VoiceSeeker: Energy-Efficient and Accurate Audio–Visual Target Speaker Separation on Mobile Devices

Juheon Yi, Kyungjin Lee, Hyunseok Oh, and Youngki Lee, *Seoul National University, Seoul, 08826, South Korea*

We present VoiceSeeker***, a mobile system that utilizes vision as complementary modality to separate the target speaker's speech from the input mixture audio. Despite recent advances in deep neural network (DNN)-based audio–visual target speaker separation, important system challenges remain unsolved: separation accuracy drop due to visual feature quality fluctuation, and high resource demand of composite-DNN execution on audio–visual inputs. We design an adaptive pipeline that flexibly switches between audio–visual, audio-only, and no separation depending on the audio–visual input content. For agile adaptation on audio–visual inputs whose content changes rapidly with low temporal correlation, we design a suite of lightweight probing techniques to analyze them on the fly and select the adequate separation path. We also develop a continuous audio-only separation pipeline capable of robustly and efficiently separating the target speaker.*** VoiceSeeker ***achieves up to 49% energy consumption reduction, 2.2 dB separation accuracy, and 41% latency gain over baselines.***

Imagine a scenario where an online streamer visits a tourist attraction and live-streams himself talking with a smartphone. A mobile system separating his voice from close-by interfering speakers will be useful. Recently, audio–visual target speaker separation (i.e., leveraging a visual input of the speaker's mouth movements as guidance to separate his voice from the input audio mixture) is getting increasing attention,[1] as it (i) achieves superior accuracy compared to the audio-only approach, (ii) overcomes the short operation range (e.g., 30 cm) of prior ultrasound-based approaches, and (iii) does not require a bulky mic array. However, the prior studies focus on improving the accuracy with deep neural network (DNN) designs, leaving the following challenges.

› *Visual Feature Quality Fluctuation:* Prior studies on audio–visual speaker separation assume ideal camera inputs captured in a recording studio.[2] Here, the target speaker looks straight at the camera, and the mouth movements are captured from the front. State-of-the-art (SOTA) models achieve high separation accuracy for such ideal settings. However, they suffer from severe performance drop in real-world settings. For instance, when the lip movements are not captured accurately (e.g., the target is looking sideways or the mouth is occluded), the audio–visual separation results in ≈10 dB worse quality compared to the ideal case, which is even worse than the unprocessed raw input.

› *Composite-DNN Execution on Audio–Visual Inputs:* It is highly challenging to execute the workload on resource-constrained mobile devices in real time. This is because i) the pipeline involves the execution of compute-intensive composite-DNNs on energy-hungry sensor inputs (i.e., camera and mic) (see Figure 1), and ii) the pipeline frequently runs on streaming input (e.g., visual feature extraction on every 30 fps camera frame, separation on every 0.5 s audio segments).[1] For instance, sequential execution of the SOTA pipeline incurs 4–5 W average power (>6 W peak) consumption (see Table 2). We present VoiceSeeker, a mobile system to support audio–visual target speaker separation in

Digital Object Identifier 10.1109/MPRV.2025.3550543
Date of publication 23 April 2025; date of current version 18 July 2025.

a highly robust and energy-efficient manner. As the core approach, we adopt *adaptive multipath activation*, which adaptively activates different branches of the composite audio–visual DNNs. In particular, it flexibly alternates among (i) *skip* separation (when the target is silent or dominant), (ii) *audio-only* separation (when visual features are not helpful due to varying face poses and mouth occlusion), and (iii) *audio–visual* separation (when both audio and visual features contribute to higher accuracy). This way, it constantly provides high separation accuracy while reducing resource and energy consumption. It is, however, nontrivial to realize the adaptive multipath activation. Notably, the adaptation should be triggered quickly in a fine-granule timescale (e.g., per 0.5 s separation window) due to the fast-changing and unpredictable nature of the audio–visual contents. In particular, speech signal content changes in the units of phoneme/words (e.g., 0.5–1 s) and face pose variations or mouth occlusion can occur abruptly in a short timescale (temporarily looking sideways). This likely makes the adaptation decision at the current separation window no longer optimal in the next window. Thus, it is infeasible to adopt the common *predictive adaptation* approach used in continuous mobile vision systems (e.g., in the work of Liu et al.[3]) where a selected logic is executed for an extended duration utilizing contextual continuity and temporal correlations in adjacent input data. We tackle the challenges with following techniques.

› *Lightweight Multistage Probing:* It is crucial to develop lightweight yet accurate probes to quickly analyze the fast-changing audio–visual inputs on the fly and activate different execution paths. In this light, we design lightweight multistage probing techniques based on two useful features: facial landmarks and low-frequency audio input. As the first probe, the facial landmarks are efficiently extracted as a part of the audio–visual pipeline and help make an early adaptation decision in two folds. First, we train a simple but effective personalized random forest classifier to predict the speaker's voice activity and skip separation when silent. Second, we quickly switch to the audio-only separation when the facial features indicate mouth occlusion or profile faces of the speaker. For the second probe, we leverage low-frequency input audio to estimate the dominance of the target speaker at a low cost. This gives a further opportunity for skipping separation when the target is already dominant.

› *Continuous Audio-Only Target Separation:* Opposed to the full audio–visual pipeline, we identify that audio-only separation cannot distinguish which branch of the separated outputs belongs to the target (i.e., *label permutation problem*). Conventional audio-only separation techniques adopt an additional DNN dedicated to speaker identification. This approach, however, incurs additional overhead to the pipeline and suffers from low accuracy when the input window length is short. To address the problem, we adopt a *lightweight target matching* technique that places partial overlap between consecutive separation windows and utilizes the commonality between them. We discover that target matching can be effectively done with only 200 ms (40% for 0.5 s window) overlap (in phoneme level).

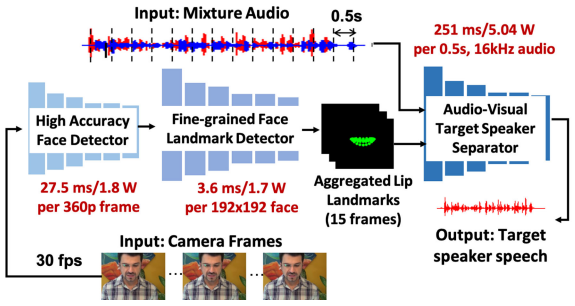

FIGURE 1. Audio–visual target speaker separation pipeline. Inference times and power consumption are measured on TensorFlow-Lite over Google Pixel 4 (Adreno 640 GPU).

BENEFITS OF AUDIO–VISUAL SEPARATION

Audio–visual target speaker separation pipeline consists of two stages, as shown in Figure 1: (i) visual feature extraction (face detection + landmark estimation), and (ii) audio–visual target separation. It has following merits compared to audio-only and recently studied ultrasound-aided speaker separation.

Visual Guidance: The speaker's lip movements give guidance for the separation process, which provides benefits compared to audio-only separation.[4] First, knowing the moment of mouth opening and closing helps determine which out of the multiple separated speeches belongs to the target. This is highly challenging for the audio-only separation due to the *label permutation problem*. Specifically, audio-only separation models are trained with *permutation-invariant training*[5] to maximize separation accuracy; it computes the training loss for all possible mappings between the ground truth inputs and the separated outputs, and takes the combination

TABLE 1. SISDR improvement on TCD-TIMIT three-speaker mixture for different visual qualities.

Audio-only	Frontal face	Profile face	Occluded mouth	Occluded mouth+jaw
10.52 dB	12.67 dB	9.47 dB	6.69 dB	-1.07 dB

that minimizes the loss. Second, analyzing the per-frame lip movements enables lip reading to infer the target speaker's speech.[6]

*Long-Range Operation:*Recent studies use ultrasound to sense the speaker's lip movements for speech enhancement,[7] but suffer from a short operating range (e.g., 30 cm) and is vulnerable to noise (e.g., hand gestures). Mobile cameras, however, can capture high-resolution video at a high frame rate (e.g., commodity wearable glasses can capture 1080p videos at 30 fps), thus being able to robustly capture the target's lip movements even from a distance.

Single Channel Mic Support: Recent smartphones utilize multimic array and beamforming techniques to filter sound from a specific direction (e.g., Galaxy S20 Zoom-in Mic). While such an approach achieves a similar goal of target speaker separation, it requires a bulky mic array, mainly because each pair of mics should be separated farther than half the wavelength of the speech (i.e., $\approx$ 5 cm for 4 kHz voice).

MOTIVATION AND CHALLENGE FOR ADAPTATION

Continuous, always-on audio–visual separation is suboptimal both in terms of resource and accuracy. We detail why adaptive switching between (i) audio–visual, (ii) audio-only, and (iii) skip separation is necessary.

Visual Feature Quality Fluctuation

Prior works mostly evaluated the audio–visual separation performance in ideal settings where the target is looking toward the camera, with his mouth is unoccluded. In real-world scenarios, visual feature quality fluctuates over time as the target speaker may intermittently look sideways, or his mouth may be occluded. Running audio–visual separation over low-quality visual features incurs a significant accuracy drop as the visual feature fails to provide accurate guidance. Table 1 gives the scale-invariant signal-to-distortion ratio (SISDR) improvement for various visual qualities for 500 samples of three-speaker mixtures generated from the TCD-TIMIT[2] dataset. When the face is profile or the mouth is occluded, SISDR improvement becomes lower than the audio-only separation. When both the mouth and the jaw are occluded, SISDR improvement drops to below 0 dB (worse than raw input). In such a case, evaluating the visual feature quality and switching between (i) audio–visual and (ii) audio-only separation can help achieve stable, high accuracy.

TABLE 2. Latency and energy consumption.

Task	Model	Latency	Power
Idle	-	-	0.8 W
Face detection	RetinaFace[8]	27.5 ms per 360p frame	3.2 W
Landmark detection	FaceMesh[9]	3.6 ms per 112×112 face	2.8 W
Speaker separation	Conv-TasNet[4]	251 ms per 0.5 s audio	4.4 W

High Resource Demand

Audio–visual target separation involves a challenging, underexplored composite DNN workload (see Figure 1). Specifically, the visual feature extraction (i.e., face and landmark detection) runs at a high frame rate (e.g., 30 fps or higher) to capture the speaker's lip movements at phoneme scale (e.g., 10–100 ms). Furthermore, separation runs continuously on the input audio stream. Table 2 gives that full execution of the SOTA audio–visual separation pipeline takes 717 ms to process a 500 ms input on Google Pixel 4 (Qualcomm Adreno 640 GPU). Also, running them all drains 4–5 W power on average (>6 W peak).

To mitigate the challenge, we identify the following adaptation opportunities. First, we can skip separation if the target is not speaking (e.g., taking turns in conversation). Second, when the target speech is already dominant (e.g., small background speech/noise), we can skip the separation as the expected accuracy improvement is marginal.

Why is Adaptation Challenging?

Rapid Content Change with Low Temporal Correlation: Continuous speech signals and associated visual features (e.g., head pose and mouth movements) exhibit rapid changes with low temporal correlation, requiring a fast and frequent adaptation in a fine-granule timescale. Figure 2 shows an example of two overlapping speaker utterances. Each speaker's ON–OFF cycle and magnitude vary rapidly in the phoneme-word scale depending on

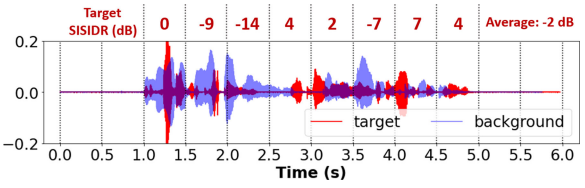

FIGURE 2. Example timeline of target speaker SISDR.

the speaking speed and accents within a sentence. As a result, the target SISDR values across 0.5 s separation windows highly fluctuate with low correlation (e.g., between -14 and 7 dB, where the average is -2 dB). Such fluctuation makes it nontrivial to keep the adaptation decision over time (e.g., skipping separation when the target was dominant in the previous separation window). In addition, visual feature quality can fluctuate abruptly in a short timescale due to the user's sudden actions.

Limitations of Prior Adaptation Techniques: Several continuous mobile vision systems (e.g., in the work of Liu et al.[3]) proposed a *predictive adaptation approach*. They utilize high temporal correlation of video (e.g., applying a high compression rate to the subarea where no objects of interest existed in the previous frame). However, it is nontrivial for audio–visual target speaker separation due to the aforementioned low temporal correlation in SISDR and abrupt visual feature quality changes.

VOICESEEKER SYSTEM OVERVIEW

We take the *reactive adaptation* approach, where we efficiently probe the current input data to determine the appropriate execution path, without any prediction based on the previous results. Figure 3 depicts VoiceSeeker's architecture. Given the audio–visual inputs, we extract two features: facial landmarks and low-frequency audio. From the extracted features, the *lightweight multistage probing* module runs a suite of probing techniques to analyze the current audio–visual inputs and determine the adequate path of the separation pipeline. We adaptively run the separation selected by the probing module to separate the target speaker; in case the audio-only path is selected, we employ a *continuous audio-only target separation* pipeline to overcome the limitations of audio-only separation in determining which branch is the target speaker.

LIGHTWEIGHT MULTISTAGE PROBING

Target Voice Activity Probe

The first probe aims to detect whether the target is silent so that the separation can be skipped. Prior works take the audio-based sound classification or speaker diarization approach[10] to detect voice activity. However, they assume only one active speaker at a time, and handling multiple overlapping speakers remains a challenge.[11] A few works aimed at audio-based target speaker detector in the presence of background speakers. However, they require a large training dataset of the target (e.g., 300k[12]) and show limited performance for unseen targets. Instead, we utilize facial landmarks of the target speaker (extracted as part of the audio–visual separation pipeline and can be leveraged without overhead) to design a lightweight yet accurate voice activity classifier.

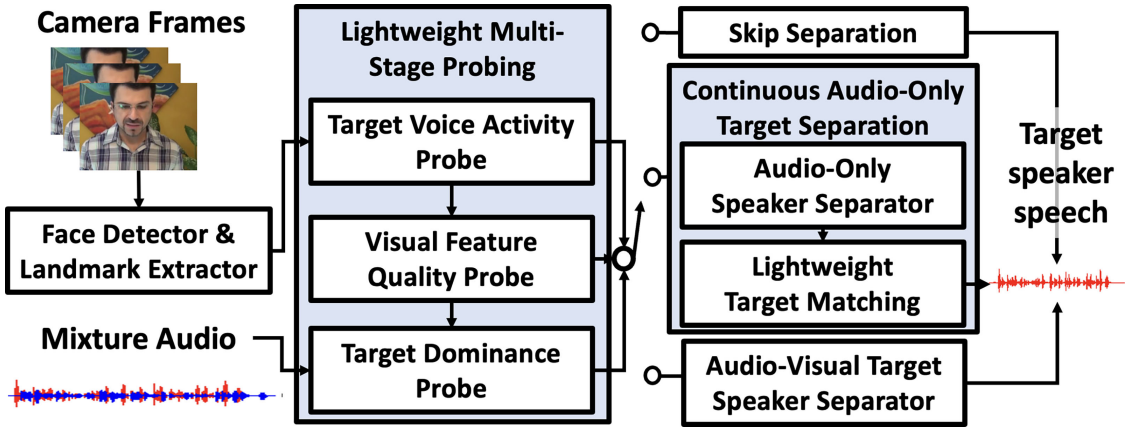

FIGURE 3. VoiceSeeker system architecture.

Landmark-Based Lightweight Voice Activity Classifier: We adopt two features that can be robustly extracted even under pose and occlusion change. i) *Inner lip height* is defined as the *y*-axis distance between the 62nd and 66th points out of the dlib 68 facial landmark model (which can be robustly calculated for both frontal and profile faces). ii) *Nose–chin distance* is defined as the 8th and 33rd points (which can be robustly calculated even when the mouth is occluded). We aggregate the two features across 0.5 s video (15 frames) and extract the difference values between consecutive frames, mean, minimum, and maximum values as final input features to a lightweight random forest classifier ($<$1 ms on Galaxy S20).

Personalization: Lip movements (e.g., lip heights when the target is speaking) vary across different speakers (e.g., some have clear pronunciation, while others may murmur). In case the video and audio samples of the target speaker are available (e.g., an online streamer can speak a few sentences in front of the camera in a quiet environment before the streaming session), we leverage them to improve accuracy. We observe that only 20–50 utterances are sufficient.

Visual Feature Quality Probe

When the target is speaking, we assess the visual feature quality to determine whether to run audio–visual or audio-only separation. We develop a lightweight visual feature quality probe to detect profile face and mouth occlusion. When either one of the two conditions is detected, we select audio-only separation.

Profile Face Detection: We analyze the face pose by analyzing the five landmarks (eyes, nose, and mouth), which are commonly extracted by SOTA face detectors (e.g., RetinaFace[8]). We take a similar approach to prior work;[13] for example, if nose is located at the right-hand side of the right eye and mouth, the face is in the right profile pose.

Mouth Occlusion Detection: Unlike pose, detecting mouth occlusion from facial landmarks is nontrivial, as the facial landmark detector infers the landmarks even when the mouth is occluded, resulting in wrong estimation. Thus, we run a lightweight Haar cascade classifier[14] to determine occlusion when it fails to detect the mouth from the frame.

Target Dominance Probe

Lastly, we opportunistically skip separation when the target is already dominant. However, predicting the target SISDR before running the separation is nontrivial due to the low temporal correlation of speech signals. We develop *low-frequency dominance probe*, which runs separation on the downsampled low-frequency audio input and predicts the target SISDR as

$$\mathrm{SISDR}_{\mathrm{pred}} = \mathrm{SISDR}(s_{\mathrm{in,down}}, s_{\mathrm{separated,down}}) \quad (1)$$

where $s_{\mathrm{in,down}}$ and $s_{\mathrm{separated,down}}$ are the downsampled input mixture and the separated output, respectively. Our key insight is that speech signals are composed of the fundamental frequency component around 100–300 Hz (determined by the vocal cord) and its harmonics. Thus, most of the energy is concentrated on low frequency (e.g., for speech samples in the TCDTIMIT dataset,[2] on average $\approx$ 81% of the energy is concentrated below 2 kHz), and it suffices to analyze the downsampled mixture to predict the target dominance.

CONTINUOUS AUDIO-ONLY TARGET SEPARATION

Unlike audio–visual separation, audio-only separation cannot tell which branch out of multiple separated outputs belongs to the target, [1] mainly due to label permutation problem. The problem was underexplored in prior works [4] as they focused on single-shot separation over a short 3–6 s sentence. However, the problem becomes more severe in our scenarios where we run audio-only separation on continuous speech and apply a short separation window (e.g., 0.5–1 s) for adaptation.

To tackle the challenge, we employ a *lightweight target matching* technique that places partial overlap between consecutive processing windows and utilizes the commonality between them to continuously track which separated output branch belongs to the target. Our key insight is that phoneme-level speech signal overlap (e.g., $\approx$ 0.1s) is sufficient for tracking the target, as it is practically very rare that multiple speakers simultaneously speak the same phoneme with same pitch, frequency, etc. We validate that 200 ms (40% overlap for 0.5 s input) achieves accurate matching.

The specific operation is as follows. When we run audio–visual separation (inference i), we save the extracted target speaker speech in a buffer. At the next inference, when we run audio-only separation (inference $i+1$), we take the input segment to partially overlap with that of the previous inference. After separation, we choose the output branch that

TABLE 3. Overall accuracy and resource usage.

	Target dominant scenario (TCD-TIMIT)				Visual Quality fluctuation scenario (LRS3-TED)			
	Two speakers		Three speakers		Two speakers		Three speakers	
	Full-AV	VoiceSeeker	**Full-AV**	VoiceSeeker	**Full-AV**	VoiceSeeker	**Full-AV**	VoiceSeeker
SISDR	15.98 dB	14.41 dB	14.78 dB	12.83 dB	9.71 dB	11.16 dB	3.54 dB	6.44 dB
PESQ	3.02	2.92	2.98	2.83	2.71	2.86	2.16	2.41
# separations (AV, AO, dominance probe)	34 (34,0,0)	19.4 (13.3, 0,24.6)	34.5 (34.5, 0, 0)	23 (16.6, 0,25.4)	39.1 (39.1,0,0)	26.2 (14.9, 7.2, 16.1)	39 (39, 0, 0)	32.4 (19.9, 8.7,15.2)

has the highest SISDR with the buffered target speaker speech

$$s_{\text{target}} = \text{argmax}_i \text{SISDR}(s_{\text{separated},i}, s_{\text{buffer}}) \quad (2)$$

where $s_{\text{separated},i}$ is the ith separated branch and s_{buffer} is the buffered previous separated target speech. The matched output s_{target} is updated in the buffer, and the process repeats for subsequent audio-only separation.

EVALUATION

Setup

Device and DNNs

We implemented VoiceSeeker on two commodity smartphones: Samsung Galaxy S20 (Qualcomm Snapdragon 865, Adreno 650 GPU) and Google Pixel 4 (Snapdragon 855, Adreno 640 GPU). We observe similar results and report performance obtained on S20 unless specified otherwise. We implement and train the DNNs listed in Table 2 using TensorFlow and convert them to TensorFlow-Lite. We use Monsoon Power Monitor for power measurement.

Datasets

› *TCD-TIMIT*[2] is composed of 62 speakers (59 verbal speakers and three lipspeakers) reading a total of 69 k phonetically rich English sentences. It provides 1080p@30 fps video recordings from two angles: frontal and 30° profile. We use 59 verbal speakers: 49 for training and 10 for testing. As each utterance is only 3–5 s, we concatenate four utterances from the same speaker (≈ 15–20 s in total, with ≈50% ON–OFF cycle) and add background speakers and noises for testset generation. We use this dataset for controlled evaluation on two scenarios. (i) *Dominant target.* The magnitudes of the background speakers and noise are scaled to set the target's input SISDR as 5, 5, 15, and 15 dB for the four utterance samples. The face poses remain frontal without any occlusion. (ii) *Visual feature quality fluctuation.* The face pose changes to profile for the second utterance. The mouth is occluded for the third utterance, with a 5-dB target input SISDR.

› *LRS3-TED*[15] is collected from TED videos. It contains complex English sentences (vocabulary size of 51 k) with diverse accents, along with 224×224, 25 fps face-cropped videos. We use this dataset to evaluate VoiceSeeker's robustness in in-the-wild scenarios. We use the *trainval* dataset (4004 speakers, 32 k utterances, 30 h in total) for separation model training. As the test utterances are also only 2–3 s long, we search and collect five videos (8–22 min) from YouTube with "English speech" keyword with diverse audio–visual contents (accents, face pose, and mouth occlusion). We split them into 20 s chunks for testset mixture generation similar to TCD-TIMIT.

Baselines

› *Full-AV* always run audio–visual separation regardless of the input.

› *Papez*[16] is the SOTA DNN for lightweight speaker separation. It utilizes *adaptive token pruning*, which adaptively adjusts the number of processing iterations depending on input difficulty. We only compare desktop latency and accuracy due to lack of GPU support for Transformer layer ops in TensorFlow-Lite/PyTorch Mobile.

Separation Accuracy and Resource Usage

Table 3 gives the overall separation accuracy and the number of separations for the two scenarios. For number of separations, we count dominance probe as 1/4 as it operates on 4× downsampled 4 kHz input. First, for the *dominant target* scenario, VoiceSeeker efficiently skips separation when the target is silent or already dominant, reducing the

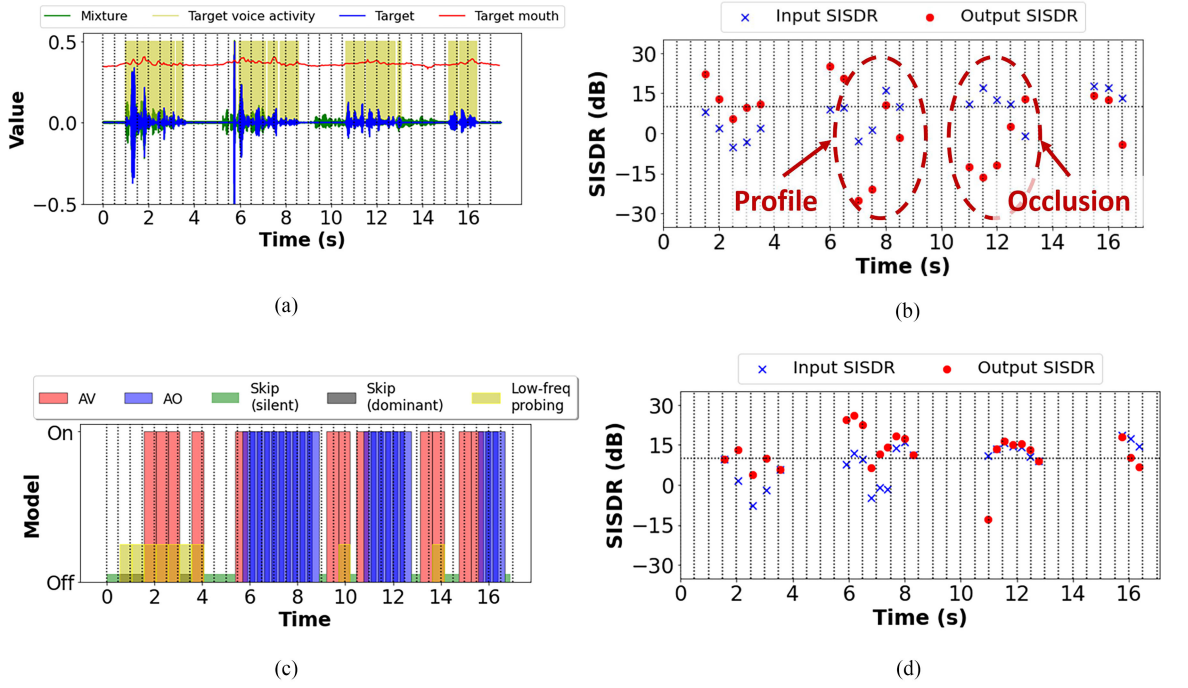

FIGURE 4. Example operation of VoiceSeeker in visual feature quality fluctuation scenario. (a) Input mixture. (b) Full-AV separation accuracy. (c) VoiceSeeker inference timeline. (d) VoiceSeeker separation accuracy.

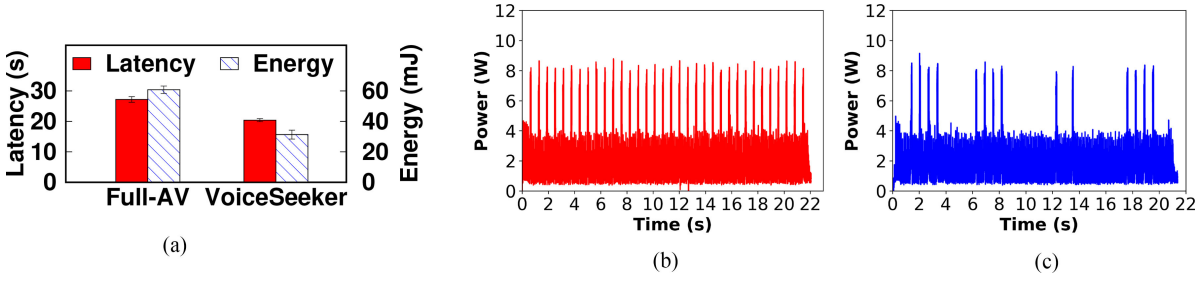

FIGURE 5. Power consumption timeline comparison. (a) Overall latency and energy. (b) Full-AV timeline (59.5 mJ). (c) VoiceSeeker timeline (32.6 mJ).

number of separations by 43% and 33% for two-speaker and three-speaker mixtures, respectively, with a minimum accuracy drop.

Next, we evaluate VoiceSeeker for the *visual quality fluctuation* scenario using LRS3-trained models and in-the-wild TED videos (we observe similar results for TCD-TIMIT dataset). Overall, we observe consistent performance gain: VoiceSeeker achieves 1.45 and 2.9 dB SISDR gain compared to Full-AV for two-speaker and three-speaker mixtures, respectively, while achieving 33% and 17% less separations.

Figure 4(a)–(d) shows an example inference timeline: while Full-AV fails in the 6–9 s and 11–13 s intervals due to profile face and mouth occlusion, VoiceSeeker efficiently switches to audio-only separation to recover accuracy.

Latency and Energy Consumption

Figure 5(a) shows the end-to-end latency and energy consumption of *Full-AV* and VoiceSeeker on Google Pixel 4 for *Dominant Target* scenario 2 speaker mixture: we observe a 25% latency reduction (enabling real-time on-device processing) and 49% energy

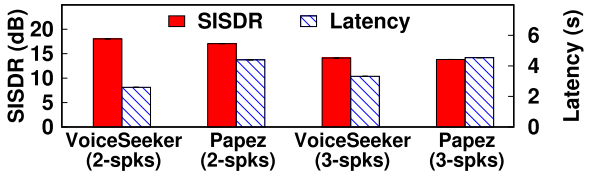

FIGURE 6. Latency and accuracy performance comparison with SOTA Papez..[16]

reduction. Figure 5(b) and (c) shows an example power consumption timelines: VoiceSeeker efficiently skips separations and removes power peaks.

Comparison With SOTA Model

Figure 6 compares the latency and accuracy performance of VoiceSeeker with Papez[16] for the *dominant target* scenario (we observe similar results for the *visual feature quality fluctuation* scenario as well). VoiceSeeker achieves similar SISDR with lower processing latency than Papez (e.g., 41% and 27% less for two and three speakers, respectively), showing the effectiveness of our adaptation pipeline. This is mainly because we run probe separation on 4 kHz input and opportunistically skip separation on full 16 kHz input, whereas Papez always runs on 16 kHz input.

Probing Overhead

Lastly, we observe that the probing overhead is very small. For each 0.5 s audio and video inputs, the base DNN inference pipeline takes in total 129 ms (face detection 28 ms, landmark estimation 54 ms, separation 47 ms). For lightweight multistage probing, the target voice activity (7 ms) and visual quality (8 ms) probes are pipelined with the face detection and landmark estimation process, effectively adding no latency overhead to the pipeline. The target dominance probe on 4 kHz downsampled audio input incurs 18 ms latency, keeping the overhead small (14%) even when the target is not dominant and audio–visual target separation runs on every input.

LIMITATIONS AND FUTURE WORKS

We identify the following limitations of VoiceSeeker and plan to address them in future work. (i) *Separation model failure.* As our main goal is the adaptation between multiple separation paths, the performance of VoiceSeeker heavily depends on the performance of the individual separation models. Especially, we observe cases where the audio-only separation model occasionally fails for a three-speaker mixture; in such a case, the target matching process fails to correctly identify the target branch, leading to consecutive failures in the future separation window. We plan to employ further improved backbones (e.g., transformer-based) to address the issue. (ii) *Separation quality fluctuation.* As VoiceSeeker dynamically switches between audio–visual, audio-only, and no separation, the quality of the separated output may fluctuate over time (e.g., due to an accuracy gap in audio–visual and audio-only separation). We plan to investigate the issue in future work to enable smooth quality in continuous separation.

CONCLUSION

We presented VoiceSeeker, a mobile audio–visual target speaker separation system. We solved the challenges of conventional pipelines that are agnostic to mobile device resource constraints and diverse real-world camera capture settings by (i) designing lightweight probing techniques to select the adequate pipeline in a fine-granule manner, and (ii) designing an end-to-end continuous, adaptive audio–visual target speaker separation pipeline that can efficiently run on mobile devices. Evaluations showed that VoiceSeeker achieves up to 49% energy consumption reduction, 2.2 dB separation accuracy improvement, and 41% latency gain over baselines.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF), Korean Government (MSIT) under Grant 2022R1A2C3008495 and Grant RS-2024-00463802.

REFERENCES

1. A. Ephrat et al., "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. Graph.*, vol. 37, 2018, Art. no. 112.
2. N. Harte and E. Gillen, "TCD-timit: An audio-visual corpus of continuous speech," *IEEE Trans. Multimedia*, vol. 17, no. 5, pp. 603–615, May 2015.
3. L. Liu et al., "Edge assisted real-time object detection for mobile augmented reality," in *Proc. ACM MobiCom*, 2019, pp. 1–16.

4. Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.
5. M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.
6. J. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6447–6456.
7. K. Sun and X. Zhang, "UltraSE: Single-channel speech enhancement using ultrasound," in *Proc. ACM MobiCom*, 2021, pp. 160–173.
8. J. Deng et al., "RetinaFace: Single-stage dense face localisation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020.
9. Y. Kartynnik et al., "Real-time facial surface geometry from monocular video on mobile GPUS," 2019, *arXiv:1907.06724*.
10. Y. Lee et al., "SocioPhone: Everyday face-to-face interaction monitoring platform using multi-phone sensor fusion," in *Proc. ACM MobiSys*, 2013, pp. 375–388.
11. H. Bredin and A. Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," in *Proc. Interspeech*, 2021, *arXiv:2104.04045*.
12. S. Ding et al., "Personal VAD: Speaker-conditioned voice activity detection," 2019, *arXiv:1908.04284*.
13. J. Yi, S. Choi, and Y. Lee, "EagleEye: Wearable camera-based person identification in crowded urban spaces," in *Proc. ACM MobiCom*, 2020, pp. 1–14.
14. P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2001, pp. I–I.
15. T. Afouras et al., "LRS3-TED: A large-scale dataset for visual speech recognition," 2018, *arXiv:1809.00496*.
16. H. Oh, J. Yi, and Y. Lee, "Papez: Resource-efficient speech separation with auditory working memory," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.

JUHEON YI received his Ph.D. degree in CS from Seoul National University, Seoul, 08826, South Korea. His research interests include edge AI systems and video analytics. Contact him at johnyi0606@snu.ac.kr.

KYUNGJIN LEE is currently working toward the Ph.D. degree in computer science and engineering with Seoul National University (SNU), Seoul, 08826, South Korea. Her research focuses on building 3D networked graphics systems at the intersection of 3D computer graphics, immersive video streaming, and mobile systems. Lee received her master's degree in electrical engineering from SNU. Contact her at jin11542@snu.ac.kr.

HYUNSEOK OH is currently working toward the Ph.D. degree with the Department of Computer Science and Engineering, Seoul National University, Seoul, 08826, South Korea. His research interests include human-centered AI and privacy-preserving machine learning. Oh received his B.S. degree in computer science and engineering from Seoul National University. Contact him at ohsai@snu.ac.kr.

YOUNGKI LEE is an associate professor at the Computer Science and Engineering Department, Seoul National University, Seoul, 08826, South Korea. He has been building experimental and innovative software systems, which involve multidimensional considerations across systems, applications, and users. His research interests include mobile and pervasive computing, applied machine learning, and human–computer interaction. He is the corresponding author of this article. Contact him at youngkilee@snu.ac.kr.