ReVISE: Self-Supervised Speech Resynthesis with Visual Input for Universal and Generalized Speech Regeneration

Wei-Ning Hsu[1], Tal Remez[1], Bowen Shi[1,3], Jacob Donley[2], Yossi Adi[1,4]

[1]FAIR, Meta AI Research   [2]Meta Reality Labs Research
[3]Toyota Technological Institute at Chicago   [4]The Hebrew University of Jerusalem

{wnhsu,talr,bshi,jdonley,adiyoss}@meta.com

Abstract

Prior works on improving speech quality with visual input typically study each type of auditory distortion separately (e.g., separation, inpainting, video-to-speech) and present tailored algorithms. This paper proposes to unify these subjects and study Generalized Speech Regeneration, where the goal is not to reconstruct the exact reference clean signal, but to focus on improving certain aspects of speech while not necessarily preserving the rest such as voice. In particular, this paper concerns intelligibility, quality, and video synchronization. We cast the problem as audio-visual speech resynthesis, which is composed of two steps: pseudo audio-visual speech recognition (P-AVSR) and pseudo text-to-speech synthesis (P-TTS). P-AVSR and P-TTS are connected by discrete units derived from a self-supervised speech model. Moreover, we utilize self-supervised audio-visual speech model to initialize P-AVSR. The proposed model is coined ReVISE. ReVISE is the first high-quality model for in-the-wild video-to-speech synthesis and achieves superior performance on all LRS3 audio-visual regeneration tasks with a single model. To demonstrates its applicability in the real world, ReVISE is also evaluated on EasyCom, an audio-visual benchmark collected under challenging acoustic conditions with only 1.6 hours of training data. Similarly, ReVISE greatly suppresses noise and improves quality. Project page: https://wnhsu.github.io/ReVISE/.

1. Introduction

Unlike anechoic studio recordings, speech in-the-wild is rarely clean: outdoor recordings are corrupted with all sorts of natural and non-natural sounds like wind and traffic noise [6]. Speech recorded indoor often contains reverberation, mechanical noise, and overlapping speech from non-target speakers [54]. On top of those, recording devices and network may also introduce other types of dis-
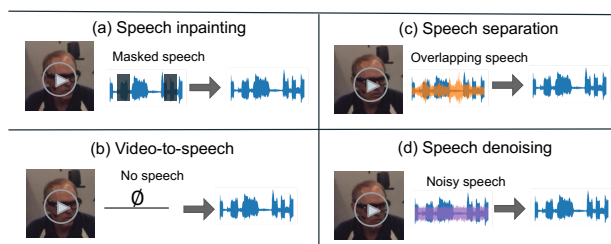

Figure 1. Illustration of AVSE with various distortion.

tortion, such as amplitude clipping, band-pass filtering, and package loss [1]. Distortion makes it hard for both human and machines to comprehend speech [10, 31]. Improving the quality and the intelligibility of corrupted speech is essential for assistive listening and robust speech processing. Generating clean speech signal based on its corrupted version is herein referred to as speech enhancement.

In speech enhancement, one line of research uses visual speech to provide auxiliary information [14, 17, 19, 57], which is known as audio-visual speech enhancement. Audio-visual speech (e.g., talking-head videos) can be seen as a multimodal view of the speech. Since visual modality is immune to acoustic noise, combining both views enables more robust estimation of shared generating factors such as textual content. Meanwhile, despite sharing the same goal of recovering corrupted speech, prior work often treats enhancement from each type of distortion as a separate problem: speech denoising and dereverberation addresses additive and convolutive non-speech noises [17], speech separation focuses on speech noises that exhibit similar characteristics to the target speech [19], speech inpainting aims to recover dropped audio frames [38], and video-to-speech synthesis is the extreme case of inpainting where all the frames are dropped [15, 36]. As a result, algorithms designed for one type of distortion may not be effective for another.

In this paper, we advocate a more holistic approach to audio-visual speech enhancement, where an algorithm

should be evaluated on all types of distortion, and a single model should also be effective on all types of corrupted data, shifting from building distortion-specific models to a universal model. Following [48], we coin the concept **universal speech enhancement**. In turn, we also argue that exact reconstruction of the reference clean speech is not an appropriate objective especially when the level of distortion is high. To address the issue, we propose to relax the objective and solve the **generalized speech regeneration** (GSR) problem: instead of focusing on exact reconstruction and measuring metrics like signal-to-noise ratios (SNRs), the goal of GSR is to enhance a predefined set of attributes, such as content intelligibility that can be measured by word error rates (WERs). In contrast, the model does not need to preserve other attributes.

This paper focuses on recovering intelligibility, synchronicity and quality. The task of improving those could be broken down into two steps: predicting the frame-level content and synthesizing high quality audio from it. Inspired by the resemblance to audio-visual speech recognition and speech synthesis, we propose ReVISE, short for **Re**synthesis with **V**isual **I**nput for **S**peech R**E**generation. ReVISE is composed of a pseudo audio-visual speech recognition model (P-AVSR) and a pseudo text-to-speech synthesis model (P-TTS); instead of using text as the output/input of the two models, self-supervised speech units that encode speech content [23, 44] are adopted to bridge them, making the system free of text supervision. Furthermore, observing the gain on speech recognition brought by self-supervised learning, we also initialize the P-AVSR with a self-supervised audio-visual speech model, AV-HuBERT [49], which significantly improves the performance, especially on low-resource setups.

To demonstrate the universality and compare with the literature, we construct four types of corrupted speech using Lip-reading Sentences 3 (LRS3) [2] and AudioSet [20], including audio-visual denoising, separation, inpainting, and video-to-speech. Results suggest that ReVISE is the first model capable of high-quality in-the-wild video-to-speech synthesis, while prior models fail to produce intelligible content [21] or generate low-quality audio for in-the-wild videos [36]. Compared to a strong masking-based method [19] on denoising and separation, ReVISE achieves comparable performance on mid-/high-SNR conditions (0-20dB), and are significantly stronger on lower SNR conditions, reducing WERs by up to 37.5% absolute and improving MOS by up to 1.09. Finally, we also show that a single ReVISE model can tackle all four types of distortion with similar performance to distortion-specific models. To further show the data efficiency and effectiveness of ReVISE on real data, we evaluate it on EasyCom [13], an audio-visual speech dataset addressing the cocktail party problem which contains clean close-talking recordings and noisy distant recordings with background noise, loud interfering speech, and room reverberation. Results show that ReVISE still shines in this challenging setup, reducing WER by up to 32% while other methods fail.

2. Background

2.1. Audio-visual speech enhancement tasks

In general, audio-visual speech enhancement is the task of improving the quality of corrupted speech $\tilde{x}_a$ given its corresponding talking head video x_v. During training, tuples of $(\tilde{x}_a, x_v, x_a)$ are provided where x_a is the reference clean speech. The tasks are further divided depending on the type of distortion applied to $\tilde{x}_a$: speech denoising and dereverberation considers $\tilde{x}_a[t] = x_a[t] * h[t] + n[t]$ with impulse response h and additive noise n (t indexes discrete time steps). Source separation considers $\tilde{x}_a[t] = x_a[t] + x'_a[t]$ where x'_a is an interfering speech sampled from the same marginal distribution as x_a. For these two subtasks, masking-based methods (in magnitude spectrogram domain [16, 41], complex spectrogram domain [14, 19, 55], or on learned basis [33, 34]) are widely adopted where the enhancement model f predicts a mask $\hat{M} = f(\tilde{x}_a, x_v)$. It is feasible because the target speech can be written as $X = M \cdot \tilde{X}$ where X and $\tilde{X}$ are the clean and noisy speech in the target domain and M is the ideal ratio mask. On the other hand, speech inpainting and video-to-speech synthesis considers $\tilde{x}_a[t] = m[t] \cdot x_a[t]$ where $m[t] \in \{0, 1\}$ for inpainting and $m[t] = 0$ for the latter. Masking-based methods are not feasible for these tasks. Hence, previous studies directly predict spectrogram [36, 38] or waveform [7, 37, 39] as $\hat{x}_a = f(\tilde{x}_a, x_v)$ and optimize the model with regression and adversarial losses. Nevertheless, previous models have not been able to generate realistic samples with in-the-wild datasets like LRS3, because these models are deterministic while the underlying mapping is highly stochastic, especially when there is a large portion of frames dropped.

To accommodate general distortions, universal enhancement models have to adopt generation-based method [32, 43, 45, 48, 50], otherwise they cannot handle distortions like package loss (i.e., inpainting). For the same reason, our proposed model is also generation-based. However, it differs from previous works in two main aspects. First, prior universal enhancement models are audio-based, which do not leverage auxiliary input like visual speech. Hence, source separation and enhancement from silence (i.e., cross-modal generation) are not attainable and hence have not been included. Second, regardless of the training objective (regression [40], adversarial [43, 45, 50], or score-matching [48]), previous universal models predict reference clean speech directly. We consider a different paradigm based on self-supervised speech resynthesis [44]: our model predicts a self-supervised representation $z = k(x_a)$ of the reference

clean speech. A separate model is used to convert z back to audio, which is described in the next section.

2.2. Self-supervised speech resynthesis

Previous studies show that discretized Self Supervised Learned (SSL) speech features from the Hidden Unit BERT (HuBERT) model [23] encode mostly phonetic information and less about speaker and noise characteristics. HuBERT is pre-trained with a masked cluster prediction objective, where spans of input waveform are randomly masked and the model is asked to predict cluster assignment for those frames. Unlike VQ-VAE [52], HuBERT does not learn to encode all the factors; instead, it encodes information pertinent to inferring the unseen content, such as phonetic information. These properties make HuBERT units great candidates for replacing speech for text-free speech generative models, because they encode less nuisance variation and a model can predicts these units more easily compared to predicting speech directly [29, 30]. Such needs motivates the development of speech resynthesis models that convert units back to audio, because these generative models still have to produce speech as their final output.

[44] proposed the first end-to-end model directly generating waveform from SSL units by adapting HiFi-GAN [27] which was originally designed as a vocoder converting spectrogram to waveform. The adpated HiFi-GAN takes SSL units, pitch, and speaker embedding as input, the latter two of which are added to capture variation of training data not encoded in the SSL units. Empirically, [30] demonstrates that speaker embedding and F0 can be removed if the model is trained on a high-quality single-speaker dataset.

3. Method

3.1. Generalized Speech Regeneration

Without loss of generality, we assume that original speech x_a and its auxiliary view x_v are generated with a bijective mapping $g_{m,Y}$ as $x_m = g_{m,Y}(\{y_i\}_i), m \in \{a, v\}$ given a set of factors $Y = \{y_i\}_i$ (e.g., pitch, speed, textual content) sampled from p_Y. Corrupted speech $\tilde{x}_a$ is generated with a corruption function g_d as $\tilde{x}_a = g_d(x_a, d)$ given x_a and distortion parameter d sampled from p_d. Audio-visual speech enhancement can be seen as estimating $p(Y \mid g_d(x_a, d), x_v)$. When the level of distortion is low, there are fewer Y that can result in the same observed $\tilde{x}_a$ after corruption. Formulating it as a regression problem and measuring the performance of reconstructing *all factors* with metrics like SNR is reasonable. However, when the level of distortion is high, there exist multiple sets of Y rendering the same noisy speech $\tilde{x}_a$. Take video-to-speech as an example: while the content and the timing of each word can be more accurately inferred, the exact pitch is hard to infer from the video. In other words, $p(y_i|\tilde{x}_a, x_v)$ is more

deterministic for some y_i but not for the others.

We argue that for these scenarios reconstructing the exact clean reference signal x_a is an ill-posed problem. Instead, one should consider the generalized speech regeneration problem, where the goal is to generate an enhanced signal $\hat{x}_a = f(\tilde{x}_a, x_v)$ that is on the manifold of clean speech $g_{a,Y}(\cdot)$ (high quality) and preserves the factors of interest (partial faithfulness). Let $g_{a,Y,i}^{-1}$ be the inverse mapping from speech to factor y_i, and $\{y_i\}_{i \in I^+}$ be the list of factors of interest (which should be a subset of factors that can be inferred from corrupted speech). Faithfulness is measured by the discrepancy between $g_{a,Y,i}^{-1}(\hat{x}_a)$ and $g_{a,Y,i}^{-1}(x_a)$ for $i \in I^+$. In case of y_i being the textual content, such discrepancy can be measured by the word error rate produced by an off-the-shelf speech recognizer for example. Quality can be measured by metrics commonly used for text-to-speech synthesis such as mean opinion scores, where human raters rate how realistic and clean the audio sounds regardless of the textual content. In this paper, we consider $\{y_i\}_{i \in I^+}$ to be textual content and timing (synchronization with the video) and do not address other factors such as voice, which is not of primary importance for communication.

3.2. ReVISE

We propose to solve the problem in two stages: (i) predicting $\{y_i\}_{i \in I^+}$ of the clean reference x_a given $(\tilde{x}_a, x_v)$, and (ii) resynthesizing speech conditioned on $\{y_i\}_{i \in I^+}$. The two steps are handled by a pseudo audio-visual speech recognition module (P-AVSR) and a pseudo text-to-speech synthesis module (P-TTS), respectively, as shown in Fig. 2. The discrete SSL units derived from HuBERT [23] as described in Sec. 2.2 are used to represent $\{y_i\}_{i \in I^+}$. Using SSL units have two main benefits: not restricted by text availability and better encoding of non-verbal information.

The P-AVSR module performs a task very similar to speech recognition. Inspired by the recent success of self-supervised pre-training for audio [4, 23] and audio-visual speech recognition [49], we initialize P-AVSR with AV-HuBERT [49], an SSL model achieving state-of-the-art performance on audio, visual, and audio-visual speech recognition. It is composed of a video encoder and an audio encoder as the modality-specific front-end, followed by a stack of transformer layers as the shared back-end taking concatenated audio and video features as input (Fig. 2 right). The model encodes waveform into features at a frame rate of 25Hz, while the SSL unit from HuBERT are encoded at 50Hz. To match the rate, a lightweight transposed convolution layer followed by a softmax layer is added as the prediction head, which is randomly initialized. The P-AVSR module is trained with a cross-entropy loss $L = \sum_t \sum_{j=1}^{C} z_t^j \log f_t^j(\tilde{x}_a, x_v)$, where $\mathbf{z}_t$ denotes the one-hot discrete unit label of the t-th frame, $f_t(\tilde{x}_a, x_v)$ is the predicted distribution over discrete units from the en-
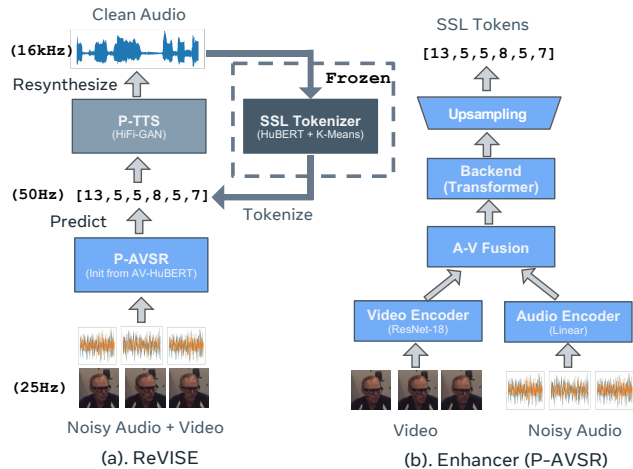
Figure 2. Model diagram.

hancer for the t-th frame and C unit vocabulary size.

We use the unit HiFi-GAN described in Sec. 2.2 for converting predicted z into speech, which aims to faithfully produce the content and timing encoded in units. In fact, P-TTS can be trained with a different dataset from P-AVSR and this dataset determines the audio quality and the voice of ReVISE. With privacy consideration in mind, we use a high quality single speaker dataset with the speaker's consent to train this model. Consequently, ReVISE does not preserve voice and other fine-grained details and could generate speech of even higher quality than the original speech. Note that the P-TTS can be easily extended following [44].

4. Experimental Setup

4.1. Datasets

We consider two datasets for our experiments. The first one is Lip Reading Sentences 3 (LRS3) [2], which is a clean dataset based on TED talk videos. It contains 433 hours of audio-visual speech data and their corresponding text transcripts. Following prior work on speech enhancement, speech data are artificially corrupted to simulate enhancement tasks. Four tasks are created: (i) **Speech denoising:** in which we artificially mix a clean speech signal with a noise sample using randomly sampled Signal-to-Noise Ratio (SNR) in the range of [-20,20]. Noisy samples were selected from AudioSet [20]; (ii) **Speech source separation:** in which we follow a similar procedure to the speech denoising setup but instead of mixing the speech signal with a noisy sample, we mix it with another speech signal produced by a different speaker; (iii) **Speech inpaiting:** under this setup we randomly dropped (i.e., zero out) spans of $s \in \{20, 30, 40\}$ frames, with probability $p \in \{0.3, 0.4, 0.5\}$. Each frame corresponds to 20ms, hence in case $s = 20$ we masked 400ms of audio. This process

can be seen as a packet loss simulation; and (iv) **Video-to-speech synthesis:** in this setup we use the LRS3 dataset without any augmentations, where our goal is to synthesize the audio given a silent video only.

The second data is Easy Communications (Easy-Com) dataset [13], which includes real-life noisy speech and is designed to study the cocktail party problem in conversational augmented reality. We preprocess the data using official time-aligned transcripts and crop image ROIs based on lip bounding boxes, resulting in 1.64/0.59/0.35 hours for train/valid/test splits. The dataset is challenging due to motion blur, barrel distortion, interfering speech, noise, and reverberation[1]. Close-talking microphones are used for clean reference speech, and beamformed audio is used as model input, following [5,13,18]. ReVISE is trained using merged audio from all six channels, including close-talking microphone audio. More details can be found in the Appendix.

4.2. Model

SSL speech tokenizer We use a BASE HuBERT model which is composed of a convolutional encoder and 12 Transformer layers. Each layer has embedding dimension of 768, feed-forward layer dimension of 3072, and 12 self-attention heads. The model is pre-trained for three iterations on 32 GPUs with 400K updates per iteration, using clusters of MFCC/6-th layer feature from iteration 1/9-th layer feature from iteration 2 as the target, with a codebook size of 100/500/1000, respectively, following the recipe of [30]. A total of 221K hours of unlabeled speech data combining multilingual Librispeech [46], Common Voice [3], and VoxPopuli [53] in eight languages (En, Es, Fr, De, Nl, It, Pl, Pt) are used for pre-training. The final SSL units z used in ReVISE are generated by clustering the third iteration feature at the last layer with a codebook size of 2000 with K-means.

P-TTS We use the unit HiFi-GAN described in Sec. 2.2. The model is trained for 400K updates on 8 GPUs on the LJSpeech [25] dataset resampled to 16kHz to match the sample rate of other datasets. Similar to [30], F0 and speaker embedding are not used.

P-AVSR We use the LARGE AV-HuBERT model by default from [49]. It takes video and/or audio as input, where video is head-crop image frames at 25fps, and audio is 23-dimensional Mel filterbank sequence (FBank) computed with a 10ms frame shift and stacked every four frames to match the video. The video encoder is a ResNet-18 [35] model, while the audio-encoder is simply a linear projection. The features are concatenated frame-by-frame and passed to a stack of 24 Transformer layers, which has embedding and feed-forward layer dimension of 1024 and

---

[1] We refer readers to the supplementary material to listen to the samples.

4096 with 16 attention heads. For all but EasyCom experiments we use the publicly available checkpoints from [49] that were pre-trained on LRS3 and filtered VoxCeleb2 for initialization. For Easycom, a u-HuBERT model [24] pre-trained additionally with Librilight [26] is used for initialization, which improves upon using [49]. The up-sampling module is one transposed convolution layer with stride of 2, kernel size of 4, and 768 channels, followed by GeLU activation [22]. The P-AVSR models are fine-tuned on 8 GPUs for less than 45K updates. Detailed fine-tuning configurations for each task are presented in the appendix.

4.3. Evaluation

We evaluate the proposed model and baselines on the following axes: *content*, *synchronization*, *quality*, and *low-level detail* reconstruction. As discussed in Sec. 3.1, we focus on the first three axes, and the results for the last metric are included in the appendix for completeness.

For content, we use the WER computed with a speech recognition model to measure the intelligibility quantitatively similar to [36, 37]. The public model from [56] is used, which reports a WER of 5.6% on LRS3 test split and 35.7% on the EasyCom close-talking validation set. For synchronization, following [21] we use SyncNet [9] metrics, the predicted temporal distance between audio and video (LSE-D) and the prediction's confidence (LSE-C) are averaged over the entire test set. For quality, we follow the tradition of text-to-speech synthesis evaluation and conduct subjective mean opinion score (MOS) studies with a scale from 1 to 5 and a 0.5 increment. We evaluate 50 randomly sampled files from the test set, where each sample was evaluated by at least 15 raters using the CrowdMOS package [47]. Finally, for reconstruction of low-level details as typically done in speech denoising studies, we include ESTOI [51] and Mel cepstral distortion (MCD) [28]. While MCD provides an estimate of the similarity of two signals by measuring the mel cepstra difference, ESTOI put more emphasis on speech properties.

5. Results

Quantitative results are presented in the section. We strongly recommend readers to watch the samples provided in the supplementary material for better understanding of the sample quality compared to the baselines.

5.1. Ground truth and resynthesis performance

For reference, we first evaluate reference clean speech (Tgt. audio) and resynthesized clean speech (Tgt. audio resynthesis) with the proposed metrics and report the results in Tab. 1. The gap between the two is an approximation of how much information is lost when tokenizing speech into SSL units. It shows that intelligibility, synchronization, and quality are all slightly degraded. As ReVISE

Method	Cont WER↓	Sync LSE-C↑	Sync LSE-D↓	Qual MOS↑	Low-Level ESTOI↑	Low-Level MCD↓
Tgt. audio	5.6%	5.20	6.33	4.38±0.02	-	-
Tgt. audio resynthesis	10.6%	5.02	6.40	4.16±0.02	41.59	10.01
Silent video to speech						
SVTS-L, LRS3	81.5%	5.07	6.46	2.12±0.04	26.87	8.13
SVTS-L, LRS3+VoxEn	67.4%	5.53	6.00	2.23±0.03	30.98	7.59
ReVISE (Ours)	33.9%	5.03	6.44	4.13±0.02	28.51	10.73

Table 1. Full video-to-speech synthesis results on LRS3 test set.

is trained to predict the tokens inferred from clean speech, if the ReVISE model have 100% prediction accuracy then the performance will be identical to the "Tgt. audio resynthesis" row here. Hence, the performance on the resynthesized speech roughly upper-bounds that of the ReVISE model.

5.2. Video-to-speech synthesis

In Tab. 1, we compare ReVISE with SVTS [36],[2] which is the state-of-the-art model on the video-to-speech synthesis task. SVTS is composed of a video-to-spectrogram predictor and a separately trained neural vocoder. Its video-to-spectrogram predictor is trained with a regression loss, which takes as input a lip video and a speaker embedding extracted from a pre-trained fixed speaker encoder. Speaker embedding is needed because voice cannot be accurately inferred from the video. The model demonstrate strong results on two constrained datasets (LRW [8] and GRID [11]) which contain isolated words and fixed-length sentences with limited word choices, respectively.

As shown in Tab. 1, because the audio quality generated by SVTS is mediocre (2.12 and 2.23 MOS),[3] these samples are not very intelligible by machines, reflected by the high WER (81.5% and 67.4% when trained on LRS3 and LRS3+VoxCeleb2). In contrast, our proposed model generates much higher quality audio (4.13 MOS) and yields significantly lower WER (33.9%). Through manual inspection, we found a confidence score above 5 and distance below 6 implies good synchronization on LRS3. All methods generate audio fairly synchronized to the video.

5.3. Audio-visual speech inpainting

We next evaluate ReVISE on the audio-visual speech inpainting task with three test splits, where 30%, 40%, and 50% of the frames are dropped with dropped spans longer than 400ms. While there have been several works studying audio-based speech inpainting or audio-visual inpainting for music [58], [38] is the only work we are aware of studying the audio-visual setup. However, their model was only evaluated on the constrained GRID dataset, showing a phone error rate of 13.7%. SVTS achieves a WER of 17.9%, hinting that [38] is likely worse than SVTS even

---

[2]We obtained test set samples of SVTS from the authors for evaluation.
[3]https://sites.google.com/view/scalable-vts

Method	Inp Mod	Test	Cont WER↓	Sync LSE-C↑	LSE-D↓	Low-Level ESTOI↑	MCD↓
Audio-visual speech inpainting							
Inp. audio	A	30%	44.1%	4.37	7.05	60.75	36.54
		40%	53.5%	4.01	7.36	51.69	44.71
		50%	60.3%	3.78	7.60	43.16	52.82
Demucs	A	30%	43.9%	4.37	7.04	59.40	35.95
		40%	53.5%	4.04	7.34	50.55	43.96
		50%	60.0%	3.82	7.57	42.20	52.09
Resynthesis	A	30%	51.2%	4.20	7.14	25.27	11.57
		40%	60.1%	3.94	7.38	21.55	11.91
		50%	66.3%	3.68	7.63	17.97	12.09
ReVISE (Ours)	AV	30%	14.9%	5.01	6.42	38.15	10.23
		40%	18.1%	4.99	6.44	37.05	10.31
		50%	19.2%	5.04	6.41	35.99	10.33

Table 2. Full speech inpainting results on LRS3 test set with 30%/40%/50% of frames dropped.

Method	Mod	Test	Cont WER↓	Sync LSE-C↑	LSE-D↓	Qual MOS↑	Low-Level ESTOI↑	MCD↓
Tgt. audio	-	-	5.6%	5.20	6.33	4.36±0.018	-	-
Audio-visual speech denoising								
Inp. audio	A	lvl 1	7.8%	4.66	6.77	2.95±0.03	87.21	5.07
		lvl 2	20.0%	3.80	7.52	2.82±0.03	69.93	8.27
		lvl 3	63.9%	2.41	8.68	2.79±0.02	45.91	11.27
		lvl 4	87.7%	1.54	9.39	2.71±0.04	25.55	13.47
Resynthesis	A	lvl 1	13.3%	4.98	6.46	4.33±0.02	41.18	10.02
		lvl 2	25.9%	4.78	6.64	4.37±0.02	39.16	10.19
		lvl 3	69.4%	3.64	7.61	3.63±0.02	25.71	11.27
		lvl 4	90.8%	2.11	8.85	3.34±0.03	9.20	12.56
Demucs [12]	A	lvl 1	6.9%	5.15	6.38	3.58±0.04	92.59	3.73
		lvl 2	15.1%	4.95	6.57	3.63±0.03	85.61	4.81
		lvl 3	48.0%	4.14	7.28	3.36±0.02	65.94	6.81
		lvl 4	81.3%	2.43	8.75	2.89±0.04	31.52	10.25
VisualVoice [19]	AV	lvl 1	6.6%	5.16	6.35	4.11±0.03	92.20	3.89
		lvl 2	8.8%	5.19	6.35	3.94±0.02	86.92	5.07
		lvl 3	23.4%	5.07	6.45	3.80±0.02	74.16	6.45
		lvl 4	58.0%	4.53	6.86	3.25±0.03	50.48	8.01
ReVISE (Ours)	AV	lvl 1	9.4%	5.03	6.42	4.31±0.02	41.22	10.04
		lvl 2	9.7%	5.04	6.41	4.36±0.02	40.93	10.06
		lvl 3	11.7%	5.04	6.41	4.41±0.02	39.79	10.14
		lvl 4	20.5%	5.07	6.40	4.34±0.02	35.80	10.36

Table 3. Full speech denoising results on LRS3 test set. SNR ranges for lvl 1/2/3/4 are [10,20]/[0,10]/[-10,0]/[-20,-10] dB.

when having partial audio as input. Accordingly, we decide not to compare with [38] but compare with a publicly available strong generation-based audio enhancement model, Demucs [12], which was trained on DNS-Challenge, VCTK, WHAM, and WHAMR, containing 700hr of clean speech and 6,000hr of noise in total. In addition, we consider resynthesis as another baseline, which encodes and decodes the corrupted speech with the same SSL tokenizer and P-TTS model.

"Inp. audio" row presents the results evaluated on the corrupted audio in Tab. 2. We see that intelligibility degrades (WER increases from 5.6% to 60.3%) as the percentage of dropped frames increases, and so does synchronization (5.20 to 3.78 for confidence and 6.33 to 7.60 for distance). The two audio-based baselines are not effective, showing similar or even worse results compared to the corrupted samples before enhancement. The little gain from Demucs over the inputting-audio baseline shows that existing enhancement approaches fails to generalize to inpainting. In contrast, our proposed model is effective on improving both intelligibility and synchronization. The WER drops from 60.3% to 19.2% after regeneration when half of the audio is missing in the corrupted audio. Comparing enhancing from no audio at all (video-to-speech synthesis) with enhancing from observing half of the audio (inpainting, 50%), we observe a 14.7% WER reduction (33.9% → 19.2%), showing that the model use the additional audio information effectively to improve content reconstruction. As for synchronization, ReVISE restore a similar level regardless of the percentage dropped.

5.4. Audio-visual speech denoising

Tab. 3 presents results of audio-visual speech denoising on four test splits: lvl 1-4, which contain noisy samples of SNR in [10, 20], [0, 10], [-10, 0], and [-20, -10]. In addition to the two audio-based baselines from the previous section, the proposed model is also compared with VisualVoice [19] in this and the next section, which is a state-of-the-art model for audio-visual source separation and denoising. VisualVoice predicts a complex IRM given noisy audio, lip video, and speaker face image as input and the model is optimized with a combination of mask prediction loss, cross-modal matching loss (image and enhanced audio embeddings from the same speaker should be close) and consistency loss (embeddings of enhanced audio for the same speaker should be close). To have a fair comparison, we re-train VisualVoice models with the same dataset as ReVISE using a similar setup provided by the authors. We use the same public checkpoint for Demucs.

In terms of intelligibility, we observe that the performance of resynthesizing corrupted audio is still bad. Demucs performs well for high SNR setup (lvl 1) but lags behind audio-visual models for the rest setups. VisualVoice achieves the best results on the [0, 20] dB SNR range, but ReVISE is much more robust to even a higher level of noise, reporting a WER of 20.5% at lvl 4. Synchronization scores follow a similar trend to WERs in terms of ranking across models and test splits. Quality of the proposed model is significantly higher (4.31-4.41 MOS) than all baselines, and is on par with the reference clean speech (4.36 MOS).

5.5. Audio-visual source separation

Following Sec. 5.4, four noisy test sets are created with different levels of SNR. Comparing the speech source separation results in Tab. 4 and denoising results in Tab. 3, we observe that audio model (the public Demucs) performs worse with speech noise than with non-speech noise at the same SNR level. In contrast, audio-visual models report similar performance on the two tasks at high SNR, and better performance on separation at low SNR. These results suggest that it is easier to remove speech noise if the model have auxiliary information that it can use to identify the tar-

Method	Mod	Test	Cont WER↓	Sync LSE-C↑	Sync LSE-D↓	Low-Level ESTOI↑	Low-Level MCD↓
Audio-visual speech separation							
Inp. audio	A	lvl 1	15.7%	4.73	6.74	85.56	3.29
		lvl 2	65.7%	3.85	7.48	65.52	5.95
		lvl 3	102.3%	2.93	8.27	40.97	8.67
		lvl 4	105.5%	2.46	8.66	22.33	10.51
Demucs [12]	A	lvl 1	11.7	5.06	6.46	89.68	3.76
		lvl 2	55.7	4.10	7.25	66.80	6.27
		lvl 3	101.2	2.77	8.42	26.32	10.22
		lvl 4	106.2	2.15	8.95	10.15	12.14
VisualVoice [19]	AV	lvl 1	6.2%	5.13	6.37	94.69	2.43
		lvl 2	7.8%	5.12	6.39	90.29	3.36
		lvl 3	15.3%	4.91	6.57	78.82	4.80
		lvl 4	51.8%	4.20	7.17	53.84	7.02
ReVISE (Ours)	AV	lvl 1	9.9%	5.05	6.41	41.34	10.05
		lvl 2	10.2%	5.05	6.41	40.90	10.07
		lvl 3	11.4%	5.05	6.41	40.07	10.11
		lvl 4	15.7%	5.04	6.41	37.97	10.21

Table 4. Full source separation results on LRS3 test set. SNR ranges for lvl 1/2/3/4 are [10,20]/[0,10]/[-10,0]/[-20,-10] dB.

get speech. Comparing VisualVoice and ReVISE which are trained on the same data, it reveals a similar trend where ReVISE is still substantially better at low SNR.

5.6. Universal audio-visual speech regeneration

In this section, we build a single ReVISE model trained on all four types of distortion and compare it with corruption-specific models presented in earlier sections (Tab. 5) in terms of WER. Surprisingly, the universal model beats or matches the distortion-specific model on almost all tasks. The only exception is video-to-speech synthesis, on which the universal model is 0.5% WER worse.

5.7. AV regeneration on real data — EasyCom

We next study how effective ReVISE on EasyCom, a real noisy dataset that is much more challenging than LRS3 and has much fewer hours of training data. Tab. 6 compares ReVISE with several baseline methods enhancing single channel speech (ch2) or beamformed speech (bf) from multi-channel sources. Note that we attempted to re-train VisualVoice on EasyCom but were not successful on getting any meaningful results, likely due to the lack of data as well as the difficulty. In addition, we use a P-AVSR model that takes lip crops instead of head crops as input for this dataset, which provides slightly better performance. Ablation studies are included in the Appendix.

ReVISE substantially enhances the intelligibility of noisy speech in both setups (WER: 87.5% → 55.0% on ch2 and 71.5% → 52.1% on bf), while Demucs only managed to reduce the WER by 0.7% and 1.7%, respectively. Moreover, the WER of the enhanced beamformed speech from ReVISE is only 4.4% away from resynthesized clean speech: the target ReVISE is trained to predict. One can expect the gap to clean speech (14.5% WER) will be reduced if the quality of the SSL units improves.

Tab. 7 presents the subjective quality evaluation results for EasyCom. We first note that the quality of resynthesized

target audio (4.13 MOS) is in fact higher than the target audio itself (3.76 MOS). This is because the P-TTS module is trained on the high-quality LJSpeech, which has better quality than the EasyCom audio recorded by the close-talking microphones, as the latter still contain mild background noise and overlapping speech. Similarly, the predicted audio from the proposed ReVISE model also delivers similar levels of audio quality (4.19 and 4.11 with ch2 and beamformed input, respectively), which are significantly better than the original distant recordings (2.95 and 2.67) and the baseline methods (2.95 and 2.39 from Demucs).

The results suggests a crucial advantage of the proposed ReVISE model compared to previous studies: for ReVISE, the audio quality is not bounded by the enhancement training data (LRS3 and EasyCom), but determined by the quality of P-TTS training data. In contrast, prior works such as SVTS or VisualVoice could only produce audio that is as good as the enhancement target.

5.8. Limitations

We see from Tabs. 1 to 4 that while ReVISE outperforms or is on par with baseline methods on intelligibility, quality, and synchronization, it performs much worse than SVTS, Demucs, and VisualVoice on recovering low-level details. This is because ReVISE does not focus on reconstructing the exact reference signal; moreover, ReVISE does not aim to infer the speaker's voice either since the SSL units are shown to encode very little speaker information.

In fact, it is not a deficiency that the proposed ReVISE model does not reconstruct the reference signal exactly. As shown in Tab. 7, the quality of the reference signal may be mediocre for real world datasets such as EasyCom. In that scenario, reconstructing the exact reference signal would lead to inferior performance compared to the proposed approach, where the training data for P-AVSR and P-TTS are decoupled, such that higher quality data can be used for the unit-to-speech module.

5.9. Ablation studies

We conduct ablation studies to evaluate the importance of pre-training P-AVSR, P-AVSR model size, selection of audio SSL units, and visual input. WER are reported here because intelligibility exhibits the highest variation across models reported earlier. Tab. 8 compares different SSL speech units on two tasks: clean speech resynthesis and video-to-speech generation. The two alternative types of units are derived from the BASE AV-HuBERT model trained on LRS3+VoxCeleb2 [49] and the BASE HuBERT model trained on Librispeech [23]. The corresponding P-TTS model are also trained on the same LJSpeech dataset. We see that performance on the two tasks are correlated, where units that loss less content information (lower Resyn WER) are better prediction targets for the P-AVSR model (lower

	video-to-speech WER↓	speech inpainting WER↓			speech denoising WER↓				source separation WER↓				Avg.↓
		20%	30%	40%	lvl 1	lvl 2	lvl 3	lvl 4	lvl 1	lvl 2	lvl 3	lvl 4	
Separate ReVISE	33.9%	14.9%	18.1%	19.2%	9.4%	9.7%	11.7%	20.5%	9.9%	10.2%	11.4%	15.7%	19.0%
Universal ReVISE	34.5%	14.9%	17.9%	19.2%	8.7%	9.4%	11.5%	20.3%	9.1%	9.7%	10.9%	15.5%	18.9%

Table 5. Comparing corruption-specific ReVISE with universal ReVISE on intelligibility measured by WER (%).

Method	Mod	WER↓ (ch2) (dev/test)	WER↓ (bf) (dev/test)
Tgt. audio	A	35.7% / 37.6%	
Tgt. audio resynthesis	A	43.2% / 47.7%	
Audio-visual speech enhancement			
Inp. audio	A	74.2% / 87.5%	60.7% / 71.5%
Demucs [12]	A	76.9% / 86.8%	63.4% / 69.8%
Resynthesis	A	79.6% / 91.0%	68.7% / 77.6%
ReVISE (Ours)	AV	50.3% / 55.0%	47.6% / 52.1%

Table 6. Audio-visual speech enhancement results on EasyCom (ch2: channel 2, bf: beamforming). Synchronization results are not reported because SyncNet fails even on clean reference x_a.

Method	Mod	MOS↑ (ch2)	MOS↑ (bf)
Tgt. audio	A	$3.76_{\pm 0.04}$	
Tgt. audio resynthesis	A	$4.13_{\pm 0.03}$	
Audio-visual speech enhancement			
Inp. audio	A	$2.95_{\pm 0.04}$	$2.67_{\pm 0.03}$
Demucs [12]	A	$2.95_{\pm 0.04}$	$2.39_{\pm 0.03}$
Resynthesis	A	$3.97_{\pm 0.04}$	$3.98_{\pm 0.03}$
ReVISE (Ours)	AV	$4.19_{\pm 0.02}$	$4.11_{\pm 0.04}$

Table 7. Subjective quality evaluation of audio-visual speech enhancement results on EasyCom test set (ch2: channel 2, bf: beamforming).

Unit	Resyn WER↓	V2S WER↓
AV-HuBERT (LRS3+VC2) [49]	20.7	41.2
HuBERT (LS960) [23]	17.4	35.2
HuBERT (VP+MLS+CV)	10.6	33.9

Table 8. Ablation studies on SSL unit choices. Clean audio resynthesis (Resyn) and video-to-speech (V2S) WERs are reported.

P-AVSR Size	V2S WER↓	
	scratch	PT
BASE	81.0	42.9
LARGE	78.6	35.5

		WER↓			
Mod		Den	Sep	Inpaint	V2S
AV		12.6	11.4	17.4	34.8
A		29.2	39.8	44.6	n/a

Table 9. (Left) Ablation studies on model sizes and pre-training. Performance on video-to-speech is reported. (Right) Comparison of ReVISE and its audio-only counterpart.

We observe that the model using both audio and video as input outperform the counterpart consistently with a large margin. Specifically, the level of degradation varies substantially across task: denoising is least effected (16.6%), followed by inpainting (27.2%) and separation (28.4%) — this shows that models rely heavily on visual information to determine the target speaker as well as infill the missing information from dropped frames. Last but not least, we note that the audio-visual ReVISE perform decently on video-to-speech synthesis despite not being trained on the task at all, suggesting zero-shot generalization to unseen distortion.

6. Conclusion

This paper presents ReVISE, a novel paradigm for audio-visual speech regeneration that is universal and benefits greatly from the recent advances in self-supervised speech pre-training. Empirical studies demonstrate effectiveness of ReVISE on a popular benchmark with synthetic noise (LRS3) as well as on a conversational benchmark with real noisy data collected from extremely challenging acoustic conditions (EasyCom). In particular, audio enhanced by ReVISE is significantly more intelligible and of higher quality compared to those from the previous work.

While this paper studies universal regeneration, it only concerns audio but not video distortion. For future work, we hope to study more general multimodal speech regeneration where the model can recover distortion in both modalities.

V2S WER). Tab. 9 (Left) shows pre-training brings significant gains (38.9% for BASE and 43.1% for LARGE) and LARGE model performs better with a 7.4% WER reduction.

Finally, we study the importance of visual input by comparing an audio-visual ReVISE with an audio-only version of it. We consider the universal speech regeneration model setup but remove video-to-speech data as it is incompatible to the audio-only model. Tab. 9 (Right) presents aggregated results averaged over testing splits for each task.

References

[1] Amir Adler, Valentin Emiya, Maria G Jafari, Michael Elad, Rémi Gribonval, and Mark D Plumbley. Audio inpainting. IEEE Transactions on Audio, Speech, and Language Processing, 20(3):922–932, 2011. 1

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. arXiv preprint arXiv:1809.00496, 2018. 2, 4

[3] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. arXiv preprint arXiv:1912.06670, 2019. 4

[4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33:12449–12460, 2020. 3, 13

[5] Hanan Beit-On, Moti Lugasi, Lior Madmoni, Anjali Menon, Anurag Kumar, Jacob Donley, Vladimir Tourbabin, and Boaz Rafaely. Audio signal processing for telepresence based on wearable array in noisy and dynamic scenes. In International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 8797–8801, 2022. 4, 12

[6] Juan P Bello, Claudio Silva, Oded Nov, R Luke Dubois, Anish Arora, Justin Salamon, Charles Mydlarz, and Harish Doraiswamy. Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution. Communications of the ACM, 62(2):68–77, 2019. 1

[7] Shlomo E Chazan, Lior Wolf, Eliya Nachmani, and Yossi Adi. Single channel voice separation for unknown number of speakers under reverberant and noisy settings. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3730–3734. IEEE, 2021. 2

[8] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In Asian conference on computer vision, pages 87–103. Springer, 2016. 5

[9] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In Asian conference on computer vision, pages 251–263. Springer, 2016. 5

[10] Martin Cooke. A glimpsing model of speech perception in noise. The Journal of the Acoustical Society of America, 119(3):1562–1573, 2006. 1

[11] 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. 5

[12] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. Real time speech enhancement in the waveform domain. arXiv preprint arXiv:2006.12847, 2020. 6, 7, 8

[13] Jacob Donley, Vladimir Tourbabin, Jung-Suk Lee, Mark Broyles, Hao Jiang, Jie Shen, Maja Pantic, Vamsi Krishna Ithapu, and Ravish Mehra. Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments. arXiv preprint arXiv:2107.04174, 2021. 2, 4, 12

[14] 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

[15] 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. 1

[16] Hakan Erdogan, John R Hershey, Shinji Watanabe, and Jonathan Le Roux. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 708–712. IEEE, 2015. 2

[17] Aviv Gabbay, Asaph Shamir, and Shmuel Peleg. Visual speech enhancement. arXiv preprint arXiv:1711.08789, 2017. 1

[18] Sharon Gannot, Emmanuel Vincent, Shmulik Markovich-Golan, and Alexey Ozerov. A consolidated perspective on multimicrophone speech enhancement and source separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(4):692–730, 2017. 4, 12

[19] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 15490–15500. IEEE, 2021. 1, 2, 6, 7

[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. 2, 4

[21] Michael Hassid, Michelle Tadmor Ramanovich, Brendan Shillingford, Miaosen Wang, Ye Jia, and Tal Remez. More than words: In-the-wild visually-driven prosody for text-to-speech. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10587–10597, 2022. 2, 5

[22] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415, 2016. 5

[23] 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 Transactions on Audio, Speech, and Language Processing, 29:3451–3460, 2021. 2, 3, 7, 8

[24] Wei-Ning Hsu and Bowen Shi. u-hubert: Unified mixed-modal speech pretraining and zero-shot transfer to unlabeled modality. In NeurIPS, 2022. 5

[25] Keith Ito and Linda Johnson. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017. 4

[26] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. Libri-light: A benchmark for asr

with limited or no supervision. In ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7669–7673, 2020. https://github.com/facebookresearch/libri-light. 5

[27] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. Advances in Neural Information Processing Systems, 33:17022–17033, 2020. 3

[28] Robert Kubichek. Mel-cepstral distance measure for objective speech quality assessment. In Proceedings of IEEE pacific rim conference on communications computers and signal processing, volume 1, pages 125–128. IEEE, 1993. 5

[29] 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. 3

[30] Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, et al. Direct speech-to-speech translation with discrete units. arXiv preprint arXiv:2107.05604, 2021. 3, 4

[31] Richard Lippmann. Speech perception by humans and machines. In Proceedings of the ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception, pages 309–316. Keele, England, 1996. 1

[32] Haohe Liu, Qiuqiang Kong, Qiao Tian, Yan Zhao, DeLiang Wang, Chuanzeng Huang, and Yuxuan Wang. Voicefixer: Toward general speech restoration with neural vocoder. arXiv preprint arXiv:2109.13731, 2021. 2

[33] Yi Luo and Nima Mesgarani. Tasnet: time-domain audio separation network for real-time, single-channel speech separation. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 696–700. IEEE, 2018. 2

[34] Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. IEEE/ACM transactions on audio, speech, and language processing, 27(8):1256–1266, 2019. 2

[35] Pingchuan Ma, Stavros Petridis, and Maja Pantic. End-to-end audio-visual speech recognition with conformers. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7613–7617. IEEE, 2021. 4

[36] Rodrigo Mira, Alexandros Haliassos, Stavros Petridis, Björn W Schuller, and Maja Pantic. Svts: Scalable video-to-speech synthesis. arXiv preprint arXiv:2205.02058, 2022. 1, 2, 5, 12

[37] Rodrigo Mira, Konstantinos Vougioukas, Pingchuan Ma, Stavros Petridis, Björn W Schuller, and Maja Pantic. End-to-end video-to-speech synthesis using generative adversarial networks. IEEE Transactions on Cybernetics, 2022. 2, 5

[38] Giovanni Morrone, Daniel Michelsanti, Zheng-Hua Tan, and Jesper Jensen. Audio-visual speech inpainting with

deep learning. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6653–6657. IEEE, 2021. 1, 2, 5, 6

[39] 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

[40] Arun Asokan Nair and Kazuhito Koishida. Cascaded time+ time-frequency unet for speech enhancement: Jointly addressing clipping, codec distortions, and gaps. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7153–7157. IEEE, 2021. 2

[41] Arun Narayanan and DeLiang Wang. Ideal ratio mask estimation using deep neural networks for robust speech recognition. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pages 7092–7096. IEEE, 2013. 2

[42] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. arXiv preprint arXiv:1904.08779, 2019. 13

[43] Santiago Pascual, Joan Serrà, and Antonio Bonafonte. Towards generalized speech enhancement with generative adversarial networks. arXiv preprint arXiv:1904.03418, 2019. 2

[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. 2, 3, 4

[45] Adam Polyak, Lior Wolf, Yossi Adi, Ori Kabeli, and Yaniv Taigman. High fidelity speech regeneration with application to speech enhancement. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7143–7147. IEEE, 2021. 2

[46] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. Mls: A large-scale multilingual dataset for speech research. arXiv preprint arXiv:2012.03411, 2020. 4

[47] Flávio Ribeiro, Dinei Florêncio, Cha Zhang, and Michael Seltzer. Crowdmos: An approach for crowdsourcing mean opinion score studies. In 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP), pages 2416–2419. IEEE, 2011. 5

[48] Joan Serrà, Santiago Pascual, Jordi Pons, R Oguz Araz, and Davide Scaini. Universal speech enhancement with score-based diffusion. arXiv preprint arXiv:2206.03065, 2022. 2

[49] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. arXiv preprint arXiv:2201.02184, 2022. 2, 3, 4, 5, 7, 8, 13

[50] Jiaqi Su, Zeyu Jin, and Adam Finkelstein. Hifi-gan-2: Studio-quality speech enhancement via generative adversarial networks conditioned on acoustic features. In 2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pages 166–170. IEEE, 2021. 2

[51] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Transactions on Audio, Speech, and Language Processing, 19(7):2125–2136, 2011. 5

[52] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Transformer vq-vae for unsupervised unit discovery and speech synthesis: Zerospeech 2020 challenge. arXiv preprint arXiv:2005.11676, 2020. 3

[53] Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. arXiv preprint arXiv:2101.00390, 2021. 4

[54] Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, Desh Raj, et al. Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. arXiv preprint arXiv:2004.09249, 2020. 1

[55] Donald S Williamson and DeLiang Wang. Time-frequency masking in the complex domain for speech dereverberation and denoising. IEEE/ACM transactions on audio, speech, and language processing, 25(7):1492–1501, 2017. 2

[56] Qiantong Xu, Alexei Baevski, Tatiana Likhomanenko, Paden Tomasello, Alexis Conneau, Ronan Collobert, Gabriel Synnaeve, and Michael Auli. Self-training and pre-training are complementary for speech recognition. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3030–3034. IEEE, 2021. 5

[57] Karren Yang, Dejan Marković, Steven Krenn, Vasu Agrawal, and Alexander Richard. Audio-visual speech codecs: Rethinking audio-visual speech enhancement by re-synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8227–8237, 2022. 1

[58] Hang Zhou, Ziwei Liu, Xudong Xu, Ping Luo, and Xiaogang Wang. Vision-infused deep audio inpainting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 283–292, 2019. 5