Audio-Visual Speech Codecs: Rethinking Audio-Visual Speech Enhancement by Re-Synthesis

Karren Yang[1] Dejan Marković[2] Steven Krenn[2] Vasu Agrawal[2] Alexander Richard[2]
[1]MIT [2]Meta Reality Labs Research
karren@mit.edu {dejanmarkovic,stevenkrenn,vasuagrawal,richardalex}@fb.com

Abstract

Since facial actions such as lip movements contain significant information about speech content, it is not surprising that audio-visual speech enhancement methods are more accurate than their audio-only counterparts. Yet, state-of-the-art approaches still struggle to generate clean, realistic speech without noise artifacts and unnatural distortions in challenging acoustic environments. In this paper, we propose a novel audio-visual speech enhancement framework for high-fidelity telecommunications in AR/VR. Our approach leverages audio-visual speech cues to generate the codes of a neural speech codec, enabling efficient synthesis of clean, realistic speech from noisy signals. Given the importance of speaker-specific cues in speech, we focus on developing personalized models that work well for individual speakers. We demonstrate the efficacy of our approach on a new audio-visual speech dataset collected in an unconstrained, large vocabulary setting, as well as existing audio-visual datasets, outperforming speech enhancement baselines on both quantitative metrics and human evaluation studies. Please see the supplemental video for qualitative results[1].

1. Introduction

Humans have the remarkable ability to extract speech content from visual information such as lip movement. Studies show that viewing speakers' faces improves human listening in noisy environments [41, 57], and that individuals naturally learn to read lip movements when their hearing is impaired [24]. Inspired by these observations, audio-visual speech enhancement methods leverage the visual input of a speaker to isolate their voice in a noisy environment [35]. By integrating facial frames of a target speaker with a noisy audio spectrogram, for example, recent deep learning models can generate a mask for the spectrogram that suppresses irrelevant voices and background sounds from the output [1, 11, 18]. These models prove useful for reducing noise and improving speech intelligibility of videos for downstream applications.

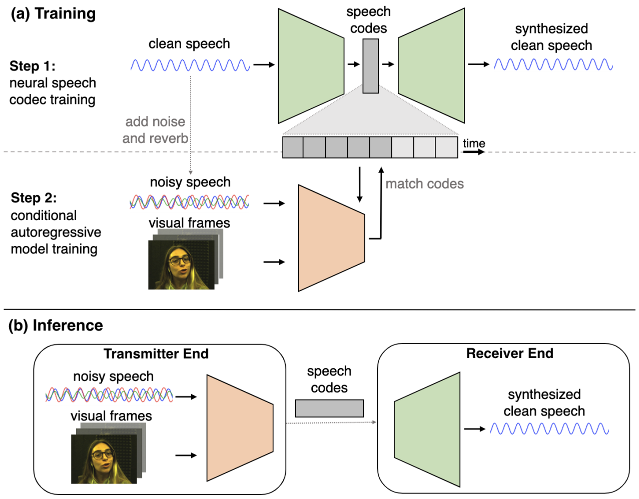

Figure 1. **Audio-Visual Speech Codecs.** Our model performs speech enhancement by leveraging audio-visual speech cues to synthesize the discrete codes of a neural speech codec. (a) During training, we first learn a codebook of natural speech for a target speaker by training a neural speech codec to compress and decode their clean speech signal. We then train an auto-regressive probabilistic model over the codes conditioned on noisy audio and visual inputs. (b) During inference, we use the auto-regressive model to generate a sequence of speech codes, which are then synthesized into speech using the decoder module of the speech codec.

However, there are a growing number of telecommunications applications where the quality and realism of the output speech, beyond speech intelligibility, are paramount. One example is social telepresence in AR/VR, which aims to enable realistic face-to-face conversations between people in a virtual setting [34, 62]. Immersive virtual conversations require extremely high-quality speech signals: each speaker's voice must sound clean and realistic when rendered in the virtual environment, as if a real conversation were taking place there. Current state-of-the-art methods fall short of these applications for two main reasons. First,

[1] https://github.com/facebookresearch/facestar/releases/download/paper_materials/video.mp4

they generate speech by using the noisy audio as a template [1, 11, 18] rather than by explicitly modeling the distribution of speech, which can lead to bleed-through noise and other unnatural distortions that disrupt the sense of immersion. Second, they focus on learning audio-visual speech cues that generalize well across a large population, but that may fail to capture speaker-specific cues needed for a higher-fidelity model [45].

Main Contributions. In this work, we take a different approach from existing work that overcomes these two limitations. Our main contributions are the following:

(1) We propose **audio-visual (AV) speech codecs**, a novel framework for AV speech enhancement. Rather than using noisy audio input as a template for producing enhanced output, AV speech codecs explicitly model the distribution of speech and re-synthesize clean speech conditioned on audio-visual cues. Our approach is summarized in Figure 1. During training, we first learn the building blocks of natural speech by training a neural speech codec to compress and decode clean speech signal through a discrete codebook. Subsequently, we learn an auto-regressive probabilistic model over the codes conditioned on noisy audio and visual inputs. At test time, we obtain speech codes from the auto-regressive model and use the decoder module of the neural speech codec to synthesize clean speech. Our approach is analogous to high-quality, two-stage image generation techniques that learn a probabilistic prior over a pre-trained vocabulary of image components [13, 38].

(2) Rather than adopting a speaker-agnostic framework as done in most recent work, we focus on *personalized models* that leverage speaker-specific audio-visual cues for higher fidelity speech enhancement. To this end, we introduce **Facestar**[2], a high-quality audio-visual dataset containing 10 hours of speech data from two speakers. Existing audio-visual datasets used for vision-based speech synthesis tasks are either captured in a clean, controlled environment with a small and constrained vocabulary [7, 21] or curated from "in-the-wild" videos with variable audio quality and unreliable lip motion [45]. In contrast, Facestar contains unconstrained, large vocabulary natural speech recorded with high audio and visual quality and enables development of high-quality personalized speech models.

(3) Empirically, our personalized AV speech codecs outperform audio-visual speech enhancement baselines on quantitative metrics and human evaluation studies while operating on only 2kbps transmission rate from transmitter to receiver. To the best of our knowledge, our work is the first to enable audio-visual speech enhancement at the quality required for high-fidelity telecommunications in AR/VR, even when the transmitter is in a highly noisy and reverberant environment.

Addressing Scalability. Beyond introducing high-quality personalized AV speech codecs, we also take steps towards addressing their scalability. While personalized models are commonly used in high-fidelity applications– for example, personalized visual avatars enable extremely photo-realistic visual representations of humans in VR that overcome the uncanny valley [34, 62] – a downside is that they typically require training on hours of data from the target individual. The question naturally arises of how we can obtain high-quality personalized models with less data, in order to scale high-fidelity telecommunications to a large volume of users. As a first step, we propose a simple strategy for personalizing AV speech codecs to new individuals with minimal new data, based on a similar approach used to scale personalized text-to-speech models [4]. Specifically, we introduce a multi-speaker extension of AV speech codecs that features a speaker identity encoder, which can be pre-trained on a multi-speaker dataset and then fine-tuned for a new speaker with only a small sample of their data. We demonstrate this personalization strategy on the GRID dataset [7]. An additional benefit of our multi-speaker model is that it enables voice conversion from one speaker to another, and thereby opens up creative applications in AR/VR.

2. Related Work

Audio-only Speech Enhancement. The ability of humans to isolate a target speaker from a noisy environment [41, 57] has inspired extensive study into computational approaches for speech separation and enhancement. While early formulations of the problem assumed input from multiple microphones [9, 65], recent approaches have also considered the monaural setting [23, 50, 53, 54]. This includes monaural speech separation methods, which address the problem of separating a mixture of speakers from a single audio track [23, 36, 60, 66], as well as monaural speech enhancement methods, which tackle the problem of removing non-speech background sounds [8, 30, 42, 55, 56, 63] and reverberation [8, 56] from noisy speech. Our work also focuses on the task of monaural speech enhancement, but we diverge from audio-only studies in that we utilize an additional visual stream to guide speech synthesis.

Audio-Visual Source Separation. The correspondence between audio and visual cues in video has led to sound source separation approaches that leverage audio-visual information [14, 40, 47, 52]. Recently, deep learning frameworks have been developed for audio-visual separation for speech [1, 2, 6, 11, 15, 18, 39] and music [16, 64, 69, 70]. In the speech domain, these approaches rely on facial recognition [6, 18, 70] and/or lip motion [1, 18, 39] to suppress sounds that do not correspond to the speaker in the visual stream. We similarly consider the task of audio-visual speech enhancement, but our framework differs from these works in that we perform speech synthesis conditioned on the audio-visual inputs rather than using a sound separation framework (*e.g.*, generating a spectrogram mask). Our re-

[2] https://github.com/facebookresearch/facestar

sults show that our approach leads to higher-quality, more natural sounding speech output.

Neural Speech Codecs. Modern audio communication systems rely on speech codecs to efficiently compress and transmit speech. Neural speech codecs use neural networks to compress input speech into a low-bit rate representation that can be transmitted over a network and decoded into audio waveform on the receiver end [59]. The low-bit rate representation is typically a discrete representation of human speech learned through autoencoding [19, 26, 28, 38] or speech features obtained through self-supervised training [44]. While some of these approaches consider the impact of noisy speech [28, 67], their primary focus is compression of clean speech. In our work, we propose an audio-visual speech enhancement framework that generates the clean speech codes of a neural audio codec conditioned on noisy audio and visual inputs.

Neural Speech Synthesis. High-quality neural speech synthesis is generally based on a two-stage pipeline [29], which first generates a low-resolution intermediate representation of speech from the input and subsequently synthesizes audio waveform from this representation [29, 31, 37, 46]. Prominent examples of speech synthesis include text-to-speech synthesis and video-to-speech synthesis, where the first stage consists of generating the intermediate representation from text [33,43,51,61] or silent video [10,12,32,45] input. Our approach is similarly based on a two-stage pipeline, but we use learned speech codes from a neural speech codec as our intermediate representation, and we condition the generation of the speech codes on noisy audio and visual inputs, rather than text or silent videos.

3. Approach

Let $\mathbf{S}'$ and $\mathbf{V}$ respectively denote the audio and visual streams of an individual's speech, where $\mathbf{S}'$ contains various sources of environmental noise (*i.e.*, interfering speakers, background sounds, reverberation). Our goal is to synthesize a high-quality, clean version of the speech $\mathbf{S}$. Our approach consists of two learned components, which are summarized in Figure 2:

1. **Generative speech codes**: We learn a discrete speech codebook that captures a rich vocabulary of the target speaker's utterances, which can be used to synthesize high-quality audio in the speaker's voice.
2. **Conditional auto-regressive model**: Conditioned on the speaker's audio-visual inputs, we train an auto-regressive model to generate speech codes, ensuring that the sequence of codes follow the natural distribution of the speaker's speech.

Our framework is inspired by high-quality two-stage generative models, such as VQ-VAE [38], which first learn a discrete encoding of the data and subsequently learn to generate a code sequence using a probabilistic model. Here, we condition the code generation on the speaker's audio-visual inputs for speech synthesis.

3.1. Generative Speech Coding

We represent an individual's clean speech as a sequence of entries from a codebook $\mathcal{Q} = \{q_k\}_{k=1}^{K}$, where each q_k is an N-dimensional vector. Based on these codes, any clean speech segment $\mathbf{S} \in \mathbb{R}^T$ can be approximately synthesized from a code $\mathbf{Z} \in \mathbb{R}^{T' \times N}$, where T' is the temporal extend of the length T input audio. Since the encoder compresses the input signal along the temporal axis, we typically have $T' < T$. We first map from speech to codes using the encoder network $\tilde{\mathcal{E}}$, which operates on the mel-spectrogram representation of $\mathbf{S}$:

$$\tilde{\mathbf{Z}} = \tilde{\mathcal{E}}(\mathbf{melspec}(\mathbf{S})) \in \mathbb{R}^{T' \times K}. \tag{1}$$

Then, we transform this encoding into a sequence of T' codes by sampling from the Gumbel-softmax distribution [25] and selecting the code from $\mathcal{Q}$ with the corresponding index. For a temporal encoding $\mathbf{Z} = [\mathbf{Z}_1, \cdots, \mathbf{Z}_{T'}]$, the t-th code is therefore given by

$$\mathbf{Z}_t = q_k, \quad k = \text{Gumbel}(\tilde{\mathbf{Z}}_{t,1:K}). \tag{2}$$

We denote the transformation from continuous valued embeddings $\tilde{\mathbf{Z}}$ to codes $\mathbf{Z}$ by $\mathbf{h}_{\mathcal{Q}}$. In practice, a codebook that captures the full range of an individual's speech may require a prohibitively large number K of codebook entries. To increase the expressiveness of the speech codebook, we follow a commonly used concept of *multi-head* codes [25, 49] and replace each code in the codebook $\mathcal{Q}$ by a set of H subcodes

$$\mathcal{Q}^{(h)} = \{q_k^{(h)}\}_{k=1}^{\tilde{K}}, \quad h = 1, \ldots, H. \tag{3}$$

In other words, instead of using one large codebook of size K, we use H smaller codebooks of size $\tilde{K}$ each. This enables the size of our speech codebook to grow exponentially in H, *i.e.*, $K = \tilde{K}^H$, increasing the expressiveness of our speech codebook without an exponential increase in encoding size. In this case, the dimensionality of the encoding $\tilde{\mathbf{Z}}$ is $T' \times H \times \tilde{K}$, and each temporal code is

$$\mathbf{Z}_t = \left[q_k^{(h)}\right]_{h=1}^{H}, \quad k = \text{Gumbel}(\tilde{\mathbf{Z}}_{t,h,1:\tilde{K}}). \tag{4}$$

Finally, the decoder that reconstructs speech from the learned codes is composed of a mel-spectrogram decoder $\tilde{\mathcal{D}}$ followed by a neural vocoder $\mathcal{G}$ that transforms the decoded mel-spectrogram back into the wave-domain. Our speech codec architecture therefore consists of an encoder $\mathcal{E} = \mathbf{h}_{\mathcal{Q}} \circ \tilde{\mathcal{E}}$ that maps from mel-spectrograms of the input speech to codes, and a decoder $\mathcal{D} = \mathcal{G} \circ \tilde{\mathcal{D}}$ that maps

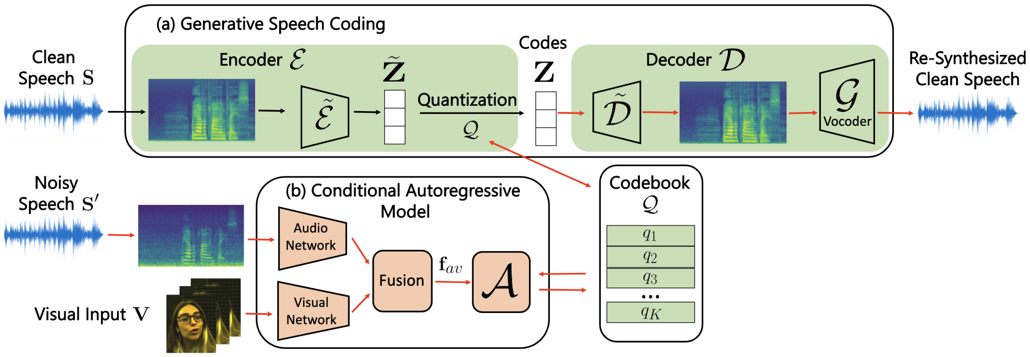

Figure 2. **Schematic of our approach**. (a) We train a Gumbel-Softmax autoencoder on the clean speech of a target speaker to obtain a personalized speech codec. The codec compresses speech into discrete codes that can be re-synthesized into the original speech with high fidelity (Section 3.1). (b) To synthesize clean speech from noisy input, we train an auto-regressive model to generate speech codes conditioned on the noisy audio and lip motion of the target speaker (Section 3.2). The red arrows indicate information flow at test time.

from codes back to speech. The codebook $\mathcal{Q}$ is learned by jointly optimizing it with the encoder $\mathcal{E}$ and the spectrogram decoder $\tilde{\mathcal{D}}$ to minimize the ℓ_2-loss between the mel-spectrogram of the clean speech signal and the reconstructed mel-spectrogram,

$$\mathbb{E}_{\mathbf{S}}||\mathbf{melspec}(\mathbf{S}) - \tilde{\mathcal{D}}(\mathbf{Z})||_2^2. \tag{5}$$

For the neural vocoder, *i.e*. the module that transforms the decoded mel-spectrograms back into a waveform, we use a HiFi-GAN [29] conditioned on the predicted spectrograms $\tilde{\mathcal{D}}(\mathbf{Z})$. The vocoder $\mathcal{G}$ is trained with a combination of a multi-scale GAN loss, a mel-spectrogram loss, and a feature-matching loss as described in Kong *et al*. [29]. We let $\mathcal{E}$ denote full transformation $\mathbf{S} \mapsto \mathbf{Z}$ and $\mathcal{D}$ denote the full generative speech model $\mathbf{Z} \mapsto \tilde{\mathbf{S}}$.

3.2. Conditional Auto-Regressive Modeling of Speech Codes

Having learned $\mathcal{E}$ and $\mathcal{D}$, any clean speech segment $\mathbf{S}$ can be represented by a sequence of codes $\mathbf{Z}$ that can be used to reconstruct the clean speech $\mathbf{S}$ through $\mathcal{D}$. Therefore, audio-visual speech enhancement can be formulated as an auto-regressive modeling problem in the latent space. Given a sequence of codes $\mathbf{Z}_{1:t-1}$, our auto-regressive model $\mathcal{A}$ predicts the distribution of the next codes conditioned on the corresponding audio-visual features $\mathbf{f}_t^{av}$. We use the categorization operator $\mathbf{h}_Q$ to map the log-probability values output by $\mathcal{A}$ to codes in $\mathcal{Q}$, *i.e*.

$$\mathbf{Z}_t = \mathbf{h}_Q(\mathcal{A}(\mathbf{Z}_{1:t-1}, \mathbf{f}_t^{av})). \tag{6}$$

The audio-visual network that extracts the audio-visual features $\mathbf{f}^{av}$ is composed of a visual stream, an audio stream, and an audio-visual fusion module. The visual stream takes $\mathbf{V}$ as input and produces an intermediate representation of visual features of dimension $T' \times H \times \tilde{K}$ that is useful for shaping speech synthesis. To produce the audio-visual features $\mathbf{f}^{av}$, the visual features are fused with the audio stream along the temporal axis, then passed through a fusion module. The auto-regressive model $\mathcal{A}$ and the audio-visual feature extraction network are optimized to minimize the error in the mel-spectrogram reconstruction,

$$\mathbb{E}_{(\mathbf{S},\mathbf{S}',\mathbf{V})}||\mathbf{melspec}(\mathbf{S}) - \tilde{\mathcal{D}}(\mathbf{Z})||_2^2. \tag{7}$$

where $\mathbf{S}'$ and $\mathbf{V}$ are the noisy input speech and the visual frames, and $\mathbf{Z}$ is the latent code obtained from the conditional auto-regressive model as in Equation (6).

Summary: Training vs. Inference. Training occurs through a two-stage procedure. First, we learn a discrete speech codebook by optimizing Equation (5) on clean speech data. Any sequence of latent codes can therefore be decoded into (clean) speech using the speech decoder $\mathcal{D}$. Second, we train an auto-regressive model over the codes by optimizing Equation (7) using noisy audio and visual data. Note that the speech codebook and decoder are fixed during this step. At inference time, information flow follows the red arrows in Figure 2. We first predict speech codes $\mathbf{Z}$ in an auto-regressive fashion through Equation (6). Then, the codes are passed through the speech decoder $\mathcal{D}$ to re-synthesize the output speech. Note the advantage of this two-step approach: since the decoder is trained on clean speech only and the speech codebook has limited capacity, the decoder is incapable of producing non-speech-like outputs. In contrast to existing approaches, bleeding-through of noise from the noisy input speech can therefore be avoided entirely.

4. Facestar Dataset

Existing audio-visual datasets tend to be either (i) captured in a clean, controlled environment with a small and constrained vocabulary such as GRID [7] or TCD-TIMIT [21]; or (ii) curated from “in-the-wild” videos with variable

AV Dataset	# Hours per Speaker	High-Quality Audio	Reliable Lip Motion	Unconstrained Natural Speech
GRID	0.8	✓	✓	✗
TCD-Timit	0.5	✓	✓	✗
Lip2Wav	20	✗	✗	✓
Facestar	5	✓	✓	✓

Table 1. **Comparison of different audio-visual speech synthesis datasets**. Our Facestar dataset contains significantly more content per speaker compared to GRID [7] and TCD-Timit [21] and contains higher-quality data compared to the YouTube-Lip2Wav [45].

audio quality and unreliable lip motion such as Lip2Wav [45]. The former datasets are constrained and do not cover the range of natural speaker content, while the latter does not have sufficiently high audio quality for training high-quality speech synthesis models. The necessity for a dataset with high-quality audio and video in a conversational, large-vocabulary setting becomes apparent for applications such as video calls, where background noise poses a significant limitation to a socially engaging experience. Therefore, we collect and introduce the Facestar dataset, which consists of 10+ hours of audio-visual speech data collected in a video-conferencing setting. The dataset was recorded in a low-noise acoustically treated 2.5m^3 environment with pseudo uniform lighting. Two participants, one male and one female, spoke freely in front of a video-conferencing device equipped with visual and audio sensors. For technical details see supplemental material. Each participant was captured for 5+ hours, resulting in 10+ hours of high-quality audio-visual data containing frontal face view and natural unconstrained speech that simulates typical video call settings. Table 1 shows a comparison of the Facestar dataset with existing single-speaker datasets used for audio-visual tasks. Compared to other datasets, the Facestar dataset contains unconstrained speech captured in a clean, video-conferencing like environment, providing the type of high-quality audio needed for training speech synthesis models.

5. Experiments on Single-Speaker Datasets

In this section, we train and evaluate personalized models on large, single-speaker datasets and demonstrate significant quantitative and perceptual gains over baseline methods trained on the same data. We introduce a multi-speaker extension that addresses scalability later in Section 6.

5.1. Evaluation Setup

Datasets. We train and evaluate our approach on the *Facestar* dataset described in Section 4. Additionally, we evaluate our approach on the *Lip2Wav* dataset [45]. This dataset consists of videos downloaded from YouTube channels with approximately 20 hours of speech available per speaker.

Model Architecture. We describe the architecture of the generative speech coding model (Figure 2a) and the conditional auto-regressive model (Figure 2b).

(a) Generative Speech Coding. Our encoder network $\mathcal{E}$ consists of a 1D convolutional block (512 filters, kernel size=5, stride=1) with batch normalization and ReLU activation, followed by three 1D residual blocks with the same hyperparameters. The projection to $\tilde{\mathbf{Z}}$ is performed by a final 1D convolutional layer (kernel size=1, stride=1) where the filter size depends on the size of the discrete latent space. For our discrete latent space, we use $\tilde{K} = 256, H = 4, N = 64$. The decoder network $\mathcal{D}$ resembles the encoder network, except there are three additional LSTM modules after the three residual blocks. For our neural vocoder $\mathcal{G}$, we use the architecture of HiFi-GAN [29].

(b) Conditional Auto-Regressive Model. Our audio processing network consists of a 1D convolutional block (512 filters, kernel size=5, stride=1) with batch normalization and ReLU activation, followed by three 1D residual blocks with the same hyperparameters. Our visual processing network consists of a 3D convolutional block (64 filters, kernel size=(5,7,7), stride=(1,2,2)) with batch normalization and ReLU activation, followed by a max pooling layer (kernel size=(1,3,3), stride=(1,2,2)). The resulting 4D tensor is passed through the feature extractor of a 2D ResNet [22], which acts independently on each temporal frame. The output is upsampled to match the temporal frequency of the processed audio and passed through a series of 1D convolutional and residual blocks, similar to the audio processing network. Finally, the audio and visual features are fused by matching their temporal axes and passed through an autoregressive module, which consists of two LSTMs with a PreNet (2 fully-connected layers with dropout) for processing previous frames [51], and a fully-connected layer for mapping the output of the LSTMs to codes.

Training Details. We train our model on 3-second clips of clean speech randomly sampled from the target dataset, which corresponds to 75-90 video frames. The video frames are pre-processed using the S^{3}FD face detector [68] to obtain face crops as done in Prajwal *et al.* [45]. To simulate noisy speech at training time, we first add room reverberation by convolving clean speech with impulse responses from the MIT Impulse Response Survey [58]. The MIT Impulse Response Survey consists of 270 impulse responses collected from different locations that volunteers encountered in their daily lives and reflect a variety of standard acoustic settings. Subsequently, we add random audio samples from the VoxCeleb2 [5] dataset to simulate interfering speakers, and random audio samples from Audioset [20] to simulate background noise. For the background noise, we adjust the signal-to-noise ratio randomly between 0 and 40 dB with respect to the clean signal.

5.2. Baselines

State-of-the-art deep learning-based AV speech enhancement approaches generally fall into two camps: *spectro-*

Model	PESQ ↑	STOI ↑	Facestar F-SNR ↑	MCD ↓	Mel-ℓ_2 ↓	PESQ ↑	STOI ↑	Lip2Wav F-SNR ↑	MCD ↓	Mel-ℓ_2 ↓
Demucs [8]	1.251	0.554	5.602	5.003	0.0106	1.383	0.672	7.644	4.724	0.0109
AV-Masking [18]	1.257	0.593	5.991	5.184	0.0093	1.438	0.689	7.873	5.167	0.0093
AV-Mapping [15]	1.332	0.626	2.802	4.885	0.0059	1.417	0.661	6.892	4.643	0.0062
Ours	**1.354**	**0.661**	**7.322**	**3.815**	**0.0056**	**1.482**	**0.740**	**8.801**	**4.072**	**0.0055**

Table 2. **Quantitative Evaluation of Audio-Visual Speech Separation and Enhancement.** Our approach consistently outperforms the baselines on both datasets. For PESQ, STOI, F-SNR, higher is better. For MCD and Mel-ℓ_2, lower is better.

gram masking approaches that leverage the visual modality to mask the noisy audio spectrogram [17, 18], and *direct mapping* approaches that leverage the audio and visual modalities to directly generate a denoised spectrogram [15]. We compare our approach to a model from each camp:

Audio-Visual Spectrogram Masking (AV-Masking). For the spectrogram masking baseline, we use a U-Net model that integrates visual information in the bottleneck layer and outputs a mask for a complex spectrogram [17, 18]. The mask is applied to the complex spectrogram of the input audio and the result is transformed back into time domain using an inverse FFT to generate the output waveform.

Audio-Visual Direct Mapping (AV-Mapping). For the direct mapping baseline, we use an encoder-decoder architecture that directly outputs a denoised spectrogram from audio-visual input and uses the Griffin-Lim algorithm to generate the output waveform [15].

In addition to these audio-visual baselines, we also compare against a well-known **Audio-Only Model (Demucs)** [8] to demonstrate the importance of the visual stream for the speech enhancement task. All models are trained and evaluated on the same datasets for fair comparison.

5.3. Quantitative Evaluation

Metrics. We evaluate all approaches using the following metrics: Perceptual Evaluation of Speech Quality, which measures speech quality (**PESQ**, higher is better); Short-Time Objective Intelligibility, which measures speech intelligibility (**STOI**, higher is better); Frequency-Weighted Segmental Signal-to-Noise Ratio (**F-SNR**, higher is better); and the ℓ_2 distance between the mel-frequency cepstrum coefficients and mel-spectrograms of predicted and ground truth audio (**MCD**, **Mel-**ℓ_2, lower is better).

Results. The results on Facestar and Lip2Wave are shown in Table 2. Our approach outperforms the baselines on all of the objective metrics used to evaluate enhanced speech. The audio-visual models (AV-Masking, AV-Mapping, ours) generally outperform the audio-only model (Demucs), demonstrating the importance of leveraging information from the visual modality regardless of the specific speech enhancement framework.

5.4. Human Evaluation Studies

Although the objective metrics shown in Table 2 are widely used in literature, it is important to note that no objective metric precisely reflects how humans perceive speech quality [27]. Therefore, we also conduct a user study using the Facestar dataset to compare our model results to the two audio-visual baselines as well as ground truth recordings. In the study, participants were presented two clips of the same sequence from two different approaches. The clips were presented in random order to ensure an unbiased evaluation, and participants were asked to decide which of the two clips sounded more natural, with three answer options: first clip, second clip, or can not tell. Overall, 100 participants ranked around 1000 clips. The results in Table 3 show that our approach is strongly preferred over the baselines. Notably, over 50% of the time, study participants could not differentiate the outputs of our model from the ground truth recording, which indicates the high quality of our personalized approach.

Ours	GT recordings	Can not tell
4.1%	44.5%	51.4%

Ours	AV Encoder Decoder	Can not tell
73.3%	6.0%	20.7%

Ours	AV Masking	Can not tell
78.5%	5.7%	15.8%

Table 3. **Perceptual Evaluation**. Participants were presented two video clips and asked to tell which of the two sounds more natural.

Model	reverb + noise + interfering spkr	only reverb + noise
Vision-Only	0.0085	
Audio-Only	0.0091	0.0056
No Auto-Regressive Module	0.0051	0.0036
Full Model	**0.0043**	**0.0033**

Table 4. **Ablation Results**. The values shown are the mean ℓ_2 errors between predicted and ground truth mel-spectrograms for ablation models trained on the Facestar dataset (Speaker 1); lower is better. See text for details.

5.5. Ablation Studies

Importance of Visual Modality. Since the noisy audio input contains significant corruptions (*i.e.*, overlapping speakers, high-intensity noise, reverberation), the visual modality is key to synthesizing the target speech components. Table 4 shows ablation results for a vision-only model (row 1) and audio-only model (row 2) compared to the full audio-visual model (row 4). Note that model performance declines significantly without the visual modality (compare row 4 to row 2). The visual modality also plays a larger role when the noisy audio contains interfering speakers (column 1) compared to when it contains only background noise and reverberation (column 2), as visual information is needed to disambiguate between speakers. To further support this point, in the presence of interfering speakers, we find that

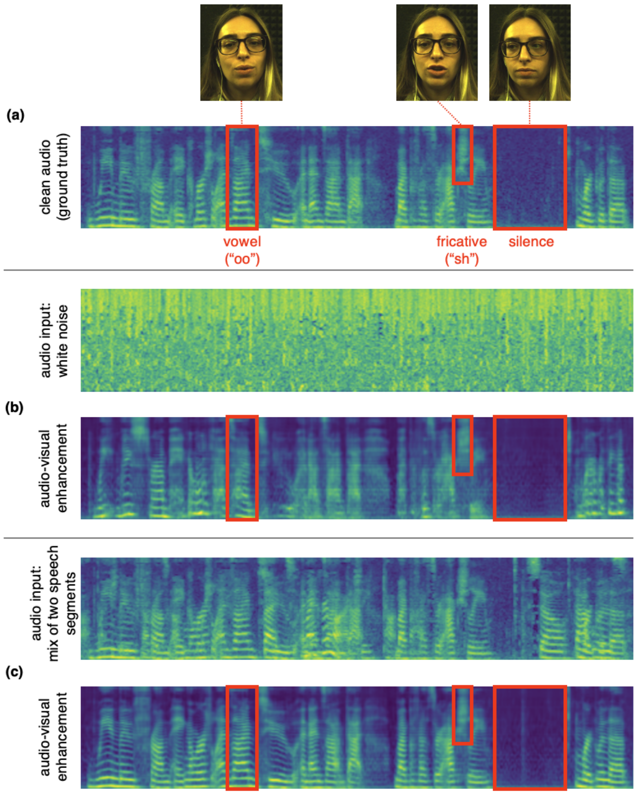

Figure 3. **Importance of visual modality.** (a) Ground truth mel-spectrogram and frames from visual input corresponding to specific vocal sounds. (b) Presented with white noise as audio input, the model relies on the visual modality to synthesize speech. (c) Presented with a speech mixture from the same speaker as audio input, the model relies on the visual modality to disambiguate and separate speech signals. See supplemental video for examples.

the visual-only model outperforms the audio-only model (compare row 1 to row 2, column 1).

Figure 3 illustrates how the visual modality is used by our model. The visual input of the target speaker contains specific mouth articulations that correspond to sounds (*e.g.*, vowels and fricatives) in the ground truth speech (Figure 3a). Our model is able to approximate these sounds from the visual cues alone, when only white noise is provided to the audio stream (Figure 3b). This suggests that the visual modality is primarily responsible for defining the structure of the synthesized speech from our model. When presented with noisy audio input, our model is able to further refine the pitch of the synthesized speech, since this information is not available from visual cues (Figure 3c). Note that the noisy audio example in Figure 3c contains interfering speech from the same speaker, *i.e.* real and interfering speech share the *same voice* and the visual modality is required in order to disambiguate between target and interfering speech signals. The model successfully leverages the visual cues to suppress the distractor speech and only keeps the original speech signal.

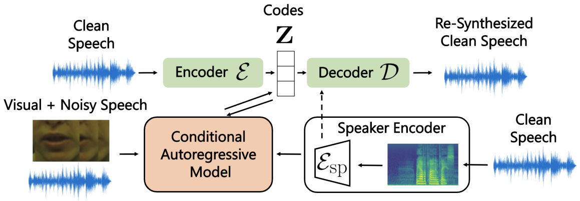

Figure 4. **Multi-speaker model**. A speaker encoder is added to the pipeline from Figure 2. Restricting the size of the codebook forces the model to disentangle speech content and speaker identity as shown in [48].

	GRID Speaker			
	Sp. 1 (M)	**Sp. 3 (M)**	**Sp. 11 (F)**	**Sp. 15 (F)**
Single-speaker model	0.00509	0.00794	0.00746	0.00781
Multi-speaker model	0.00657	0.00909	0.00960	0.01594
Multi-speaker model personalized to new speaker with k minutes of data				
5 min	0.00481	0.00682	0.00625	0.00681
12.5 min	0.00457	0.00620	0.00589	0.00655
25 min	0.00443	0.00595	0.00570	0.00621
50 min	0.00425	0.00561	0.00553	0.00596

Table 5. **Performance of multi-speaker models that are personalized to new speakers by fine-tuning on different quantities of target speaker data**. The personalized (*i.e.*, fine-tuned) models outperform the single-speaker models even when the amount of data for fine-tuning is greatly reduced. Values shown represent mean ℓ_2 distances between predicted and ground truth mel-spectrograms; lower is better.

Importance of Auto-Regressive Modeling. Our conditional auto-regressive model generates speech codes conditioned on previous speech codes, ensuring that the sequence of speech codes is temporally consistent. To determine the contribution of this model component, we perform an ablation study using a model without the auto-regressive component. As shown in Table 4 (compare row 3 to row 4), this leads to a significant decrease in model performance.

Efficiency. Discretized neural speech codecs generally allow for highly efficient signal transmission. For instance, Soundstream [67] demonstrates compelling speech reconstruction with bitrates from 6-12kbps. In contrast, our approach is a personalized model and therefore allows for an even stronger data compression with no or barely noticeable loss of quality. Results shown in the supplemental video have a bitrate of 2kbps, which we found to be sufficient for personalized speech reconstruction.

6. Scalability with Multi-Speaker Extension

So far we have demonstrated the efficacy of AV speech codecs as a personalized model when they are trained on single-speaker datasets comprised of hours of data. In real-world applications involving a high volume of telepresence users, however, one must be able to obtain high-quality personalized models with less individualized data. In this section, we extend our work to the multi-speaker setting, demonstrating two advantages to such a model: (i) efficient

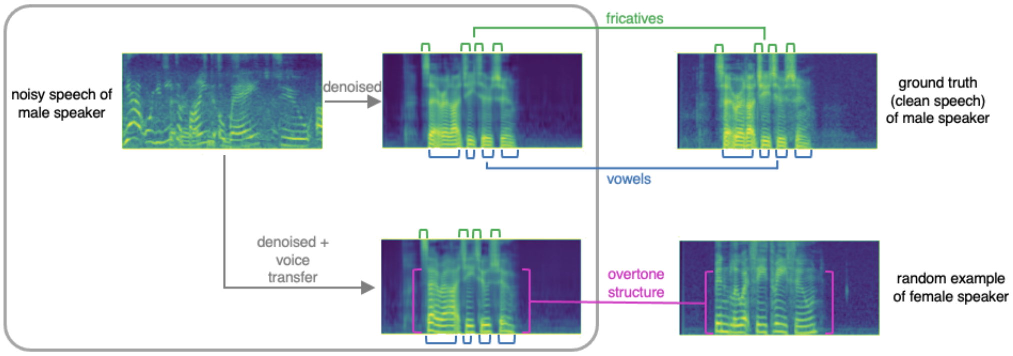

Figure 5. **Voice Transfer Examples.** By swapping the speaker code at the decoder stage, we can synthesize clean audio in a different target speaker's voice. Images shown are mel-spectrogram representations of audio. Note how the linguistic content (*i.e.*, vowels and fricatives) are carried over from the original male speaker, while the pitch and overtone structure are changed to that of the female speaker.

model personalization and (ii) voice controllability.

Efficient Model Personalization. For practical applicability, there is a need for personalized speech models that can synthesize speech for new users given a small sample of audio-visual data. We extend our framework to the multi-speaker setting by adding a speaker identity encoder to the model as shown in Figure 4, following the approach of personalized text-to-speech synthesis models [4]. Our multi-speaker AV speech codec can be pretrained on a larger, multi-speaker dataset and finetuned on a small amount of a new speaker's data to personalize the model to their speech. Table 5 shows the results of pretraining on 30 speakers from the GRID dataset [7] and finetuning on different amounts of data for four held-out speakers (last four rows). For comparison, we train four single-speaker models on all data for the held-out speakers (row 1). We find that personalizing our multi-speaker model to new speakers significantly reduces the amount of data needed to achieve the same performance of a single-speaker model: with only 5 minutes of data per speaker, the fine-tuned multi-speaker model already outperforms the pure single-speaker models. Note that this performance gain is in fact attributed to the fine-tuning process; the multi-speaker model alone does not generalize well to held-out speakers and is therefore strictly worse than single-speaker models.

Voice Controllability. An additional benefit of the multi-speaker extension is the ability to transfer from the source speaker's voice to a different target speaker's voice. During training of the speech codec, adding a path for information flow from the speaker identity encoder directly to the decoder as shown in Figure 4 and restricting the size of the codebook disentangles speech content from speaker identity; the information bottleneck forces speech codes to reflect speech content [48]. Voice transfer is therefore achieved by swapping the speaker identity embedding with the identity embedding of the new target speaker. Figure 5 shows an example of such a voice transfer. Denoising alone restores the linguistic content, *e.g.* the vowels and fricatives, of the noisy input speech. Denoising with simultaneous voice transfer (by swapping the speaker embedding to another speaker) produces a result in which the same speech content (*e.g.* vowel and fricatives in Figure 5) are maintained but additionally the overtone structure, which determines the sound of one's voice, is adjusted according to the new speaker identity.

7. Conclusion

We presented a novel speech enhancement framework that maps video and noisy audio inputs onto discrete speech codes, from which clean speech can be re-synthesized without bleeding-through of acoustic noise or unnatural distortions. To train and evaluate our model, we introduced a novel audio-visual dataset containing more than 10 hours of unconstrained, natural speech with large-vocabulary and high-quality audio and visual recordings. Experiments show that our approach outperforms existing frameworks both in quantitative evaluation and human perceptual studies. In the same way that personalized photo-realistic codec avatars are pushing 3D face representations beyond the uncanny valley, we show that personalized audio-visual speech codecs enable a similar leap forward in audio-visual speech enhancement for VR telepresence applications.

Limitations. (1) Personalized AV speech codecs require a separate model for each user, which comes at a higher computational cost than speaker-agnostic methods. We propose a first step towards scaling up in Section 6, but a large-vocabulary multi-speaker dataset with high-quality audio is needed to investigate this direction further. (2) When our model fails (*e.g.*, in very noisy settings), the outputs may be realistic but do not faithfully represent the user's speech. In extreme cases, our model can hallucinate plausible mumbling that was not part of the user's original speech.

Ethical Considerations. As with all speech synthesis systems that enable voice conversion, our approach has to be handled responsibly to avoid audio deep-fakes. Audio watermarking [3] is one strategy for protecting against misuse.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement. *arXiv preprint arXiv:1804.04121*, 2018. 1, 2

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. My lips are concealed: Audio-visual speech enhancement through obstructions. *arXiv preprint arXiv:1907.04975*, 2019. 2

[3] Michael Arnold. Audio watermarking: Features, applications and algorithms. In *2000 IEEE International conference on multimedia and expo. ICME2000. Proceedings. Latest advances in the fast changing world of multimedia (cat. no. 00TH8532)*, volume 2, pages 1013–1016. IEEE, 2000. 8

[4] Yutian Chen, Yannis Assael, Brendan Shillingford, David Budden, Scott Reed, Heiga Zen, Quan Wang, Luis C Cobo, Andrew Trask, Ben Laurie, et al. Sample efficient adaptive text-to-speech. *arXiv preprint arXiv:1809.10460*, 2018. 2, 8

[5] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 5

[6] Soo-Whan Chung, Soyeon Choe, Joon Son Chung, and Hong-Goo Kang. Facefilter: Audio-visual speech separation using still images. *arXiv preprint arXiv:2005.07074*, 2020. 2

[7] 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. 2, 4, 5, 8

[8] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. Real time speech enhancement in the waveform domain. In *Interspeech*, 2020. 2, 6

[9] Ngoc QK Duong, Emmanuel Vincent, and Rémi Gribonval. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1830–1840, 2010. 2

[10] Ariel Ephrat, Tavi Halperin, and Shmuel Peleg. Improved speech reconstruction from silent video. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 455–462, 2017. 3

[11] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018. 1, 2

[12] Ariel Ephrat and Shmuel Peleg. Vid2speech: speech reconstruction from silent video. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5095–5099. IEEE, 2017. 3

[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021. 2

[14] John W Fisher III, Trevor Darrell, William T Freeman, and Paul A Viola. Learning joint statistical models for audio-visual fusion and segregation. In *Advances in neural information processing systems*, pages 772–778, 2001. 2

[15] Aviv Gabbay, Asaph Shamir, and Shmuel Peleg. Visual speech enhancement. *arXiv preprint arXiv:1711.08789*, 2017. 2, 6

[16] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10478–10487, 2020. 2

[17] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2019. 6

[18] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. *arXiv preprint arXiv:2101.03149*, 2021. 1, 2, 6

[19] Cristina Gârbacea, Aäron van den Oord, Yazhe Li, Felicia SC Lim, Alejandro Luebs, Oriol Vinyals, and Thomas C Walters. Low bit-rate speech coding with vq-vae and a wavenet decoder. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 735–739. IEEE, 2019. 3

[20] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017. 5

[21] Naomi Harte and Eoin Gillen. Tcd-timit: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia*, 17(5):603–615, 2015. 2, 4, 5

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[23] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. Deep learning for monaural speech separation. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1562–1566. IEEE, 2014. 2

[24] Lisa I Iezzoni, Bonnie L O'Day, Mary Killeen, and Heather Harker. Communicating about health care: observations from persons who are deaf or hard of hearing. *Annals of internal medicine*, 140(5):356–362, 2004. 1

[25] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 3

[26] Srihari Kankanahalli. End-to-end optimized speech coding with deep neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2521–2525. IEEE, 2018. 3

[27] Keisuke Kinoshita, Marc Delcroix, Sharon Gannot, Emanuël AP Habets, Reinhold Haeb-Umbach, Walter Kellermann, Volker Leutnant, Roland Maas, Tomohiro Nakatani, Bhiksha Raj, et al. A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP Journal on Advances in Signal Processing*, 2016(1):1–19, 2016. 6

[28] W Bastiaan Kleijn, Andrew Storus, Michael Chinen, Tom Denton, Felicia SC Lim, Alejandro Luebs, Jan Skoglund,

and Hengchin Yeh. Generative speech coding with predictive variance regularization. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6478–6482. IEEE, 2021. 3

[29] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *arXiv preprint arXiv:2010.05646*, 2020. 3, 4, 5

[30] Anurag Kumar and Dinei Florencio. Speech enhancement in multiple-noise conditions using deep neural networks. *arXiv preprint arXiv:1605.02427*, 2016. 2

[31] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. *arXiv preprint arXiv:1910.06711*, 2019. 3

[32] Yaman Kumar, Rohit Jain, Khwaja Mohd Salik, Rajiv Ratn Shah, Yifang Yin, and Roger Zimmermann. Lipper: Synthesizing thy speech using multi-view lipreading. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2588–2595, 2019. 3

[33] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. Neural speech synthesis with transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6706–6713, 2019. 3

[34] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Transactions on Graphics (TOG)*, 37(4):1–13, 2018. 1, 2

[35] Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021. 1

[36] Eliya Nachmani, Yossi Adi, and Lior Wolf. Voice separation with an unknown number of multiple speakers. In *International Conference on Machine Learning*, pages 7164–7175. PMLR, 2020. 2

[37] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016. 3

[38] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017. 2, 3

[39] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 2

[40] Sanjeel Parekh, Slim Essid, Alexey Ozerov, Ngoc QK Duong, Patrick Pérez, and Gaël Richard. Motion informed audio source separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6–10. IEEE, 2017. 2

[41] Sarah Partan and Peter Marler. Communication goes multimodal. *Science*, 283(5406):1272–1273, 1999. 1, 2

[42] Santiago Pascual, Antonio Bonafonte, and Joan Serra. Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*, 2017. 2

[43] 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 preprint arXiv:1710.07654*, 2017. 3

[44] Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. Speech resynthesis from discrete disentangled self-supervised representations. *arXiv preprint arXiv:2104.00355*, 2021. 3

[45] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. Learning individual speaking styles for accurate lip to speech synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13796–13805, 2020. 2, 3, 5

[46] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621. IEEE, 2019. 3

[47] Jie Pu, Yannis Panagakis, Stavros Petridis, and Maja Pantic. Audio-visual object localization and separation using low-rank and sparsity. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2901–2905. IEEE, 2017. 2

[48] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*, pages 5210–5219. PMLR, 2019. 7, 8

[49] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando de la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 3

[50] Sam T Roweis. One microphone source separation. In *NIPS*, volume 13, 2000. 2

[51] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 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. IEEE, 2018. 3, 5

[52] Paris Smaragdis and Michael Casey. Audio/visual independent components. In *Proc. ICA*, pages 709–714, 2003. 2

[53] Paris Smaragdis, Bhiksha Raj, and Madhusudana Shashanka. Supervised and semi-supervised separation of sounds from single-channel mixtures. In *International Conference on Independent Component Analysis and Signal Separation*, pages 414–421. Springer, 2007. 2

[54] Martin Spiertz and Volker Gnann. Source-filter based clustering for monaural blind source separation. In *Proceedings of the 12th International Conference on Digital Audio Effects*, volume 3, 2009. 2

[55] Daniel Stoller, Sebastian Ewert, and Simon Dixon. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *arXiv preprint arXiv:1806.03185*, 2018. 2

[56] Jiaqi Su, Zeyu Jin, and Adam Finkelstein. Hifi-gan: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks. *arXiv preprint arXiv:2006.05694*, 2020. 2

[57] William H Sumby and Irwin Pollack. Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, 26(2):212–215, 1954. 1, 2

[58] James Traer and Josh H McDermott. Statistics of natural reverberation enable perceptual separation of sound and space. *Proceedings of the National Academy of Sciences*, 113(48):E7856–E7865, 2016. 5

[59] A Vasuki and PT Vanathi. A review of vector quantization techniques. *IEEE Potentials*, 25(4):39–47, 2006. 3

[60] Yuxuan Wang, Arun Narayanan, and DeLiang Wang. On training targets for supervised speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 22(12):1849–1858, 2014. 2

[61] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017. 3

[62] Shih-En Wei, Jason Saragih, Tomas Simon, Adam W Harley, Stephen Lombardi, Michal Perdoch, Alexander Hypes, Dawei Wang, Hernan Badino, and Yaser Sheikh. Vr facial animation via multiview image translation. *ACM Transactions on Graphics (TOG)*, 38(4):1–16, 2019. 1, 2

[63] Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R Hershey, and Björn Schuller. Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr. In *International conference on latent variable analysis and signal separation*, pages 91–99. Springer, 2015. 2

[64] Xudong Xu, Bo Dai, and Dahua Lin. Recursive visual sound separation using minus-plus net. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 882–891, 2019. 2

[65] Ozgur Yilmaz and Scott Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on signal processing*, 52(7):1830–1847, 2004. 2

[66] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 241–245. IEEE, 2017. 2

[67] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *arXiv preprint arXiv:2107.03312*, 2021. 3, 7

[68] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 192–201, 2017. 5

[69] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1735–1744, 2019. 2

[70] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018. 2