 1D convolutions.

We use AdamW optimiser with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-9}$. The learning rate is fixed to 2×10^{-4}. For

[3] https://github.com/ladrianb/face-alignment

Table 1: Evaluation results. MOS results are presented with 95% confidence interval. 'Nat.' and 'Intel.' represent MOS for naturalness and intelligibility, respectively. Note that Multi-task [17] cannot be trained on the Lip2Wav dataset since the model requires text transcription to be trained. ↑ denotes higher is better, ↓ denotes lower is better.

Method	GRID				Lip2Wav			
	Nat. ↑	Intel. ↑	WER(%) ↓	CER(%) ↓	Nat. ↑	Intel. ↑	WER(%) ↓	CER(%) ↓
Ground truth	4.82 ± 0.04	4.82 ± 0.05	12.20	7.03	4.80 ± 0.05	4.78 ± 0.05	4.12	2.58
Vocoded	4.74 ± 0.05	4.79 ± 0.05	12.44	7.18	4.63 ± 0.06	4.73 ± 0.06	6.05	4.19
VCA-GAN	3.46 ± 0.07	4.10 ± 0.08	17.62	9.55	2.05 ± 0.08	2.71 ± 0.10	48.73	32.51
SVTS	3.35 ± 0.09	3.97 ± 0.09	23.30	13.12	1.77 ± 0.07	2.18 ± 0.10	61.09	41.01
Multi-task	2.42 ± 0.09	3.08 ± 0.12	30.56	18.14	N/A	N/A	N/A	N/A
Proposed	$\mathbf{4.46 \pm 0.07}$	$\mathbf{4.63 \pm 0.07}$	**17.07**	**9.17**	$\mathbf{4.15 \pm 0.08}$	$\mathbf{3.69 \pm 0.10}$	**34.71**	**22.57**

training the GRID dataset, we randomly sample consecutive sequence with a length of 50, and the model is trained for 400 epochs. For the Lip2Wav dataset, we sample contiguous 75 frames and train our model for 900 epochs. We apply data augmentations: horizontal flipping with probability of 50%, and random masking with fixed position throughout all frames. The masked area is randomly sampled within the range from 10×10 to 30×30.

We compare our model against the ground truth, vocoded[4], and generated samples from recent LTS models which show promising results: VCA-GAN [16], SVTS [25], and Multi-task [17]. For a fair comparison, all the LTS models are trained on the same settings, and the predicted mel-spectrograms are converted to audible speech by pre-trained Fre-GAN vocoder [15].

4. Experimental Results

The performance of the proposed method is evaluated with both qualitative and quantitative evaluation metrics. For qualitative evaluation, we conduct mean opinion score (MOS) test, wherein 30 domain-expert speakers assess the quality of 30 random speech clips for naturalness and intelligibility on a scale of 1 to 5. Naturalness of speech represents how close the speech is to that of human utterance. Intelligibility focuses solely on the successful delivery of linguistic contents; high scores are given if one can clearly identify the contents even if it sounds unnatural. Moreover, we also compute word error rate (WER) and character error rate (CER) of 300 random samples for quantitative evaluation. For error calculation, we obtain the transcriptions of speech clips by using publicly available automatic speech recognition (ASR) model [27] pre-trained on 438k hours of English corpus.

[4]Vocoded speech refers to speech reconstructed from the ground truth mel-spectrogram through a vocoder, and thus it is practically considered the upper bound quality for our evaluation.

4.1. Quality comparison

As a qualitative evaluation, we compute MOS for naturalness and intelligibility. As shown in Table 1, our proposed method achieves the highest naturalness and intelligibility on both datasets. Especially, in GRID dataset, the generated speech of the proposed method closely approximates the vocoded quality with a minor gap of 0.28 in naturalness and 0.13 in intelligibility.

Moreover, as a quantitative metric, we compare the WER and CER of the synthesised speech with those of the ground truth and vocoded speech. For the GRID dataset, the error rates are obtained by directly comparing the ASR transcriptions with the provided ground truth texts. For the Lip2Wav dataset, since the dataset does not provide text labels, we manually annotate the ground truth texts and compare them with the ASR transcription results.

The results are shown in Table 1. The proposed model clearly shows the lowest WER and CER on both GRID and Lip2Wav datasets. This demonstrates our method can effectively reduce the homophene problems and synthesise highly intelligible speech with accurate pronunciation.

5. Conclusion

In this paper, we propose a novel LTS system that generates high-quality speech speech close to human-level quality in both naturalness and intelligibility. We directly tackle the inherent one-to-many mapping problems, and address them by providing linguistic and acoustic variance information. We further refine the generated speech by enhancing modelling capability. Both qualitative and quantitative experiments clearly demonstrate that the proposed method improves the overall quality of the synthesised speech, outperforming the previous works by a notable margin. For the future work, we will continue to enhance the generated speech quality by adopting audio-visual SSL models. We also aim to simplify the overall generation pipeline, making a fully end-to-end architecture.

References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. Wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*, 2020.

[2] Murtaza Bulut, Sungbok Lee, and Shrikanth S Narayanan. Analysis of emotional speech prosody in terms of part of speech tags. In *Proc. Interspeech*, 2007.

[3] Heng-Jui Chang, Shu-wen Yang, and Hung-yi Lee. Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert. In *Proc. ICASSP*, 2022.

[4] Hyeong-Seok Choi, Juheon Lee, Wansoo Kim, Jie Lee, Hoon Heo, and Kyogu Lee. Neural analysis and synthesis: Reconstructing speech from self-supervised representations. In *NeurIPS*, 2021.

[5] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *Proc. CVPR*, 2017.

[6] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006.

[7] Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Ye Jia, Ron J Weiss, and Yonghui Wu. Parallel tacotron: Non-autoregressive and controllable tts. In *Proc. ICASSP*, 2021.

[8] Ariel Ephrat, Tavi Halperin, and Shmuel Peleg. Improved speech reconstruction from silent video. In *ICCV Workshops*, pages 455–462, 2017.

[9] Ariel Ephrat and Shmuel Peleg. Vid2speech: Speech reconstruction from silent video. In *Proc. ICASSP*, 2017.

[10] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. In *Proc. Interspeech*, 2020.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016.

[12] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

[13] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2012.

[14] Changhwan Kim, Se Yun Um, Hyungchan Yoon, and Hong Goo Kang. Fluenttts: Text-dependent fine-grained style control for multi-style tts. In *Proc. Interspeech*, 2022.

[15] Ji-Hoon Kim, Sang-Hoon Lee, Ji-Hyun Lee, and Seong-Whan Lee. Fre-GAN: Adversarial Frequency-Consistent Audio Synthesis. In *Proc. Interspeech*, 2021.

[16] Minsu Kim, Joanna Hong, and Yong Man Ro. Lip to speech synthesis with visual context attentional gan. In *NeurIPS*, 2021.

[17] Minsu Kim, Joanna Hong, and Yong Man Ro. Lip-to-speech synthesis in the wild with multi-task learning. In *Proc. ICASSP*, 2023.

[18] Minsu Kim, Jeong Hun Yeo, and Yong Man Ro. Distinguishing homophenes using multi-head visual-audio memory for lip reading. In *Proc. AAAI*, 2022.

[19] Felix Kreuk, Adam Polyak, Jade Copet, Eugene Kharitonov, Tu-Anh Nguyen, Morgane Rivière, Wei-Ning Hsu, Abdelrahman Mohamed, Emmanuel Dupoux, and Yossi Adi. Textless speech emotion conversion using discrete and decomposed representations. In *Proc. EMNLP*, 2022.

[20] Yaman Kumar, Rohit Jain, Khwaja Mohd Salik, Rajiv Ratn Shah, Yifang Yin, and Roger Zimmermann. Lipper: Synthesizing thy speech using multi-view lipreading. In *Proc. AAAI*, 2019.

[21] Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, et al. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354, 2021.

[22] Adrian Łańcucki. Fastpitch: Parallel text-to-speech with pitch prediction. In *Proc. ICASSP*, 2021.

[23] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *Proc. AAAI*, 2022.

[24] Neil Midgley. New technology catches hitler off guard. *Telegraph*, 2006.

[25] Rodrigo Mira, Alexandros Haliassos, Stavros Petridis, Björn W Schuller, and Maja Pantic. Svts: Scalable video-to-speech synthesis. In *Proc. Interspeech*, 2022.

[26] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. Learning individual speaking styles for accurate lip to speech synthesis. In *Proc. CVPR*, 2020.

[27] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*, 2023.

[28] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *Proc. ICLR*, 2021.

[29] Yi Ren, Jinglin Liu, and Zhou Zhao. Portaspeech: Portable and high-quality generative text-to-speech. In *NeurIPS*, 2021.

[30] Yi Ren, Xu Tan, Tao Qin, Zhou Zhao, and Tie-Yan Liu. Revisiting over-smoothness in text to speech. In *Proc. ACL*, 2022.

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

[32] Jack Smith. When lip reading between the lines had the subtitles beat by a long sight. *LA Times*, 1987.

[33] Guangzhi Sun, Yu Zhang, Ron J Weiss, Yuan Cao, Heiga Zen, and Yonghui Wu. Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis. In *Proc. ICASSP*, 2020.