May 29, 2019

# Audio-Recording Techniques Using Machine Learning (ML)

Snehitha Singaraju

Moonseok Kim



# Audio-Recording Techniques Using Machine Learning (ML)

**Abstract:**

This publication describes audio-recording techniques, which allow a user equipment (UE) to determine a user's intent and apply audio beamforming and wind-noise filtering based on the determined intent. This provides the user with an enhanced user experience (UX) by enabling the user to focus on the moment and use the UE to record or capture the acoustic signals that best-represent the auditory scene. Among other aspects, the invention uses machine learning (ML) to analyze the auditory scene and the environment of the user.

**Keywords:**

Mobile device, user equipment (UE), smartphone, sound receiver, acoustic receiver, audio receiver, acoustic sensor, audio sensor, machine learning (ML), noise, wind noise, background noise, recurrent neural network (RNN), neural network, dense neural network, deep learning, convolution neural network, artificial intelligence (AI), design of experiment (DOE).

**Background:**

User equipment (UE), such as smartphones, conferencing devices, notebooks, computers, headphones, wireless headphones, microphones, wireless microphones, digital cameras, hearing aids, digital audio-recorders, and the like, use various techniques to capture the intended auditory scene composed by various acoustic signals.

In some respects, UE developers try to imitate a person's brain activity responsible for processing the information of the auditory scene. The human brain is skilled in determining the sources of the various acoustic signals in the auditory scene. For example, in a 360-degree auditory scene, the person can determine whether the sound is coming from the front, back, left, or right, even if the person is blind-folded. More impressively, the person can determine whether the acoustic signal source is near, but has a weak original strength (e.g., another person whispering), or is it far, but has a strong original strength (e.g., another person shouting from across a field).

The mind also acts as a filter, damper, or amplifier when the person is in the auditory scene. Assume little Jane and little Johnny are having a dialogue and they are both speaking with comparable voice strengths and in comparable frequency bands (often children have high frequency voices regardless of gender). When Jane speaks to Johnny, her mind tends to dampen her voice and amplify Johnny's voice. And, vice-versa, Johnny's mind tends to dampen his voice and amplify Jane's voice. In Jane's mind, her voice does not sound louder than Johnny's voice, even though her vocal cords are nearer her ears. In addition, she does not lower her voice while speaking to Johnny. And, vice-versa, Johnny does not lower his voice while speaking to Jane.

Now assume, while talking to Johnny, Jane records the auditory scene. Given that Jane is holding the recording UE, without beamforming or filtering, the UE may record an auditory scene in which Jane sounds louder than Johnny. This recording is not representative to what Jane nor

Johnny are experiencing in real life. In real life, Jane and Johnny experience an auditory scene in which each of them speaks with a comparable voice strength.

To further show the challenge of recording the intended auditory scene, consider Michael is watching his daughter, Ana, playing soccer. Michael records some key moments of the match, as illustrated in Figure 1.

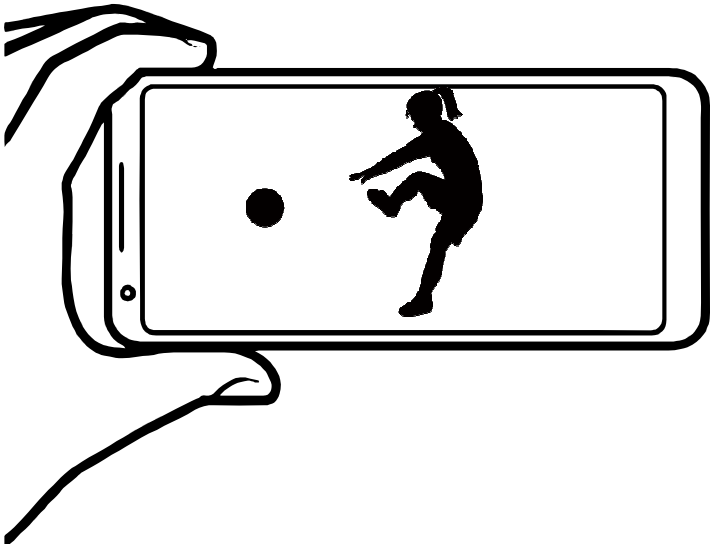


**Figure 1**

Ana's high school team is in the girls' soccer state tournament. The atmosphere is electrifying, and Michael proudly sees the fans cheering his daughter's team. As illustrated in Figure 1, Michael records some of these moments using his smartphone and wants to show the video with the accompanying auditory scene to his colleagues. As Michael is playing the video to his colleagues, he notices that the captured auditory scene does not match Michael's experience at the game. He remembers the roar of the fans, the players calling each-other, the girls forcefully kicking the ball, and the referee's whistling, but that is not how the UE captured the auditory scene. As Michael is watching and hearing the recorded video, the most prominent sound is wind noise. Michael's mind had filtered out the wind noise because Michael was focused on the interesting

part of the auditory scene — the players and the cheering fans. Michael, like most human beings, is capable of feeling the wind and his mind automatically filters out the wind noise and concentrates on the interesting part of the auditory scene.

One may have noticed this dichotomy of the recorded auditory scene and the environment of the user when watching an amateur video. Often there is wind noise that seems out of place and not representative of the auditory scene, such as a recording at a beach. This dichotomy, however, is often not present when watching a professional video. Obviously, one reason is that the professional videographer avoids background noise, such as wind noise near the recording UE, but he or she also uses beamforming and acoustic filtering that matches the user's environment. Hence, the recorded auditory scene of a professional video sounds more natural.

Currently, the user may manually activate the beamforming and the various acoustic signal capturing features (e.g., filters and amplifiers) to filter out wind noise. While these beamforming techniques and sophisticated acoustic signal filtering allow the user to capture the intended auditory scene, the user experience (UX) suffers by the fact that the user engages in manipulating the UE, instead of effortlessly capturing the intended auditory scene.

Furthermore, there is a challenge combining traditional beamforming algorithms with traditional wind-noise suppression algorithms. One of the challenges is that performing wind-noise suppression may accidentally remove spatial cues, which are necessary to create beamforming. This happens because the UE performs the beamforming algorithm and the wind-noise suppression algorithm in stages. In the first stage, the UE extracts spatial cues that tell the UE where the sound is coming from (e.g., left, right, ahead, or behind). In the second stage, the UE performs wind-noise suppression algorithms, which may accidently remove some of the features extracted in the first stage. The result may be a monophonic sound reproduction that lacks

the spatial information of the auditory scene. Therefore, it is desirable for the UE to perform beamforming and wind-noise suppression algorithms at the same stage, or *in situ*.

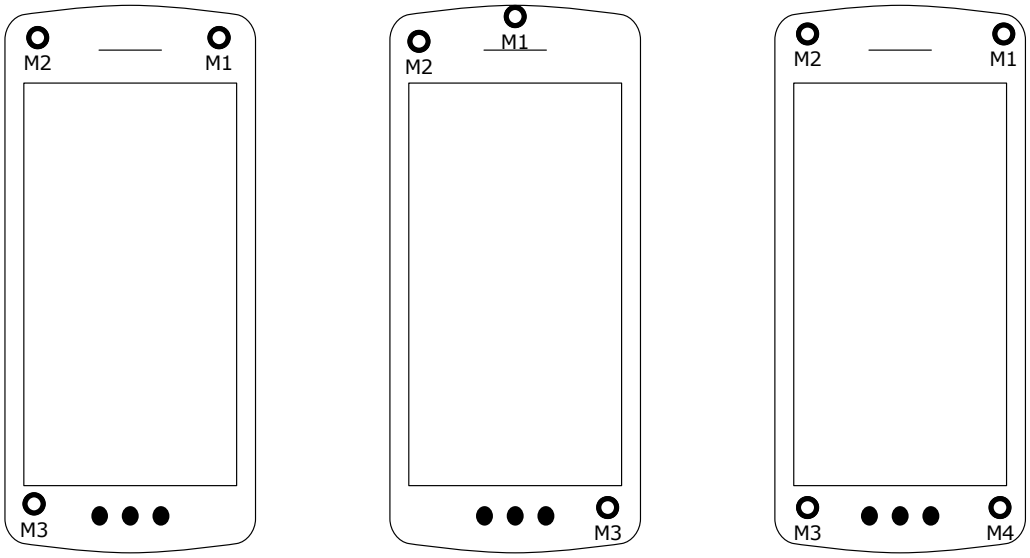
**Description:**

This publication describes audio-recording techniques, which allow a user equipment (UE) to determine a user's intent and apply audio beamforming and wind-noise filtering based on the user's intent. The UE engages in a process of evaluating the user's intent and applies the proper beamforming and acoustic filtering depending on the auditory scene. For the UE, evaluating the point of interest in the auditory scene may be a challenging process.

To enhance the quality of the captured auditory scene, various UE incorporate an array of microphones for sensing acoustic waves. UE developers try to embed the array of microphones in such a way that they cover a corresponding plurality of acoustic waves and, subsequently, develop a corresponding plurality of captured signals. Furthermore, the UE's processor is configured to perform a beamforming operation to combine microphone signals into combined signals. The number of combined signals is greater than one (1) and less in number than the number of microphone signals. In addition, the UE's processors may filter the plurality of microphone-captured signals.

As the auditory scene's various acoustic signals propagate, each microphone detects a different sound power level or sound strength, because the farther the distance from the sound source (e.g., the user's vocal cords), the weaker the sound power at the destination (microphone). Even though the sound power variation at each microphone is small given that the embedded microphones are clustered close to each other, it is still detectable. In addition, the various acoustic signals of the auditory scene reach each UE's microphone at different times. Furthermore, in some implementations, the UE may measure the phase differences of the acoustic signals in the auditory

scene as they reach each microphone. To create acoustic beamforming, the UE uses three or more non-collinear microphones embedded on the device, which detect the same sound independently. Figure 2 shows examples of these microphone placements on the UE.




**a) An example of three embedded microphones**

**b) Another example of three embedded microphones**

**c) An example of four embedded microphones**

**Figure 2**

Figure 2 shows three examples on the possible locations where the microphones (labeled M1, M2, M3, and M4) may be embedded on a smartphone. This concept, however, may apply to any UE used to record or capture an auditory scene, such as conferencing devices, notebooks, computers, headphones, wireless headphones, microphones, wireless microphones, digital cameras, hearing aids, digital audio recorders, and the like. Figures 2a, 2b, and 2c also show that the microphones are not embedded collinearly. Because two microphones can be collinear and equidistant from the user, the UE uses three or more embedded microphones. Theoretically, however, many more microphones can be used.

The exact placement of these example microphones may depend on the type of the UE and the spatial correlation of the microphones to the UE. A design of experiment (DOE) helps

determine the number of microphones and the placement of microphones at the UE. The DOE may evaluate the source of acoustic sound (e.g., wind, fan, waterfall, speech, musical instrument), the shape of the acoustic scene, the sound strength, the sound origin location, the acoustic signal's time delay at each microphone, the acoustic signal's phase at each microphone, the use of various acoustic filters, empirical evaluations of best-in-class solutions in an acoustic laboratory environment, and numerous other variables that affect capturing the intended auditory scene.

The optimal placement of the microphones helps address one aspect of the challenge to capture the intended auditory scene — beamforming. In contrast, the optimal use of the various acoustic-signal-processing features (e.g., filters and amplifiers) requires extensive manipulations by the user. In some aspects, the optimal combination of the various signal processing features is somewhat subjective — the user decides the outcome. Nevertheless, people generally agree what is the intended captured auditory scene when they hear one.

Wind-noise suppression, or reduction, poses many challenges in capturing the intended auditory scene. Wind-noise suppression is challenging because wind can be directional, not stationary, unpredictable, and varying in strength and frequency.

To this end, to determine the large number of possible acoustic signal beamforming and wind-noise filtering settings, the audio-recording technique leverages machine learning (ML), as shown in Figure 3.

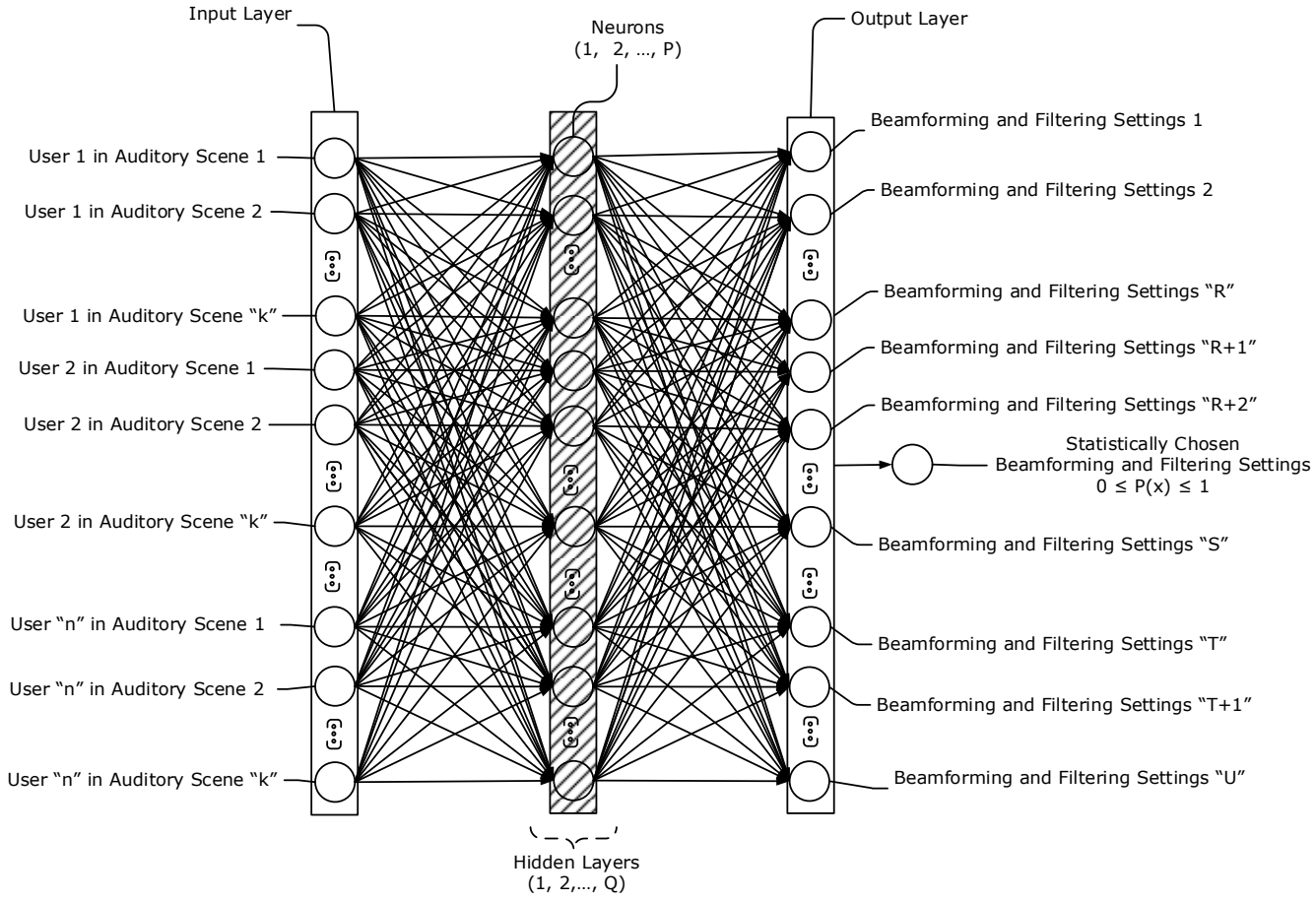



**Figure 3**

Figure 3 demonstrates a neural network, which is used to analyze the auditory scene and apply the optimal beamforming and wind-noise filtering settings that best-represents the intended auditory scene. The neural network illustrates an input layer, several hidden layers, and an output layer. The input layer includes "k" auditory scenes from "n" number of users, which are captured by the UE's microphones. The auditory scene of the input layer may represent an auditory scene without wind noise, such as indoors, or an auditory scene with wind noise. There are "Q" number

of hidden layers with up to "P" number of neurons in each layer. There can be a different quantity of neurons in each hidden layer. The output layer includes "U" number of bins with different probabilities on the beamforming and wind-noise filtering settings to capture the intended auditory scene, where $U > T > S > R > 0$. The intended auditory scene preserves the spatial cues and suppresses wind noise. The neural network interprets as the correct output the bin with the closest probability to one (1).

Given the large computational power that machine learning uses to train a model to analyze so many auditory scenes and user preferences, the model training may be performed on a cloud, server, or other capable computing device or system.

During the first stages of machine learning training, the model may ask the user for some input. For example:

- Is it windy? The model considers the sound of wind to be an undesirable sound and filters it out.
- Rate the type of wind — zero (0) for no wind or indoors, one (1) for a light breeze, two (2) for windy, three (3) for a strong wind.
- Rate the quality of the captured auditory scene — zero (0) for inadequate quality, five (5) for excellent quality.

As the machine learning training, with some initial user input, creates a more-sophisticated model, periodic model updates are sent to each UE, which allows the UE to execute the model even if that UE does not have the resources to update the model itself.

Further to the descriptions above, a user may be provided with controls allowing the user to make an election as to both if and when systems, programs, or features described herein may enable collection of user information (e.g., a user's preferences or a user's current location), and if

the user is sent content or communications from a server. In addition, certain data may be treated in one or more ways before it is stored or used, so that personally identifiable information is removed. For example, a user's identity may be treated so that no personally identifiable information can be determined for the user, or a user's geographic location may be generalized where location information is obtained (such as to a city, ZIP code, or state level), so that a particular location of a user cannot be determined. Thus, the user may have control over what information is collected about the user, how that information is used, and what information is provided to the user. The user may also choose not to take part in model training and accept the model as is.

In summary, the described audio-recording techniques use the existing acoustic beamforming and wind-noise filtering techniques, while integrating a DOE-determined optimal placing of microphones on the UE, machine learning, and some user input to aid the user or future users to effortlessly capture an auditory scene as intended.

**References:**

[1] Bouchard, Martin and Homayoun Kamkar Parsi. Method and system for a multi-microphone noise reduction. US Pub. 2011/0305345, filed February 10, 2010, and published December 15, 2011.