

## Article

# Enabling Real-Time On-Chip Audio Super Resolution for Bone-Conduction Microphones

Yuang Li <sup>1,2</sup> , Yuntao Wang <sup>1,\*</sup> , Xin Liu <sup>3</sup>, Yuanchun Shi <sup>1</sup>, Shwetak Patel <sup>3</sup> and Shao-Fu Shih <sup>4</sup>

<sup>1</sup> Key Laboratory of Pervasive Computing, Ministry of Education, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

<sup>2</sup> Department of Engineering, University of Cambridge, Cambridge CB2 1TN, UK

<sup>3</sup> Department of Computer Science and Engineering, Paul G. Allen School of Computer, University of Washington, Seattle, WA 98195, USA

<sup>4</sup> Google Inc., Mountain View, CA 94043, USA

\* Correspondence: yuntaowang@tsinghua.edu.cn

**Abstract:** Voice communication using an air-conduction microphone in noisy environments suffers from the degradation of speech audibility. Bone-conduction microphones (BCM) are robust against ambient noises but suffer from limited effective bandwidth due to their sensing mechanism. Although existing audio super-resolution algorithms can recover the high-frequency loss to achieve high-fidelity audio, they require considerably more computational resources than is available in low-power hearable devices. This paper proposes the first-ever real-time on-chip speech audio super-resolution system for BCM. To accomplish this, we built and compared a series of lightweight audio super-resolution deep-learning models. Among all these models, ATS-UNet was the most cost-efficient because the proposed novel Audio Temporal Shift Module (ATSM) reduces the network's dimensionality while maintaining sufficient temporal features from speech audio. Then, we quantized and deployed the ATS-UNet to low-end ARM micro-controller units for a real-time embedded prototype. The evaluation results show that our system achieved real-time inference speed on Cortex-M7 and higher quality compared with the baseline audio super-resolution method. Finally, we conducted a user study with ten experts and ten amateur listeners to evaluate our method's effectiveness to human ears. Both groups perceived a significantly higher speech quality with our method when compared to the solutions with the original BCM or air-conduction microphone with cutting-edge noise-reduction algorithms.



**Citation:** Li, Y.; Wang, Y.; Liu, X.; Shi, Y.; Patel, S.; Shih, S.-F. Enabling Real-Time On-Chip Audio Super Resolution for Bone-Conduction Microphones. *Sensors* **2023**, *23*, 35. <https://doi.org/10.3390/s23010035>

Academic Editors: Shyan-Ming Yuan, Zeng-Wei Hong and Wai-Khuen Cheng

Received: 2 November 2022

Revised: 15 December 2022

Accepted: 15 December 2022

Published: 20 December 2022



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The most commonly used microphones for voice communication are air-conduction microphones, which pick up sound propagating through the air. Although providing high fidelity capture in quiet scenarios, they are vulnerable to environmental noises. To improve the speech quality of air-conduction microphones in noisy environments, researchers proposed multi-microphone beamforming with noise suppression techniques [1–3] or deep-learning-based speech enhancement methods [4,5]. However, these solutions require a significant amount of additional hardware or computing resources. Further, all these methods fundamentally seek to reduce environmental noises but also inevitably corrupt speech. Moreover, these solutions are still vulnerable to boisterous environments, such as construction sites or strong wind, where extraneous noises overpower speech signals.

Bone-conduction microphones (BCMs) could achieve more robust results against ambient noises due to their physical design and fundamental conduction principles. However, BCMs only have limited frequency response with high-frequency components above 2 kHz

significantly attenuated. Reconstructing the high-frequency details can effectively increase the speech audio's quality.

A traditional method of reconstruction is to design a linear phase impulse response filter [6]. However, acoustic paths are different among speakers because their bone structures are unique. Furthermore, it is impossible to ensure uniform BCM placement, which may result in different spectral properties [7]. Therefore, a simple filter is insufficient to accommodate a variety of users.

Audio super resolution [8], also called bandwidth expansion, is the task of increasing the audio sampling rate and restoring the high-frequency components of low-resolution audios. Convolutional Neural Networks have achieved state-of-the-art performance in audio super resolution [9–11]. Additionally, similar neural-network structures have also been proven effective in reconstructing distorted spectrograms [12] and enhancing recordings from low-end microphones [13,14]. Therefore, designing an audio super-resolution model is feasible to reproduce high-fidelity speech from BCMs while maintaining their noise resistance property in multi-speaker settings. However, existing deep-learning-based audio super-resolution methods are commonly computationally intensive, making them unfit for deployment on resource-constrained embedded systems.

This paper proposes the first-ever real-time on-chip speech audio super-resolution system for BCMs. In order to achieve this goal, we first designed and compared a series of lightweight deep-learning models for speech audio super resolution. Among all the models, ATS-UNet is the most cost-efficient. We proposed an audio temporal shift module (ATSM) and introduced this module to ATS-UNet. Therefore, ATS-UNet can reduce the network to one dimension but still learn sufficient features from the temporal information flow in speech audios.

Thus, ATS-UNet can reconstruct high-fidelity speech audios but require minimal computational resources. We further quantized and deployed ATS-UNet and its variants on micro-controllers, including ARM Cortex-M4f and M7 processors, and conducted a full evaluation of our proposed method's performance regarding the audio quality and inference latency. The results show that ATS-UNet outperformed the cutting-edge audio super-resolution method [9] with the perceptual evaluation of speech quality (PESQ) [15] increased by 9% and log-spectral distance (LSD) [16] by 44%.

On the Cortex-M7 processor, our end-to-end latency, comprising model inference, feature extraction, and reconstruction, is 38 ms on average. This is less than the half-frame length (64 ms), meaning that our system can achieve real-time processing with 128 ms frames half-overlapped. To further assess our method's effectiveness in obtaining high quality speech, we recruited 20 participants, including 10 experts and 10 amateur listeners, for the perceptual audio quality evaluation. The results show that our method outperformed the original BCM solution and commodity noise reduction solution with the air-conduction microphone. To the best of our knowledge, our method is the first chip-deployable audio super-resolution solution.

#### To summarize, our contributions are as follows:

- We propose a lightweight audio super-resolution deep-learning model—ATS-UNet—that utilizes our proposed audio temporal shift module (ATSM) to form a novel one-dimensional UNet architecture. When compared with ATS-UNet's variants without ATSM, ATS-UNet was the most cost-efficient for chip deployment.
- We implement the first-ever real-time on-chip speech audio super-resolution system for the bone-conduction microphone by quantizing and deploying ATS-UNet to popular micro-controllers in commodity wearable devices. We further evaluate its computational complexity on both ARM Cortex-M4f and M7 processors and demonstrated its real-time processing capability.
- We evaluate our system's effectiveness in improving speech quality with a bone-conduction microphone through perceptual audio quality user studies. Audio samples are publicly available (<https://sites.google.com/view/audio-sr-for-bcm/home> (accessed on 1 November 2022)).

## 2. Background and Related Work

Voice communication using air-conduction microphones in noisy environments has been a challenging problem. For conventional speech communication, researchers have proposed speech enhancement methods, such as beamforming with a microphone array and blind source separation [17], for background noise removal. These algorithms only remove part of the unwanted noises and introduce the risk of damaging the voice integrity. Beamforming is based on directionality.

Therefore, it is prone to directional noise sources. In other words, when noise and voice sources are on-axis, beamforming will not effectively separate the noise. To reduce on-axis noise, noise suppression algorithms, such as [3,18,19] first estimate the noise with statistical models and then remove the noise from the captured spectrum to recover the original speech. These methods could lead to speech integrity issues due to the noise model estimation accuracy. Moreover, under extreme conditions, such as strong wind noise, air-conduction microphones will not pick up human voices due to saturation.

A bone-conduction microphone, which collects human speech propagated via human bones, naturally suppresses environmental noises with its hardware placement and FSV conduction. However, the speech captured by the BCM has a limited frequency bandwidth which attenuates quickly above 2 kHz [20]. Our motivation is to enhance the BCM's speech sound quality by recovering high-frequency details while keeping its advantage against environmental noise. In this section, we describe existing speech enhancement algorithms for BCM and then give an overview on speech super-resolution techniques.

### 2.1. Bone-Conduction Microphones

Bone-conduction microphones are commonly used as an accessorial enhancer to air-conduction microphones for capturing human speech. Researchers have proposed speech enhancement methods using BCMs [21–25]. The BCMs can be used for accurate voice activity detection due to their noise suppression characteristics [24]. BCMs can also be incorporated to increase the voice activity detector accuracy and, hence, increase the accuracy for noise model estimation to achieve better denoising results [22].

Further, BCMs can provide additional input for a multi-modal deep-learning network [25]. However, these solutions require multiple microphones that are costly and limited in capability in extreme circumstances, such as strong wind noise. Our work aims to enhance speech quality using a single bone-conduction microphone. In other words, we plan to achieve clean human speech capture while maintaining the microphone's capability against environmental noises.

Similar speech processing techniques based on BCM speech capture with audio super resolution can be found, including the following: speech enhancement approaches for bone-conduction microphones through audio signal processing [20]. Shimamura and Tamiya [6] proposed a reconstruction filter calculated from long-term spectra of human voices from both air- and bone-conduction microphones.

Shimamura et al. [26] further utilized a multi-layer perceptron to model the reconstruction filter more accurately. Rahman and Shimamura [27] excluded the need for the air-conduction microphones by introducing an analysis-synthesis method based on linear prediction. Bouzerhal et al. [28] introduced adaptive filtering along with non-linear bandwidth extension method for enhancing the speech sound quality. However, these methods require complex feature engineering and are, thus, difficult to adapt to different users and equipment setups.

Recently, researchers applied deep-learning methods for speech enhancement with BCMs. These methods aim to increase the sound quality of BCMs to be comparable to air-conduction microphones in ideal conditions. For example, Shan et al. [29] proposed an encoder-decoder network with a long short-term memory (LSTM) layer and local attention mechanism, which reconstructs an air conduction log-spectrogram from a bone-conduction log-spectrogram. This method only reconstructs frequency components below 4 kHz and is based on a specific speaker.

Liu et al. [30] introduced Mel-scale features of speech audio from a bone-conduction microphone with a deep denoise auto-encoder for speech enhancement. This work reconstructs high-frequency components up to 8 kHz. This increases the perceptual evaluation of speech quality (PESQ) by 9.38% compared with the original bone-conduction speech; however, the auto-encoder is also trained with a single speaker's speech. Hussain et al. [31] proved that, with only limited training data, the hierarchical extreme-learning machine could outperform the denoise auto-encoder. Zheng et al. [32] adapted structural similarity (SSIM)—a widely used metric in image quality assessment—as the loss function for a Bidirectional LSTM Neural Network. As a result, the model achieved higher PESQ when trained with SSIM loss compared with the standard mean square error (MSE).

Although proven effective, the aforementioned deep-learning methods were not designed to be deployed on real-time embedded systems due to exceeding computation resources and power limits. Moreover, these methods were not evaluated with cross-user validation, which limits their generalizability to adapt to individual users. Some works [29,32] used a sampling rate of 8 kHz, which is not sufficient for the Wideband Speech protocol with required sampling frequency at 16 kHz. Therefore, our work approaches BCM voice capture as a real-time resource-constrained super-resolution problem on embedded systems. Furthermore, to make our solution robust against individual users and various environments, we also introduced transfer learning to make our model generalizable.

## 2.2. Audio Super-Resolution Techniques

The audio super resolution, also known as bandwidth expansion, aims to increase the sampling rate and restore high-frequency components of the low-resolution audio. Inspired by image inpainting methods, researchers have proposed several frequency domain based deep-learning methods for audio super resolution. These methods can be trained using clean samples of BCMs as input and air-conduction microphones as references. These samples are then converted into snapshots of spectrograms in the frequency domain as snippets of audio features. The learned model restores the missing high-frequency details from BCMs based on pattern recognition at the inference time. Then, the output snippets are reconstructed back into the real-time speech stream as output. Below we describe various audio super-resolution methods, optimization strategies, and on-chip deployment methods.

Audio super-resolution methods either took raw waveforms [8–11] or spectral representations [33–35] as the input. A one-dimensional UNet [8] asymmetrical network with skip connections was the first attempt to use a deep convolutional neural network for end-to-end speech super resolution. To expand the perceptual field, TFiLM [9] utilized bidirectional LSTM as the module to build up a variant 1D-UNet for speech audio super resolution. In another variation of 1D-UNet [10], conventional convolutional layers were replaced by multi-scale convolutional layers to capture information at multiple scales. Mfnet [11] also attempted to facilitate multi-scale information exchange through multi-scale fusion block.

Other deep-learning methods utilized the spectrogram as input. For example, Li and Lee [33] proposed a three-layer fully connected network for speech audio super resolution. UNet [34] was also proven to be effective in performing speech audio super resolution using the power spectrogram. To take advantage of representations in both the time and frequency domain, TFnet [35] incorporated two network branches that operate on both the waveform and the spectrogram, respectively. Although proven effective on speech audio super resolution, the deep-learning models mentioned above have too many parameters, which causes them to exceed the computation and power budgets of micro-controllers by a factor of 100 times.

The optimization strategy for audio super resolution includes the loss function and training optimizations. One of the most commonly used loss functions is the mean square error [8,33]. Although simple to compute, the MSE does not represent human perceptual speech quality. Therefore, perceptually motivated loss [36] that calculates the L1 distance on log mel-spectrogram was proposed. Further, the log spectral distance (LSD) [8,16], which

measures the distance between the log-power spectrum of reference and reconstructed signals, was also adopted as one option for the loss function.

For training, WaveNet [37,38], an auto-regressive model, optimized the joint probability of the targeted high-resolution audio. Adversarial learning is another popular training technique. In this technique, a discriminator that works either in the time domain [10,11,39] or frequency domain [34,40,41] guides the generator to predict more realistic high-resolution audio from low-resolution inputs.

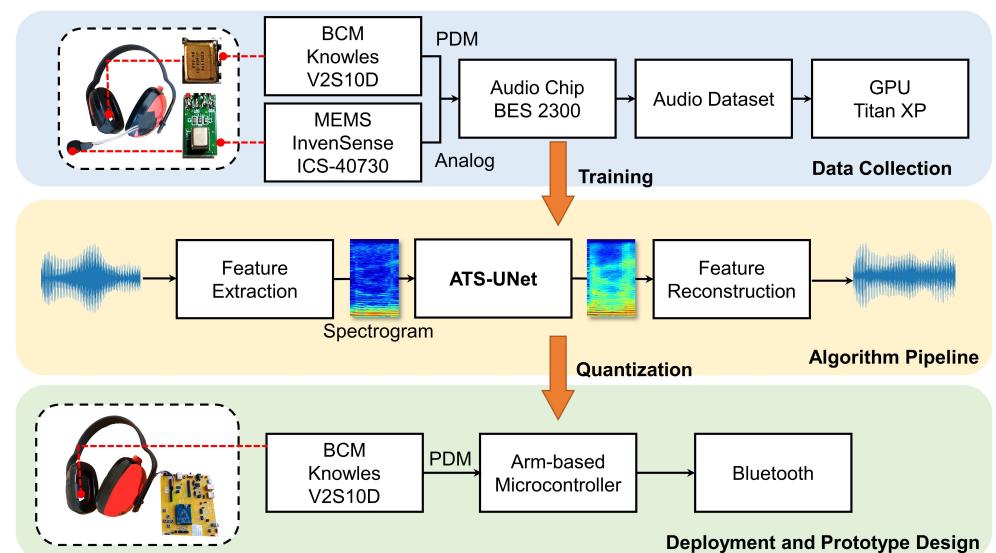
Recently, hearable devices, such as TWS earbuds, have become increasingly popular, with 233 million shipments in 2020, while deep-learning-based audio super-resolution methods have been proven effective, deploying such solutions to a resource-limited embedded system has not been fully investigated. Similar to our proposal, several super-resolution deep-learning methods [39,42,43] have proven the feasibility of applying the super-resolution method on a smartphone.

Other state-of-the-art speech super-resolution models require considerable computation resources and cause significant latency, which is not suitable for edge device deployment. This paper proposed a lightweight deep-learning model—ATS-UNet, which can run on power- and space-limited ARM Cortex-M platforms. We expect future hearables embedded with a single BCM will achieve good performance without the need for additional computation resources with the proposed method.

### 3. Overview

This paper focuses on the uplink portion of the communication system—namely, the capture side of the speech communication protocol. In particular, the capture and recovery of the BCM input as an alternative solution to the conventional air-conduction microphone. Shown in Figure 1, our proof-of-concept prototype is composed of commercially available electronic parts: a pulse density modulation (PDM) bone-conduction microphone (Knowles V2S10D), an analog MEMS air-conduction microphone (InvenSense ICS-40730), and a micro-controller development board (Bestechnic ([http://www.bestechnic.com/Home/Index/index/lan\\_type/2](http://www.bestechnic.com/Home/Index/index/lan_type/2)) (accessed on 1 November 2022)) (BES) BES2300YP).

The BES2300YP system on chip (SoC) simultaneously collects audio signals from the Knowles V2S10D and InvenSense ICS-40730, forming a dataset for audio super resolution. Then, we trained ATS-UNet using this dataset on an Nvidia Titan XP GPU (12GB RAM). The floating-point model was further quantized to the 16-bit data format and transformed from a Tensorflow [44] to CMSIS-NN [45] implementation. With the model quantization and optimization, ATS-UNet can then run efficiently on micro-controllers.



**Figure 1.** This paper's overview.

In this work, we tested two popular micro-controllers. (1) The BES2300YP with dual ARM-Cortex M4F processors operating at a frequency up to 300 MHz with 992 KB SRAM and 4 MB flash storage. The BES SoC was adopted by many popular TWS earbuds, such as JBL FREE II, Samsung Galaxy Buds Live, and Huawei FreeBuds 2 Pro, for its compact form factor and power efficiency. We only use one single processor in this work since the other processor runs the Bluetooth stack and digital signal processing (DSP)-related algorithms.

Furthermore, the two processors share the SRAM with the storage requirements from the Bluetooth and the operating system taking more than 400 KB. To prevent the memory overflow, we limited the SRAM for the machine-learning model to be below 500 KB. (2) The NXP RT1060 SoC with a single Arm-Cortex M7 processor operating at a frequency up to 600 MHz with 1 MB on-chip SRAM. In this case, only 512 KB general-purpose SRAM can be used to host the machine-learning model. Both micro-controllers support audio applications.

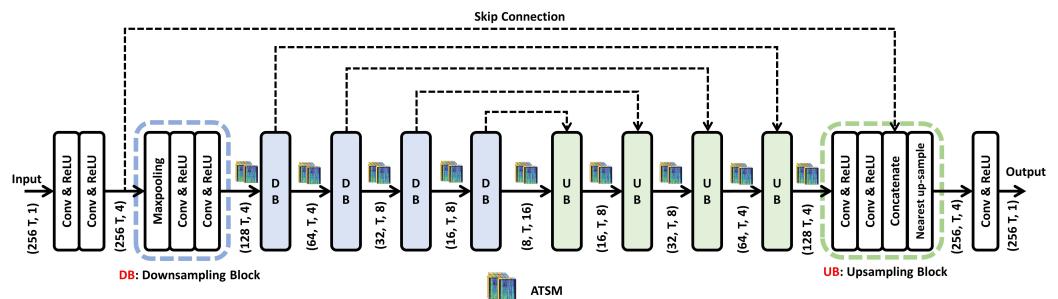
#### 4. Deep-Learning Models for Bone Conduction Speech Audio Super Resolution

In this section, we first describe our general UNet design for bone-conduction speech audio super resolution. We then describe how our models, including our proposed 2D-UNet, Hybrid-UNet, Mixed-UNet, 1D-UNet, and ATS-UNet, were derived from this UNet design. Most importantly, we present the key module called the Audio Temporal Shift Module (ATSM). Finally, we describe the pre-processing and post-processing methods for our deep-learning models.

##### 4.1. UNet Variants for Bone Conduction Speech Audio Super Resolution

The original UNet has a fully convolutional and symmetrical network structure with skip connections to facilitate information flow. Additionally, it can extract temporal and frequency information in the time–frequency domain and reconstruct high-resolution spectrograms. Compared with conventional convolutional and recurrent neural networks, UNet is more efficient, as feature maps are down-sampled, contributing to fewer floating-point operations.

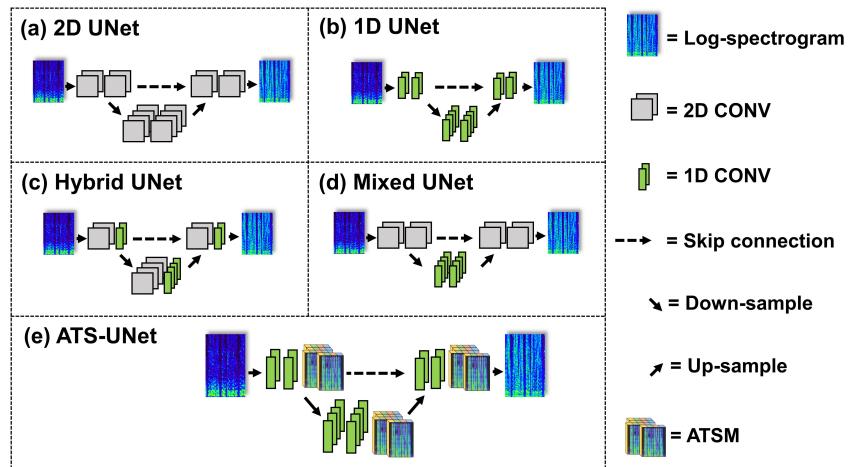
However, UNet’s large model size still introduces unfavorable computation for on-chip deployment. Therefore, we reduce the number of channels and network depth. The general UNet architecture (Figure 2) contains five down-sampling blocks (DB) and five up-sampling blocks (UB). Each DB has a max-pooling layer followed by two convolutional layers. The size of the max-pooling layer is  $2 \times 1$ . Therefore, after each DB, the length of the frequency axis of the feature map is halved, while the time dimension remains the same throughout the network.



**Figure 2.** The detailed architecture of our general UNet design. ( $F, T, C$ ) indicates  $F$  for frequency bins,  $T$  for temporal frames, and  $C$  for channels. When all convolutional layers are 1D and ATSMs are inserted after each DB/UB, it is our ATS-UNet.

In UB, feature maps are up-sampled, concatenated with skip features, and then fed into two convolutional layers. A ReLU activation function is adopted after each convolutional layer except for the last layer. **2D-UNet V1** (Figure 3a) has the same structure and channel numbers as shown in Figure 2; however, it has no ATSM, and every convolutional

layer is 2D. We also present **2D-UNet V2**, which has four times the filters of 2D-UNet v1 for comparison.



**Figure 3.** Our proposed series of novel network architectures including a variant 2D-UNet, 1D-UNet, Hybrid-UNet, Mixed-UNet, and ATS-UNet.

Although 2D-UNet V1 is significantly smaller than the original UNet for image segmentation, 2D convolutional layers still introduce unfavorable computation for on-chip deployment. Since low-latency audio super resolution requires a small frame size, the shape of the input spectrogram is narrow (the frequency axis is much longer than the time axis). Therefore, only a few 2D convolutional layers are sufficient to extract information from the full temporal range. Thus, using 2D convolutional layers throughout the network is unnecessary.

Based on the above observation, We present another architecture called **Hybrid-UNet** (Figure 3c) and **Mixed-UNet** (Figure 3d), which replace a portion of the 2D convolutional layers with 1D convolutional layers. 2D convolutional layers enable temporal information flow, while 1D convolutional layers only compute along the frequency dimension to enlarge the perceptual range. Hybrid-UNet adopts 2D and 1D convolutional layers in each DB/UB alternately, which maintains temporal information flow in the whole network. Mixed-UNet replaces 2D convolutional layers with 1D layers in the middle of the network so that temporal information exchange only exists in shallow layers.

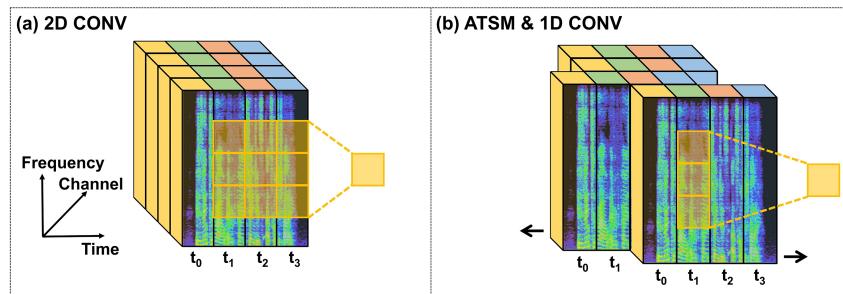
Although Hybrid-UNet and Mixed-UNet are more efficient than traditional 2D-UNet, 2D convolutional layers still introduce unfavorable computation for real-time inference on low-end embedded systems. Thus, we replace 2D convolutional layers with 1D convolutional layers completely to obtain a new architecture called **1D-UNet** (Figure 3b), a 1D version of 2D-UNet V1. However, 1D-UNet lacks temporal modeling; therefore, we inserted ATSM after each DB/UB to enable efficient and effective information exchange along the temporal axis. We called 1D-UNet with ATSM **ATS-UNet** (Figure 3e). For all the models, the kernel sizes of 1D and 2D convolutional layers are  $3 \times 1$  and  $3 \times 3$ , respectively.

#### 4.2. Audio Temporal Shift Module (ATSM)

Conventional deep-learning models for audio processing require massive 2D convolutional operations to extract meaningful features from spectrograms [4] as Figure 4a shows. However, they utilize a large number of computational resources. Therefore, the aforementioned deep-learning models are unlikely to be adopted for on-chip audio super resolution. Instead, we introduced a novel module to accelerate convolutional operations in the time-frequency domain called the Audio Temporal Shift Module (ATSM), as Figure 4b shows. ATSM was inspired by the Temporal Shift Module (TSM) [46], an effective mechanism for video understanding.

This replaces 3D convolutional operations with 2D ones while preserving high-dimensional modeling. This is achieved by shifting the feature maps among video frames to enable temporal information flow. Similarly, ATSM utilizes 1D convolution operations to replace 2D convolution operations for audio processing. In order to utilize information from a longer temporal range for 1D convolutional kernels, ATSM shifts feature maps along the temporal axis of spectrograms. In contrast to the TSM, whose input is four-dimensional feature maps extracted from video frames, the input ATSM is extracted from the 2D log-spectrogram and only has three dimensions: channel, time, and frequency.

More specifically, as illustrated in Figure 4b, feature maps are divided into two chunks along the channel dimension: (1) dynamic and (2) static. The dynamic feature maps are split evenly into two parts, with one part shifted forward (delaying time by one frame) and the other backward (advancing time by one frame). The static feature maps remain unchanged. It is worth noting that ATSM requires no additional computational resources to facilitate information exchange along the temporal dimension in spectrogram computation.



**Figure 4.** To enable information flow through  $t_1, t_2, t_3$ , we can use (a) a 2D convolutional layer or (b) the proposed ATSM with a 1D convolutional layer. The latter is more lightweight.

#### 4.3. Audio Pre-Processing and Resynthesis

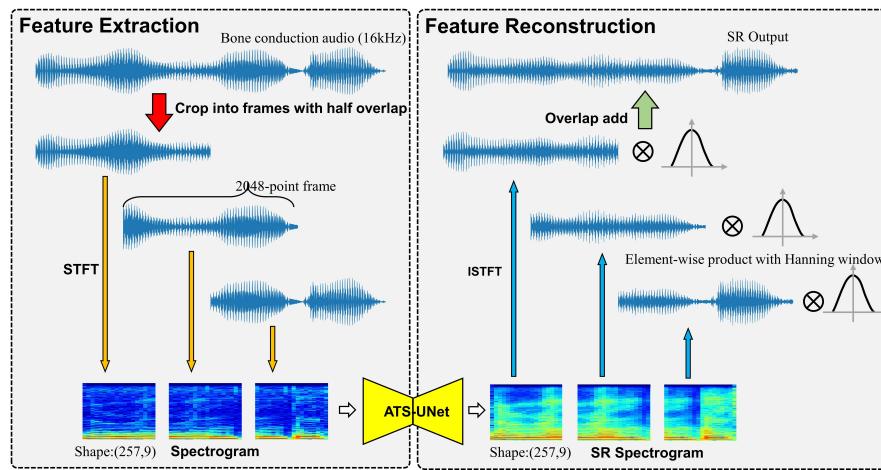
Voice communication's ideal overall latency is below 50 ms, which humans are unable to notice. As latency increases, humans start to notice lip-sync issues; however, communication latency under 150 ms is still considered acceptable. However, a latency that exceeds 400 ms [47] is unacceptable for real-time communication. Therefore, a feasible audio super-resolution system should not add too much latency to the communication process. As a result, our system requires fast computing with an appropriate frame size and short-time Fourier transform (STFT) parameter.

The pre-processing includes the feature extraction from the raw audio signal as shown in the left figure of Figure 5. ATS-UNet processes a single audio frame at a time and outputs audio frames in sequence to resynthesize the audio stream. A large frame provides more information for ATS-UNet but introduces longer latency, since the system has to wait for the time of the entire frame. Therefore, we use a 2048-point (128 ms) frame with half overlap to achieve acceptable latency while maintaining adequate information. These frames are transformed into spectrograms by STFT and fed into ATS-UNet.

The STFT parameter is another major factor in computational intensity. High frequency and time resolution spectrograms can be achieved with a larger fast Fourier transform (FFT) size and overlap between FFT windows, resulting in considerable computation load. Considering the memory and resources on the micro-controller, we adopted a window size of 512 points for the STFT. Further, we also utilized a half overlap strategy to the raw audio data. The detailed trade-off of the STFT parameter is explained in Section 5.

The audio resynthesis, also known as post-processing, converts the reconstructed spectrograms back to the time domain using the inverse short-time Fourier transform (ISTFT). The overlapped frame is then multiplied by the Hanning window (2048-point) and added to the previous frame. We adopted this resynthesis method because the data points in the center of the window are better reconstructed due to richer temporal information. Therefore, the Hanning window function gives the data samples in the center of the window higher weights but weakens the importance of the data samples by the side.

Further, we adopted half-overlapped Hanning window functions so that their summation is a constant value. Thus, it will not distort the signal and can smooth the transition between adjacent frames.



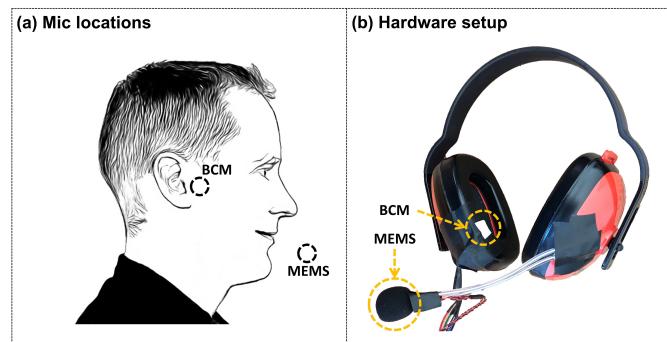
**Figure 5.** Feature extraction and reconstruction.

## 5. Model Training and Deployment

In this section, we provide experimental details, including the data collection procedure, training scheme, and quantization procedure.

### 5.1. Speech Audio Data Collection

We conducted a user experiment to collect an audio dataset using the hardware shown in Figure 6. The MEMS air-conduction microphone was placed near the mouth to collect high-quality ground-truth speech audios. The BCM was secured with an earmuff. Thus, when subjects wore the earmuff, the BCM would be pressed in front of the ear. To prevent reverberation, we placed an acoustic panel in front of the speaker. As Figure 1 indicates, we utilized a BES2300YP micro-controller to simultaneously collect speech audios from both the air- and bone-conduction microphones. The sampling rate and bit depth were set to 44.1 kHz and 16 bits. We recorded the speech audios in a recording studio that was quiet for high speech quality.



**Figure 6.** (a) The position of the air (MEMS) and the bone (BCM) conduction microphone. (b) The headphone prototype for data collection.

We recruited 20 participants (10 males and 10 females). After wearing the headphone, each subject was informed to read six paragraphs of an article, yielding approximately 12 min of speech per subject. We removed the silence clips at the beginning and the end of all audio files and normalized the volume across participants. We then down-sampled each speech audio to 16 kHz, which is sufficient for communication ([https://en.wikipedia.org/wiki/Sampling\\_\(signal\\_processing\)](https://en.wikipedia.org/wiki/Sampling_(signal_processing))) (accessed on 1 November 2022)). The processed

dataset includes 200 min of speech audios in total. Each participant received a 10 USD gift card after the experiment.

### 5.2. Implementation and Training Details

Bone-conduction audios were down-sampled to 16 kHz, cropped to 2048-point frames (128 ms), and transformed to time–frequency representations by short-time Fourier transform (STFT) [48] with 512-point Hanning window and half overlap (a stride of 256-point). We adopted the implementation of librosa (<https://librosa.org/doc/latest/index.html> (accessed on 1 November 2022)) for STFT and ISTFT. The symmetrical component is removed, resulting in 257 Fourier coefficients. Thus, the size of the input spectrogram is  $257 \times 9$ .

Then, we converted the power of each coefficient to the log scale and standardize them to normal distribution. We skipped the 0th coefficient (DC component) but fed the remaining 256 coefficients into the network. Lastly, we obtained the enhanced log-spectrogram from the super-resolution model and concatenated it with 0th coefficient; therefore, the output’s shape is also  $257 \times 9$ . The post-processing audio resynthesis includes denormalization, conversion to linear-scale, and inverse STFT. The model only predicts magnitude, and thus we kept the original phase information from the bone-conduction audio to resynthesize the enhanced speech audio.

All super-resolution models were implemented in Tensorflow [44]. We adopted cross user validation with the training dataset consisting of speech audios from 18 speakers and the remaining audios as the test dataset. We randomly initialized the model and trained it for 100 epochs using the Adam optimizer [49] with a learning rate of 0.0001 and batch size of 64.

### 5.3. Loss Function

The loss function is given by Equation (1), that consists of two parts: the least absolute deviation (L1) loss and perceptually motivated loss [36]. L1 loss measures the absolute difference between the log-spectrograms of the output speech audio and the ground-truth speech audio— $\log(s(y))$ . Perceptually motivated loss is the L1 distance calculated on log-melspectrogram  $\log(ms(y))$  considering that the mel-scale is more aligned with human hearing [50]. In Equation (1),  $s(y)$  and  $s(\hat{y})$  stand for spectrograms of the output and ground-truth audios.  $ms(y)$  and  $ms(\hat{y})$  represent melspectrograms of the output and ground-truth audios, respectively.

$$\text{Loss} = |\log(s(y)) - \log(s(\hat{y}))|_1 + |\log(ms(y)) - \log(ms(\hat{y}))|_1 \quad (1)$$

### 5.4. Model Quantization

We re-compiled each model using the CMSIS-NN [45] framework for efficient inference on Arm Cortex-M processors. First, we transformed the model from floating-point to fixed-point format. Both weights and activations were quantized to 16-bit integers, given by Equation (2). In practice, quantization is symmetrical around zero with power-of-two scaling; therefore, it can be implemented by bitwise shifts in CMSIS-NN kernels.

$$x_q = \lfloor x \times 2^{15-\log_2 \max|x|} \rfloor, \quad (2)$$

where  $x$  represents the weights of a convolutional layer.  $x_q$  is the quantized weights.

### 5.5. Noise Transfer Learning

The primary motivation behind the use of BCM is to enhance the communication quality in noisy environments. Therefore, our system should improve bone-conduction recordings in a quiet environment and in boisterous environments. To this end, we collected voices from BCM in different noisy locations. However, we observed more unwanted noises in the reconstructed speech compared with the quiet laboratory setup. This is because BCM can still pick up some external noises that are further being enhanced by the model.

We adopted transfer learning to fine-tune the model for noise reduction. By collecting pure noises using BCM and adding them to bone-conduction audio in the training dataset, the model can learn to identify the unwanted noises and only recover speech signals. In detail, we collected bone-conduction noises in three locations, including a subway station, a bus stop, and a dining hall. We instructed a participant to wear the prototype without speaking and recorded bone-conduction noises for 20 min in each location.

We then added the noises to the bone-conduction speech recorded in the quiet studio. For each audio clip, the signal-to-noise ratio (SNR) between the bone-conduction speech and the additive noise was randomly sampled from Gaussian distribution with a mean of 18 and a standard deviation of 3.5. Before being deployed in real-world environments, ATS-UNet was fine tuned on the noisy data for another 100 epochs with the same parameters in Section 5.2.

## 6. Quantitative Speech Quality Evaluation

In this section, we present the quantitative speech quality evaluation regarding the air-conduction microphone as the ground truth. We describe the evaluation metrics, baselines, and results. We then explain the reasons behind the hyper-parameter selection and benchmark the performance of ATS-UNet, UNet variants, and baselines for speech enhancement.

### 6.1. Evaluation Metrics

We considered the effectiveness, model size, latency, and power consumption to evaluate each model's performance comprehensively. Specifically, the effectiveness includes two metrics: the log spectral distance (LSD) [16], and the perceptual evaluation of speech quality (PESQ) [15]. LSD, given by Equation (3) [8], measures the distance between the log-power spectrum of reference and reconstructed signals. Therefore, a lower value indicates a better performance. PESQ was provided by Recommendation ITU-T P862 [15] for the objective assessment of speech quality. This models the mean opinion score (MOS), which ranges from 1 (bad) to 5 (excellent).

$$LSD(x, \hat{x}) = \frac{1}{T} \sum_{t=1}^T \sqrt{\frac{1}{K} \sum_{k=1}^K (X(t, k) - \hat{X}(t, k))^2}, \quad (3)$$

where  $t$  and  $k$  are the frame and frequency index, respectively.  $X$  and  $\hat{X}$  denote the log-power spectrum of  $x$  and  $\hat{x}$ , which are defined as  $X = \log|S(x)|^2$ .  $S$  stands for STFT with 2048-point frames.

### 6.2. Baselines

Birnbaum et al. [9] inserted temporal feature-wise linear modulation (TFiLM) layers into a time-domain 1D-UNet to expand the receptive field. This improved the performance of audio super resolution compared with the original 1D-UNet [8], achieving cutting-edge audio super resolution performance. Therefore, we adopted TFiLM as the baseline in this paper. We used the open-sourced code of TFiLM implementation (<https://github.com/kuleshov/audio-super-res> (accessed on 1 November 2022)).

### 6.3. Effect of the Input Hyper-Parameter

To evaluate the trade-off between frequency resolution and model performance, we first compared two sets of STFT parameters: 1024-point FFT, 256 strides, and Blackman window as well as 512-point FFT, 256 strides, and Hanning window. The experiments were performed on two models. The first model is a lightweight 2D-UNet v1. In the second model, we expanded 2D-UNet v1 by increasing the number of filters in each layer by four times to explore the optimum audio super resolution results without considering computation.

As shown in Table 1, both 2D-UNet v1 and v2 outperformed the baseline method—TFiLM [9] with significantly fewer parameters. This proves the effectiveness of 2D-UNet in speech audio super resolution. The computational intensity of the 1024-point STFT was

nearly doubled compared with the 512-point STFT; however, the performances were close. Therefore, we adopted a 512-point window size for the STFT in the following evaluation procedures considering the latency and model size.

**Table 1.** Performance comparison for different STFT parameters.

<b>Model</b>	<b>Params</b>	<b>Average LSD(dB)/PESQ</b>	
		<b>1024-Point STFT</b>	<b>512-Point STFT</b>
2D-UNet v1	11.8 k	2.028/2.713	2.013/2.790
2D-UNet v2	187.3 k	1.949/2.937	1.961/2.983
TFiLM (baseline) [9]	68,221.2 k	3.646/2.523 (time domain)	

#### 6.4. Model Performance Results and Comparison

To align the loss function with human hearing sensitivity for different frequency ranges, we incorporated perceptually motivated loss [36]. Compared with the L1 loss, this increased the accuracy for every tested architecture (Table 2).

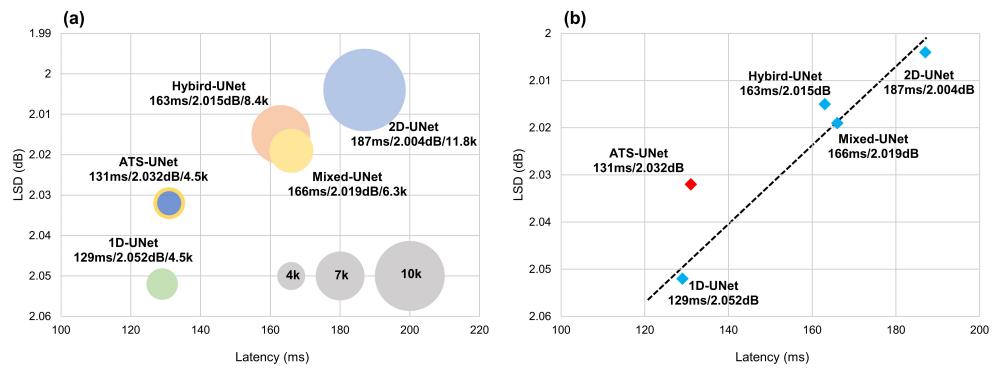
**Table 2.** Performance comparison for different UNet architectures. Latency is the model inference time on a single 2048-point frame by Arm Cortex-M4f/M7 processor. For 2D-UNet v2 and TFiLM, latency was not provided as they are too large to be deployed on our embedded system.

<b>Our Models</b>	<b>Params</b>	<b>FLOPs</b>	<b>Latency (ms)</b>		<b>Average LSD(dB)/PESQ</b>	
			<b>Cortex-M4f</b>	<b>Cortex-M7</b>	<b>L1</b>	<b>L1 + Perceptual Loss</b>
2D-UNet v2	187.3 k	133.9 M	/	/	1.961/2.983	1.954/3.030
2D-UNet v1	11.8 k	8.6 M	187	44	2.013/2.790	2.004/2.780
Hybrid-UNet	8.4 k	7.0 M	163	38	2.024/2.689	2.015 /2.733
Mixed-UNet	6.3 k	6.9 M	166	39	2.026/2.692	2.019/2.743
1D-UNet	4.5 k	4.8 M	129	31	2.063/2.664	2.052/2.717
<b>ATS-UNet</b>	<b>4.5 k</b>	<b>4.8 M</b>	<b>131</b>	<b>32</b>	<b>2.032/2.710</b>	<b>2.032/2.749</b>
TFiLM (baseline) [9]	68,221.2 k	116,420 M	/	/	3.646/2.523	

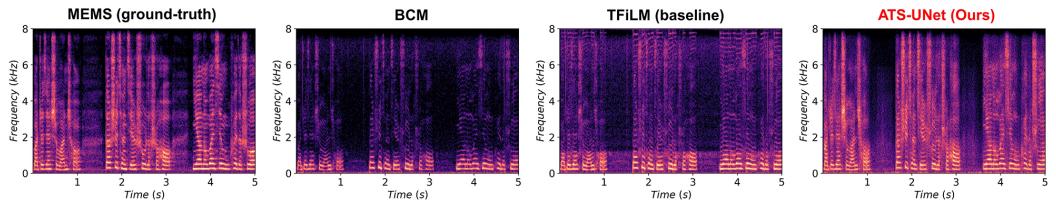
Although UNet v1 only has about 10 thousand parameters, it still requires a long inference time on a computation restricted platform, and thus we proposed Hybrid-UNet, Mixed-UNet, and ATS-UNet as described in Section 4. Benchmark latencies and accuracies are provided in Table 2 and Figure 7. ATS-UNet and 1D-UNet are the fastest networks, taking 131/32 ms and 129/31 ms, respectively, to inference a 2048-point frame.

Due to the lack of temporal modeling, the accuracy of 1D-UNet is significantly lower than ATS-UNet. ATSM effectively promotes temporal modeling while only adding negligible latency. Although Hybrid-UNet has 2000 more parameters than Mixed-UNet, the two settings achieve nearly the same latency and accuracy because their floating-point operations (FLOPs) are very close. 2D-UNet v1 is the slowest with expensive computation and slightly higher accuracy. Note that 2D-UNet v2 is too large to be run on our embedded system, and thus we leave gaps in the table.

As shown in Figure 7, ATS-UNet is the most cost-efficient model as it is on the upper left of the plot. In addition, spectrum examples in Figure 8 demonstrate that ATS-UNet outperformed TFiLM since it recovered a more accurate high-frequency structure. Considering our embedded system's restricted computational resources and memory, ATS-UNet was the best architecture to enable on-chip audio super resolution for BCM.



**Figure 7.** (a) The trade-off between on-chip latency and audio super resolution performance measured by LSD. The number of parameters is represented by circle size. (b) Without ATSM, there is a linear relationship between latency and LSD. ATS-UNet is on the upper left of the dotted line, thereby, proving its superiority.



**Figure 8.** Audio super-resolution results visualized by spectrograms.

### 6.5. Power Consumption

Since the algorithm pipeline can be run in real-time on Arm Cortex-M7, we measured the power consumption of this microcontroller under two circumstances: (1) during audio super resolution and (2) without audio super resolution. We used a power analyzer (EMK850) to measure the average current within one minute. When our super-resolution module was active, the average consumption was 498 mW (114.4 mA at 4.35 V). After the audio super-resolution module was deactivated, the microcontroller consumed 406mW (93.0 mA at 4.37 V). Therefore, ATS-UNet, feature extraction and reconstruction consumed 92 mW (498 mW–406 mW) on average.

## 7. Perceptual Speech Quality Evaluation

In this section, we present the perceptual speech quality evaluation of our method under both quiet and noisy environments. We describe the user study design, participants, and results in this section. Specifically, we conducted two user studies. The first was to compare the perceived speech audio quality of different machine-learning models.

The second user study was to evaluate our method's effectiveness against environmental noises. We utilized a within subject user study design. We utilized the Friedman test for non-parametric statistical analysis ( $p < 0.05$ ) and the Wilcoxon signed-rank test for post hoc analysis ( $p < 0.05$ ). We utilized the Mann–Whitney U test to evaluate the difference between user groups for statistical analysis ( $p < 0.05$ ).

### 7.1. Participants

We recruited 20 participants (14 males and 6 females) with an average age of 33.2 (s.d. = 4.8) separated into two groups. The “Golden Ear” (GE) group had 10 participants (6 males and 4 females) with an average age of 34.2 (s.d. = 5.0). They were specialists who were selected and trained to be able to discern subtle differences in audios. The “Non-Golden Ear” (NGE) group had the other 10 amateur listeners (6 males and 4 females) with an average age of 32.1 (s.d. = 4.6).

The study was conducted in a quiet listening room. During the test, each participant was required to wear headphones (AKG N20 model). A 5 min break was required after

10 trials. The whole study lasted for 60 min. Each participant received a 30 USD gift card for their time and effort.

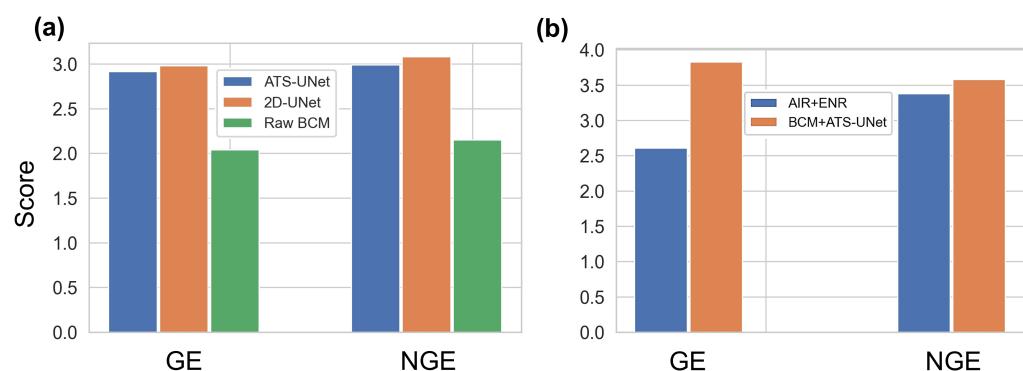
### 7.2. User Study 1: Speech Audio Quality in Quiet Environment

This user study included 20 trials. Participants listened to an audio clip from the MEMS air-conduction microphone in each trial, which produced the highest speech audio quality. Then, they listened to and compared three audio clips, including: (1) original speech audio from the BCM (Original); (2) speech audio processed by the 2D-UNet v1; and (3) speech audio processed by the ATS-UNet.

The three audio clips had the same duration, while each set of audio clips lasted between 5 and 10 s with an average duration of 8.2 s. Then, each participant rated the sound quality of these three audio clips by referring to the high-quality audio clip from the MEMS microphone. We utilized a 5-point Likert scale for the rating (5 = very good, 3 = neutral, and 1 = very bad). The three audio clips were ordered randomly in each trial before the user study. Participants were allowed to listen to and compare audio clips repeatedly. In total, each participant listened to 80 audio clips.

### Results

The results show that both 2D-UNet v1 and ATS-UNet can effectively increase the sound quality of audio from the bone-conduction microphone. Further, 2D-UNet v1 achieved better performance compared with ATS-UNet. As shown in Figure 9a, the mean score of the original audio was 2.09 (s.d. = 0.03), of the ATS-UNet audio was 2.95 (s.d. = 0.03), and of the 2D-UNet v1 audio was 3.03 (s.d. = 0.03). These differences were statistically significant according to a Friedman test ( $\chi^2(2, N = 400) = 376.6, p < 0.001$ ). Post-hoc analysis showed that both the perceived sound quality of audio processed by the ATS-UNet ( $Z = -13.6, p < 0.001$ ) and the 2D-UNet v1 ( $Z = -13.9, p < 0.001$ ) significantly outperformed the original bone-conduction speech audio. Further, 2D-UNet v1 outperformed ATS-UNet ( $Z = -2.8, p < 0.01$ ) significantly.



**Figure 9. (a,b)** The qualitative speech audio quality evaluation results.

The user group analysis results show that there was significant effect of golden ear status on the perceived sound quality when listening to the original speech audio from the BCM ( $Z = -2.1, p = 0.036$ ) but not under the 2D-UNet v1 ( $Z = -0.9, p = 0.37$ ) or ATS-UNet ( $Z = -1.2, p = 0.23$ ) conditions.

### 7.3. User Study 2: Effectiveness of ATS-UNet Against Environmental Noises

This user study is to evaluate the effectiveness of our method against environmental noises. We compared our method (BCM + ATS-UNet) to a baseline method—a single air-conduction microphone with environment noise reduction (AIR+ENR). In this evaluation, we adopted Baseus Encok TWS earbuds—WM01 (<http://www.baseus.com/product-740?lang=en-us> (accessed on 1 November 2022)) as our comparison baseline. It has a built-in signal processing method for environmental noise reduction. During our test with five

different brands of earphones with the single speech microphone solution, the Baseus WM01 earbud achieved the best performance in environmental noise reduction.

### 7.3.1. Speech Audio Data Collection in Noisy Environments

We first recruited five participants to collect speech audios under various environments, including a noisy pedestrian street, a subway station, and a moving car with the window open. They had an average age of 26.5 (s.d. = 2.5). We used the recording hardware presented in Section 5.1 to collect the speech audios from both the air-conduction microphone and the bone-conduction microphone. Further, we streamed the speech audio from the Baseus WM01 earbud to an iPhone 12 for comparison.

We used three hand-clapping events to start each recording and later synchronize the audios. During each data collection session, each participant read the same article that lasts around 2 min. The whole data collection procedure lasted 40 min. Each participant received a 20 USD gift card. As a result, we collected  $2 \text{ (min)} \times 5 \text{ (participants)} \times 3 \text{ (environments)} = 30 \text{ min}$  of speech audios with the bone-conduction microphone, the MEMS air-conduction microphone, and the Baseus WM01 earbud. Among these audio recordings, the raw audios from the MEMS microphone contained stronger environmental noises.

### 7.3.2. Speech Audio Quality Evaluation

This user study included 24 trials. In each trial, participants listened to an audio clip from the MEMS microphone with stronger background noises. Then, they listened to and compared two audio clips, including (1) speech audio from the bone-conduction microphone processed by the ATS-UNet and (2) speech audio from the Baseus WM01 earbud with noise-reduction algorithms. Each pair of audio clips had the same duration and lasted between 8 and 13 s, with an average duration of 10.96 s.

Then, each participant rated the sound quality of these two audio clips by referring to the audio clip from the MEMS microphone. We utilized a 5-point Likert scale for the rating (5 = very good, 3 = neutral, and 1 = very bad). The two audio clips were ordered randomly in each trial before the user study. Participants were allowed to listen to and compare the audio clips repeatedly. In total, each participant listened to 60 audio clips.

### 7.3.3. Results

The results show both golden ear (GE) ( $Z = -12.2, p < 0.001$ ) and non-golden ear (NGE) ( $Z = -2.2, p = 0.02$ ) raters considered the speech audio quality of the BCM + ATS-UNet outperformed the baseline method—AIR + ENR. As shown in Figure 9b, GE raters scored the speech audio quality of BCM + ATS-UNet with an average of 3.83 (s.d. = 0.86) and of the AIR + ENR with an average of 2.61 (s.d. = 0.93). NGE raters scored the speech audio quality of BCM + ATS-UNet and AIR+ENR with average values of 3.58 (s.d. = 1.17) and 3.38 (s.d. = 1.21), respectively.

User group analysis results show that there was a significant effect of the user group on the perceived sound quality of AIR + ENR ( $Z = -7.7, p < 0.001$ ) but not BCM + ATS-UNet ( $Z = -1.9, p = 0.053$ ). These results indicate that GE and NGE raters perceived similar speech audio quality regarding our method. However, GE raters gave the AIR+ENR a significantly lower score, indicating a poorer preference for AIR + ENR.

## 8. Discussion

In this work, we present the first on-chip audio super-resolution system for BCM. By integrating a novel ATSM into UNet architecture, ATS-UNet makes it possible to recover the missing high-frequency content captured by the BCM on resource-constrained wearable devices. Therefore, model inferences could be run locally on wearable devices without unwanted data transmission and lower latency. In this section, we discuss potential future works and related applications.

### 8.1. Dual Microphone System and Ambient Awareness

Even though BCM is superior to traditional microphones in noisy environments, and our system significantly improved the BCM's audio quality, air-conduction microphones still provide higher speech quality in low noisy environments. Therefore, a great deal of research [21–25] has focused on using an air-conduction microphone as the primary sensor, accompanied by a BCM for noise reduction. Conversely, low-quality bone-conduction audio is used directly in this research, and thus we hypothesize that there may be an opportunity to apply the audio super-resolution model on bone-conduction speech in conjunction with multi-microphone denoising algorithms.

BCMs and air-conduction microphones are suitable for different scenarios due to their hardware properties. For example, under strong wind noise, BCMs are highly desired, whereas, in a quiet meeting room, BCMs are unnecessary. In this case, the audio super-resolution algorithm leads to unnecessary power consumption. Therefore, another potential future research with a dual-microphone system could be ambient awareness. We anticipate that a dual-microphone system with ambient awareness could achieve the best user experience with optimal power consumption. With the ambient environment information, we could then determine an appropriate microphone and algorithm combination to be utilized at any instance.

### 8.2. Audio Super Resolution Applications

In this work, our system incorporated a single BCM, which we modeled as an audio super-resolution problem. We have also observed many other potential real-world applications. For example, recently, many people are wearing masks to prevent COVID-19. While these masks prevent the spread of the virus, it also blocks part of the speech signals. Corey et al. [51] showed different masks and microphone placements have different impacts on speech quality.

We believe the audio super-resolution model is a potential solution for recovering the attenuated frequency components from the masks. Increasingly, people pursue high-fidelity music; however, for now, the majority of music on the internet is compressed MP3 files. We anticipate that our model could be used to recover compression losses generated from lossy compression audio codecs. In general, it is encouraged to use our ATS-UNet if audio quality is degraded by frequency loss.

### 8.3. ATSM for Other Audio Applications

ATSM was designed for processing spectrograms of the audio signal, one of the most widely used input features for audio-related deep neural networks. Therefore, we believe ATSM can be easily adapt to other audio applications, such as speech separation [52] and speech emotion recognition [53]. Researchers can insert ATSM into existing models and reduce the dimension of convolutional layers making the models more lightweight and deployable. Though ATSM was not designed for input features, such as waveforms and MFCC, this provides insight on how to enable information flow and enlarge perceptual range without large convolutional kernels.

### 8.4. Limitations and Future Work

In this paper, we built a hardware prototype to evaluate the effectiveness of our method to recover high-fidelity speech audio from the bone-conduction microphone. The data collection and performance evaluation procedures were performed on the development board, as Figure 1 shows. We did not develop a wearable hardware solution designed for users to wear it comfortably. Further, during our test and evaluation, the placement of the BCM had a significant effect on the audio quality.

In our implementation, we utilized an earmuff to stabilize the BCM to the user's skin with its location shown in Figure 6. We chose this location for two reasons. (1) It can pick good quality of speech audios during our pilot study (In the pilot study, we compared two locations: in front of the ear and behind the ear). (2) We referred to the

cutting-edge design of modern bone-conduction speakers. We expect future work to investigate the optimized location and mounting mechanism. Further, we expect future work to explore sensor fusion approaches to enable better speech audio quality using air- and bone-conduction microphones.

## 9. Conclusions

In this paper, we proposed a novel real-time embedded audio super-resolution-based speech-capture system with BCM. By integrating a novel ATSM into UNet architecture, ATS-UNet efficiently processed bone-conduction speech audio signals with minimal computational resources among our proposed lightweight audio super-resolution models. Compared with the baseline method (TFLM), ATS-UNet achieved higher performance in audio quality and reduced the number of parameters by approximately 100 times. Compared to the 2D-UNet v1, ATS-UNet reduced the number of FLOPs by 44% and achieved comparable performance.

With the reduction in computation complexity, our system can achieve real-time processing on a Cortex-M7 with an average power consumption of 92 mW. User studies demonstrated that our system significantly improved the perceptual quality of bone-conduction speech. We propose that our system will promote the usage of BCM in earphones and other deep-learning-based audio-processing applications, particularly those deployed in resource-constrained embedded systems.

**Author Contributions:** Conceptualization, Y.L., Y.W., X.L., Y.S., S.P., and S.-F.S.; methodology, Y.L., Y.W., X.L., and S.-F.S.; software, Y.L., Y.W., X.L., and S.-F.S.; validation, Y.L., Y.W., X.L., and S.-F.S.; formal analysis, Y.L., Y.W., X.L., Y.S., S.P., and S.-F.S.; investigation, Y.L., Y.W., X.L., Y.S., and S.-F.S.; resources, Y.L., Y.W., X.L., Y.S., S.P., and S.-F.S.; data curation, Y.L., Y.W., X.L., and S.-F.S.; writing—original draft preparation, Y.L., Y.W., X.L., Y.S., S.P., and S.-F.S.; writing—review and editing, Y.L., Y.W., X.L., Y.S., S.P., and S.-F.S.; visualization, Y.L., Y.W., and X.L.; supervision, Y.S., S.P., and S.-F.S.; project administration, Y.S. and S.P.; funding acquisition, Y.S., S.P., and S.-F.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Natural Science Foundation of China under Grant No. 62132010 and No. 62002198, Tsinghua University Initiative Scientific Research Program, Beijing Key Lab of Networked Multimedia, and the Institute for Guo Qiang, Tsinghua University.

**Institutional Review Board Statement:** The IRB were waived for this study since this study only involved recording users' voices in usual settings such as canteen, sidewalk, or subway.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** Data are available upon request.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Capon, J. High-resolution frequency-wavenumber spectrum analysis. *Proc. IEEE* **1969**, *57*, 1408–1418. [[CrossRef](#)]
2. Pratt, W. Generalized Wiener Filtering Computation Techniques. *IEEE Trans. Comput.* **1972**, *C-21*, 636–641. [[CrossRef](#)]
3. Boll, S. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech Signal Process.* **1979**, *27*, 113–120. [[CrossRef](#)]
4. Park, S.R.; Lee, J.W. A Fully Convolutional Neural Network for Speech Enhancement. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017. [[CrossRef](#)]
5. Macartney, C.; Weyde, T. Improved speech enhancement with the wave-u-net. *arXiv* **2018**, arXiv:1811.11307.
6. Shimamura, T.; Tamiya, T. A reconstruction filter for bone-conducted speech. In Proceedings of the 48th Midwest Symposium on Circuits and Systems, 2005, Covington, KY, USA, 7–10 August 2005. [[CrossRef](#)]
7. McBride, M.; Tran, P.; Letowski, T.; Patrick, R. The effect of bone conduction microphone locations on speech intelligibility and sound quality. *Appl. Ergon.* **2011**, *42*, 495–502. [[CrossRef](#)]
8. Kuleshov, V.; Enam, S.Z.; Ermon, S. Audio super-resolution using neural nets. In Proceedings of the ICLR (Workshop Track), Toulon, France, 24–26 April 2017.

9. Birnbaum, S.; Kuleshov, V.; Enam, Z.; Koh, P.W.W.; Ermon, S. Temporal FiLM: Capturing Long-Range Sequence Dependencies with Feature-Wise Modulations. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, USA, 8–14 December 2019; pp. 10287–10298.
10. Kim, S.; Sathe, V. Bandwidth extension on raw audio via generative adversarial networks. *arXiv* **2019**, arXiv:1903.09027.
11. Hao, X.; Xu, C.; Hou, N.; Xie, L.; Chng, E.S.; Li, H. Time-Domain Neural Network Approach for Speech Bandwidth Extension. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020. [CrossRef]
12. Kegler, M.; Beckmann, P.; Cernak, M. Deep Speech Inpainting of Time-Frequency Masks. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020. [CrossRef]
13. Mysore, G.J. Can we Automatically Transform Speech Recorded on Common Consumer Devices in Real-World Environments into Professional Production Quality Speech?—A Dataset, Insights, and Challenges. *IEEE Signal Process. Lett.* **2015**, *22*, 1006–1010. [CrossRef]
14. Su, J.; Jin, Z.; Finkelstein, A. HiFi-GAN: High-Fidelity Denoising and Dereverberation Based on Speech Deep Features in Adversarial Networks. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020. [CrossRef]
15. International Telecommunication Union. *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-To-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*; ITU-T Publications: Geneva, Switzerland, 2001.
16. Gray, A.; Markel, J. Distance measures for speech processing. *IEEE Trans. Acoust. Speech Signal Process.* **1976**, *24*, 380–391. [CrossRef]
17. Hidri, A.; Meddeb, S.; Amiri, H. About Multichannel Speech Signal Extraction and Separation Techniques. *J. Signal Inf. Process.* **2012**, *03*, 238–247. [CrossRef]
18. Ephraim, Y.; Malah, D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.* **1984**, *32*, 1109–1121. [CrossRef]
19. Scalart, P.; Filho, J. Speech enhancement based on a priori signal to noise estimation. In Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, Atlanta, GA, USA, 9 May 1996. [CrossRef]
20. Shin, H.S.; Kang, H.G.; Fingscheidt, T. Survey of speech enhancement supported by a bone conduction microphone. In Proceedings of the Speech Communication; 10. ITG Symposium, Braunschweig, Germany, 26–28 September 2012; VDE: Frankfurt, Germany, 2012; pp. 1–4.
21. Liu, Z.; Zhang, Z.; Acer, A.; Droppo, J.; Huang, X. Direct filtering for air- and bone-conductive microphones. In Proceedings of the IEEE sixth Workshop on Multimedia Signal Processing, Siena, Italy, 29 September–1 October 2004. [CrossRef]
22. Lee, C.H.; Rao, B.D.; Garudadri, H. Bone-Conduction Sensor Assisted Noise Estimation for Improved Speech Enhancement. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018. [CrossRef]
23. Takada, M.; Seki, S.; Toda, T. Self-Produced Speech Enhancement and Suppression Method using Air- and Body-Conductive Microphones. In Proceedings of the 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, HI, USA, 12–15 November 2018. [CrossRef]
24. Zhou, Y.; Chen, Y.; Ma, Y.; Liu, H. A Real-Time Dual-Microphone Speech Enhancement Algorithm Assisted by Bone Conduction Sensor. *Sensors* **2020**, *20*, 5050. [CrossRef] [PubMed]
25. Yu, C.; Hung, K.H.; Wang, S.S.; Tsao, Y.; Hung, J.-w. Time-Domain Multi-Modal Bone/Air Conducted Speech Enhancement. *IEEE Signal Process. Lett.* **2020**, *27*, 1035–1039. [CrossRef]
26. Shimamura, T.; Mamiya, J.; Tamiya, T. Improving Bone-Conducted Speech Quality via Neural Network. In Proceedings of the 2006 IEEE International Symposium on Signal Processing and Information Technology, Vancouver, BC, Canada, 27–30 August 2006. [CrossRef]
27. Rahman, M.S.; Shimamura, T. Intelligibility enhancement of bone conducted speech by an analysis-synthesis method. In Proceedings of the 2011 IEEE 54th International Midwest Symposium on Circuits and Systems (MWSCAS), Seoul, Republic of Korea, 7–10 August 2011. [CrossRef]
28. Bouzerhal, R.E.; Falk, T.H.; Voix, J. In-ear microphone speech quality enhancement via adaptive filtering and artificial bandwidth extension. *J. Acoust. Soc. Am.* **2017**, *141*, 1321–1331. [CrossRef] [PubMed]
29. Shan, D.; Zhang, X.; Zhang, C.; Li, L. A Novel Encoder-Decoder Model via NS-LSTM Used for Bone-Conducted Speech Enhancement. *IEEE Access* **2018**, *6*, 62638–62644. [CrossRef]
30. Liu, H.P.; Tsao, Y.; Fuh, C.S. Bone-conducted speech enhancement using deep denoising autoencoder. *Speech Commun.* **2018**, *104*, 106–112. [CrossRef]
31. Hussain, T.; Tsao, Y.; Siniscalchi, S.M.; Wang, J.C.; Wang, H.M.; Liao, W.H. Bone-Conducted Speech Enhancement Using Hierarchical Extreme Learning Machine. In *Lecture Notes in Electrical Engineering*; Springer: Gateway East, Singapore, 2021; pp. 153–162. [CrossRef]
32. Zheng, C.; Yang, J.; Zhang, X.; Sun, M.; Yao, K. Improving the Spectra Recovering of Bone-Conducted Speech via Structural SIMilarity Loss Function. In Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019. [CrossRef]
33. Li, K.; Lee, C.H. A deep neural network approach to speech bandwidth expansion. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015. [CrossRef]

34. Eskimez, S.E.; Koishida, K.; Duan, Z. Adversarial Training for Speech Super-Resolution. *IEEE J. Sel. Top. Signal Process.* **2019**, *13*, 347–358. [[CrossRef](#)]
35. Lim, T.Y.; Yeh, R.A.; Xu, Y.; Do, M.N.; Hasegawa-Johnson, M. Time-Frequency Networks for Audio Super-Resolution. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018. [[CrossRef](#)]
36. Feng, B.; Jin, Z.; Su, J.; Finkelstein, A. Learning Bandwidth Expansion Using Perceptually-motivated Loss. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019. [[CrossRef](#)]
37. Wang, M.; Wu, Z.; Kang, S.; Wu, X.; Jia, J.; Su, D.; Yu, D.; Meng, H. Speech Super-Resolution Using Parallel WaveNet. In Proceedings of the 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), Taipei, Taiwan, 26–29 November 2018. [[CrossRef](#)]
38. Gupta, A.; Shillingford, B.; Assael, Y.; Walters, T.C. Speech Bandwidth Extension with Wavenet. In Proceedings of the 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 20–23 October 2019. [[CrossRef](#)]
39. Li, Y.; Tagliasacchi, M.; Rybakov, O.; Ungureanu, V.; Roblek, D. Real-Time Speech Frequency Bandwidth Extension. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021. [[CrossRef](#)]
40. Li, X.; Chebiyyam, V.; Kirchhoff, K. Speech Audio Super-Resolution for Speech Recognition. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019. [[CrossRef](#)]
41. Kumar, R.; Kumar, K.; Anand, V.; Bengio, Y.; Courville, A. NU-GAN: High resolution neural upsampling with GAN. *arXiv* **2020**, arXiv:2010.11362.
42. Liu, X.; Li, Y.; Fromm, J.; Wang, Y.; Jiang, Z.; Mariakakis, A.; Patel, S. SplitSR: An end-to-end approach to super-resolution on mobile devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2021**, *5*, 25. [[CrossRef](#)]
43. Lee, R.; Venieris, S.I.; Dudziak, L.; Bhattacharya, S.; Lane, N.D. MobiSR: Efficient on-device super-resolution through heterogeneous mobile processors. In Proceedings of the The 25th Annual International Conference on Mobile Computing and Networking, Los Cabos, Mexico, 21–25 October 2019. [[CrossRef](#)]
44. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. Tensorflow: A system for large-scale machine learning. In Proceedings of the 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), Savannah, GA, USA, 2–4 November 2016; pp. 265–283.
45. Lai, L.; Suda, N.; Chandra, V. Cmsis-nn: Efficient neural network kernels for arm cortex-m cpus. *arXiv* **2018**, arXiv:1801.06601.
46. Lin, J.; Gan, C.; Han, S. TSM: Temporal Shift Module for Efficient Video Understanding. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019. [[CrossRef](#)]
47. Abbas, S.; Mosbah, M.; Zemmari, A. ITU-T Recommendation G. 114, “One way transmission time”. In Proceedings of the International Conference on Dynamics in Logistics 2007 (LDIC 2007), Bremen, Germany, August 2007; Lecture Notes in Computer Sciences; Citeseer: Princeton, NJ, USA, 1996. Available online: <https://www.itu.int/rec/T-REC-G.114-200305-I/en> (accessed on 1 November 2022)
48. Allen, J.; Rabiner, L. A unified approach to short-time Fourier analysis and synthesis. *Proc. IEEE* **1977**, *65*, 1558–1564. [[CrossRef](#)]
49. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
50. Volkmann, J.; Stevens, S.S.; Newman, E.B. A Scale for the Measurement of the Psychological Magnitude Pitch. *J. Acoust. Soc. Am.* **1937**, *8*, 208–208. [[CrossRef](#)]
51. Corey, R.M.; Jones, U.; Singer, A.C. Acoustic effects of medical, cloth, and transparent face masks on speech signals. *J. Acoust. Soc. Am.* **2020**, *148*, 2371–2375. [[CrossRef](#)] [[PubMed](#)]
52. Ochiai, T.; Delcroix, M.; Kinoshita, K.; Ogawa, A.; Nakatani, T. A Unified Framework for Neural Speech Separation and Extraction. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019. [[CrossRef](#)]
53. Drakopoulos, G.; Pikramenos, G.; Spyrou, E.; Perantonis, S. Emotion Recognition from Speech: A Survey. In Proceedings of the 15th International Conference on Web Information Systems and Technologies, Vienna, Austria, 18–20 September 2019; SCITEPRESS—Science and Technology Publications: Setubal, Portugal, 2019. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.