Piezoelectric MEMS Microphones

The only way to build reliable arrays for far-field voice recognition

White Paper
Rev1.0
Author: Udaynag Pisipati

Vesper Technologies Inc.
77 Summer St, 8th floor
Boston, MA - 02110

Table of Contents

Table of Figures

Introduction

Voice Assistants are now a dominant and fast-growing product category in the Consumer Electronics industry and a key platform that is driving the smart home revolution. Every major technological product in the market is moving towards integrating voice services. From smartphones and wearable devices to smart speakers, garbage cans to refrigerators, improvement in speech recognition performance is driving the success of voice assistants. What exactly in the audio signal chain is the key attribute to achieving these improvements? If you do a search on the internet for an answer, its highly likely that you'll end up with the notion that Artificial intelligence, Machine Learning, Natural Language Processing and signal processing are the key reasons driving the improvements in speech recognition. While these data processing technologies enable efficient reconstruction of the recorded signal, the single most important component for signal processing is the microphone that captures the signal at the front end of the audio chain. An input signal that is already distorted at the microphone output can only get worse along each step of the signal chain as it gets converted to digital domain for processing. Proliferation of devices integrated with multiple microphones is therefore a significant contributor to the success of voice assistants. On the other hand, lack of significant advances in microphone technology has also prevented further improvement in voice assistants. The good news is piezoelectric microphones developed by Vesper Technologies are currently disrupting the microphone industry with a technology that is well suited for microphone arrays while also offering ruggedness and extreme durability over time. That being said, there is currently no clear consensus on what makes a microphone good or bad for far-field scenario where the user is at a distance from the device. Is it the Signal to Noise Ratio which is the most common industry metric for microphone selection? Is it the number of microphones in the array? This white paper is intended to provide insight into microphone selection for far-field performance using the

measurement data taken with Vesper's piezoelectric microphones in comparison to standard Electret Condenser microphones (ECM).

What is Far-field?

Use of distant microphones has fueled several new use cases for low power always-on smart solutions such as TV remotes, smart speakers, wearables etc. To distinguish between near-field and far-field, consider the example of a TV remote. While you are using a push to talk button on Fire TV remote, the microphone on the remote is close to your mouth. This is considered a near-field scenario within 1 wavelength of the lowest frequency of interest. For the speech signal where the frequency of interest is between 300 Hz and 3400 Hz, this wavelength translates to a range of 9.5 cm to 1.1 meters from the sources ($\lambda = c/f = 331.1/300 = 1.104$ meters, where c is the speed of the sound). Therefore, the general near-field range of speech is from 9.5 cm to 1.1 meters from the source. Now, imagine you are using Amazon Echo to directly control the Fire TV with voice commands, without using a remote. In this case, speech signal is in the far-field of speech source, beyond 1m. For the current state of the art in speech technology, far-field can be generalized to be anywhere from 1 to 7 meters from the source.

In the near-field environment, user's speech is louder than the ambient noise. However, in far-field, two main factors impact the audio quality at the input of the microphone – a) additive noise introduced by surrounding environment and b) reverberations or reflections of the original source signal on the walls and objects in the room. The incoherent nature of these noise sources adds to the complexity of noise suppression solutions such as Beamforming, Blind Source separation or Deep Learning algorithms used for speech recognition in new generation audio products. Product manufacturers have therefore moved towards microphone arrays instead of single microphone solutions for speech enhancement and advanced noise suppression. While microphone arrays improve the

performance of noise suppression algorithms, they also increase the overall system noise of the product, for example, when used in combination with Beamforming algorithms. Therefore, a microphone with good acoustic characteristics such as low self-noise is crucial for high performance in arrays. At the same time, low current drain and quick startup are also critical to meet the low power requirements and efficient wake word detection for battery powered, always listening solutions. The following sections provide background on far-field performance metrics and analyze the test results relative to microphone specifications.

Far Field Performance Tests

Several factors contribute to the performance of far-field processing as shown below.

a) Ambient noise and Reverb characteristics of the room
b) Distance and angle of the speaker relative to the far-field microphone array
c) Spectral characteristics of the distractors – level, type of distractor
d) Distance and angle of the distractor relative to the speaker
e) Speech utterances used for the evaluation
f) Speech Signal to Noise ratio
g) Speech Signal to Echo ratio for voice barge-in performance

Test Setup

For this study, far-field performance metrics are evaluated by comparing a 2-Mic VM1001 array from Vesper Technologies with the 2-Mic EM6027 array from Horn which previously shipped along with Synaptics CX20921 reference. The reference kit includes hands-free voice interaction with Synaptics' proprietary Smart Source Pickup™ for background noise suppression and an embedded low power Alexa wake word engine for voice activation from Sensory™. In addition, full duplex stereo Acoustic Echo Cancellation algorithm

integrated into the development kit enables music barge-in performance. Music barge-in provides the ability to detect Alexa wake word even when the device is playing music or voice prompts loudly.

As of the date measurements were taken, CX20921 reference platforms shipped with 2-mic slider board that has EM6027 Electret Condenser Microphone (ECM) with adjustable distance between the microphones. Microphone distance of 55mm as recommended by Synaptics' documentation is used for all the measurements. For the Vesper array, the 2-mic CX20921 slider board is replaced with Vesper's design of 2-mic VM1001 slider board which is shown in Figure 1 and Figure 2.

Figure 1: Vesper VM1001 mics on slider – Microphone lids facing on 2 mini-pcbs

Figure 2: Vesper VM1001 mics on slider – Acoustic port view

The complete test setup for Synaptics reference platform is shown in Figure 3. For simplicity through rest of the paper, EM6027 2-mic array will be denoted as ECM and VM1001 2-mic array as Vesper.

Amazon Echo dot is used as another reference device to provide comparison between 2-mic and 7-mic array. Echo dot uses Amazon's proprietary algorithms for noise suppression/echo cancellation with Alexa Voice Service and a 7-mic capacitive MEMS microphone array.

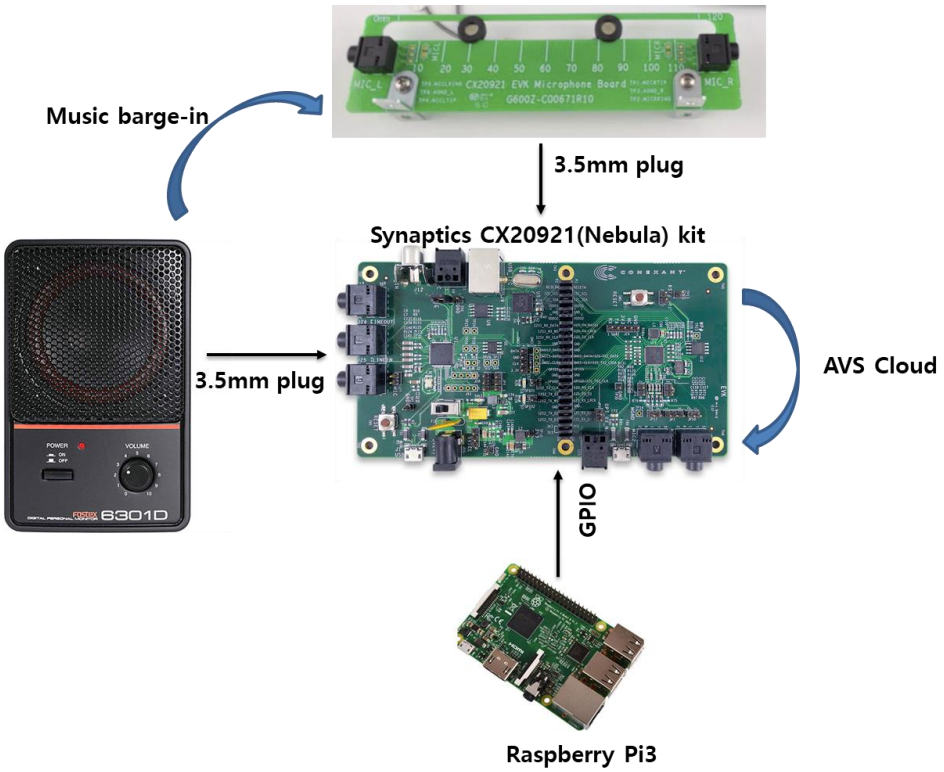

Figure 3: Device under Test configuration for Synaptics Reference platform

A datasheet comparison between Vesper VM1001 and Horn EM6027 microphone is provided in the Table 1 below.

	VM1001	**EM6027**
Type	Piezoelectric MEMS	ECM
Sensitivity	-38 dBV	-31 dBV
SNR	64 dB	70 dB
Acoustic Overload Point	127 dBSPL	115 dBSPL
Sensitivity Matching	+/- 1 dB	+/- 3 dB

Table 1: Datasheet comparison between Vesper VM1001 Vs EM6027 microphone

The ECM array uses a default gain of 24 dB in the Analog to Digital conversion. To compensate for the sensitivity difference between ECM and Vesper microphones, an additional gain of 6 dB is applied to VM1001.

Test Environment

Far-field performance tests are performed in an isolated sound environment at Mobimark Labs, an independent audio test laboratory, to ensure consistent and repeatable results for accurate evaluation. The ambient noise level of the room is measured to be 33 dB(A) and the RT60 reverberation time is measured to be 370 milliseconds.

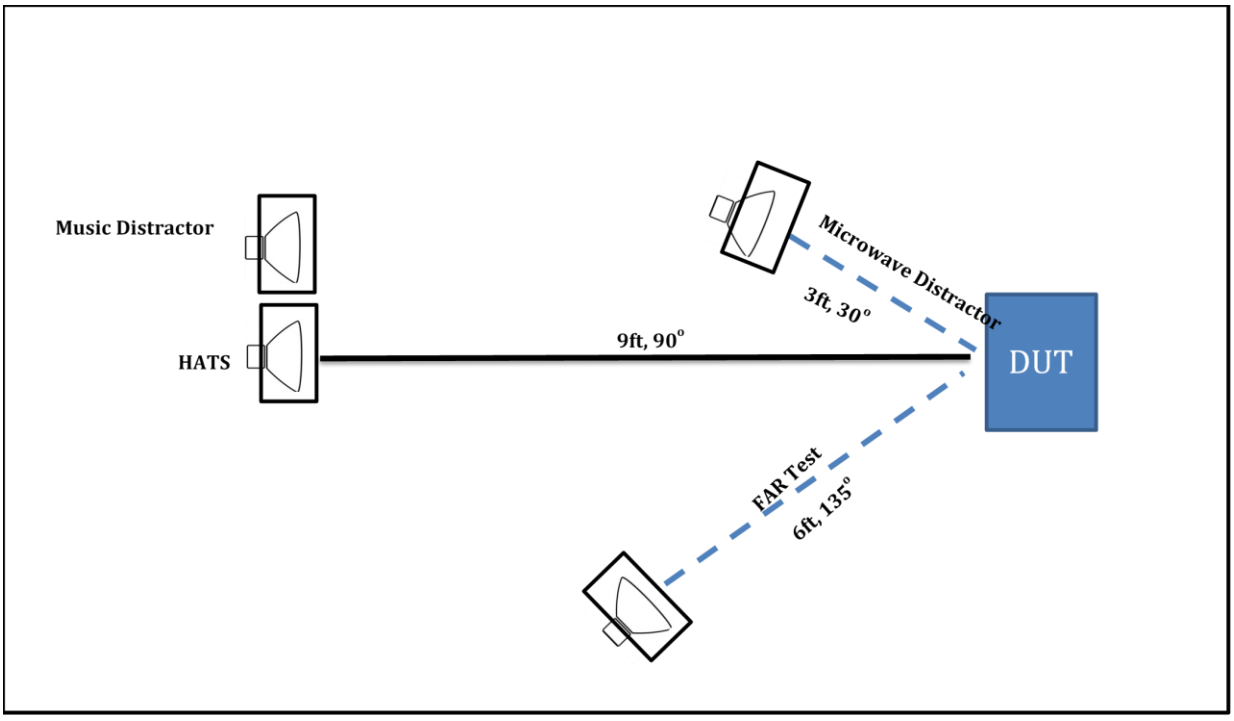

Figure 4: Test Setup for Far-field Performance Tests

Devices are tested in quiet condition as well as in the presence of distractors at different distances and angles as shown in Figure 4. Microwave Oven Noise and Music Playback

conditions are tested with a speech Signal to Noise Ratio of +5 dB played back using Fostex 6301NE powered speakers. In addition, Device playback scenario is tested at a Speech to Echo ratio of -15 dB to evaluate music barge-in performance.

A standardized speech database that consists of 96 utterances from 8 individual voices with 4 male and 4 female speakers is used for recording responses. Speech utterances are carefully chosen to cover different Alexa commands typically used in several Alexa service domains, for example, music playback, calendar/alarm setup etc. A Bruel & Kjaer Head and Torso Simulator(HATS) loudspeaker located at a distance of 9ft/90^{o} from the DUT is used to playback the utterances from a computer with high quality sound card. Adobe Audition is used as playback software.

All the test signals including speech and noise files are calibrated using pink noise at the required sound pressure level given in Table 2. A sound level meter with c-weighting and Fast integration is used to measure the required levels of the pink noise for the corresponding test condition. For example, in quiet conditions, pink noise is calibrated to sound pressure level of 63 dBC measured on a sound level meter at 1 meter from the source. Speech level is then adjusted to match the calibrated pink noise level.

	Speaker			**Distractor**			**Test Metric**
Condition	Distance (ft)	Angle	Level (dBC)	Distance (ft)	Angle	Level (dBC)	
Quiet	9’	90^{o}	63	N/A	N/A	N/A	FRR/RAR
Microwave Oven	9’	90^{o}	63	3’	30^{o}	58	
Music playback	9’	90^{o}	63	9’	90^{o}	58	
Device Music	9’	90^{o}	63	N/A	N/A	78	FRR
News	N/A	N/A	N/A	6’	135^{o}	60	FAR

Table 2: Speaker and Distractor Levels

Test Metrics

The following test metrics are used to evaluate the far-field performance of device under test

False Rejection Rate (FRR) is calculated as the ratio of missed wake words to the number of wake-words spoken. A lower value of FRR is expected for a device with good wake word detection accuracy. For example, a device that missed 1 out of 10 wake words will have a 10% FRR.

Response Accuracy Rate (RAR) is calculated as the ratio of successful Alexa responses to the total number of requests. For example, if 10 commands are requested and Alexa responds accurately to 9 commands, RAR will be 90%. A high value of RAR is therefore better.

False Acceptance Rate (FAR) is the rate of false triggers when DUT is exposed to continuous news stream from a broadcast program within a 24-hour timeframe.

For each test condition, all the 96 utterances are played to the device under test and evaluated for FRR and RAR metrics. For the analysis, an average of FRR and RAR from all the utterances is used for each test condition. False Acceptance Rate is obtained using Alexa voice service's history dialog.

Test Results

Far-field performance in Noise Conditions

FRR and RAR performance of the three devices under test for different background noise conditions are shown in Figure 5 and Figure 6.

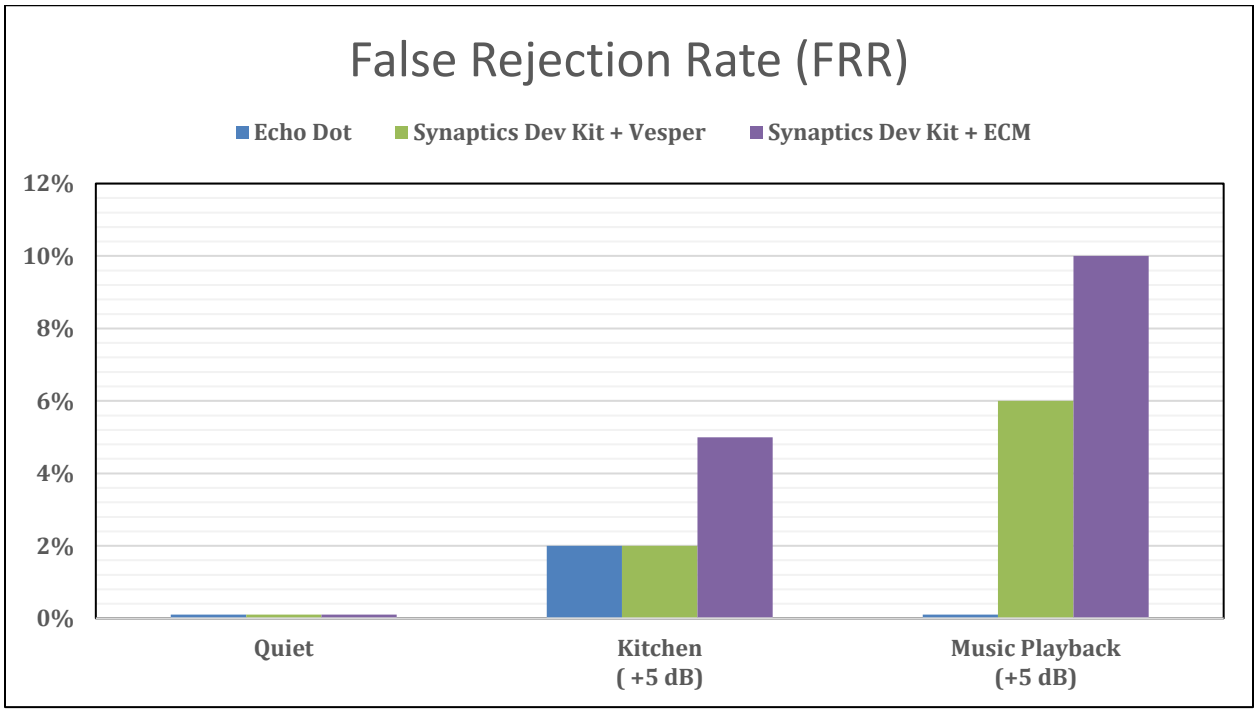

Figure 5: False Rejection Rate (FRR) comparison (lower is better)

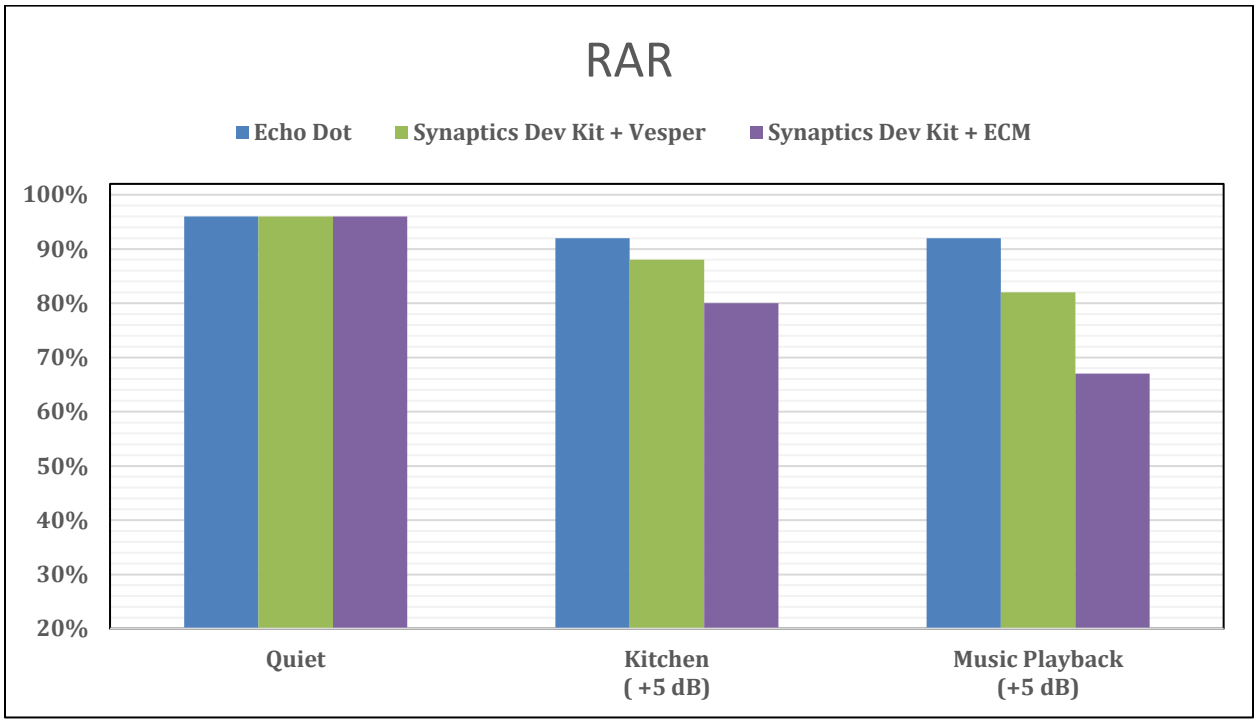

Figure 6: Response Accuracy Rate (RAR) comparison (higher is better)

Echo Dot provides the best wake word detection (False Response Rejection) and Response Accuracy in all noise conditions, which can be attributed to the seven microphone-array on the device. ECM and Vesper arrays provide 100% accuracy in wake word detection in a silent environment. In the presence of distractors, Vesper's VM1001 microphone offers 4% improvement in wake word detection compared to ECM. The same is the case with RAR, where VM1001 significantly outperforms ECM mics.

The following sections provide further insight on far-field test results with respect to applicable microphone performance metrics.

Sensitivity Matching

Spatial signal processing algorithms used in microphone arrays are very sensitive to mismatch in microphone characteristics such as sensitivity and frequency response.

Sensitivity mismatch between the microphones in an array can be interpreted by the algorithm as a directionality cue, thereby altering the orientation and the desired response of the array. Figure 7 below illustrates the impact of 3 dB shift in sensitivity of a single microphone within a 4-mic array. The array with tight sensitivity matching shown on the left exhibits excellent directionality and a hyper cardioid pattern with the null at 90 degrees for 300 Hz frequency. A 3-dB shift in sensitivity of one of the microphones in the array reverses the directionality of the beam steering it away from the expected orientation

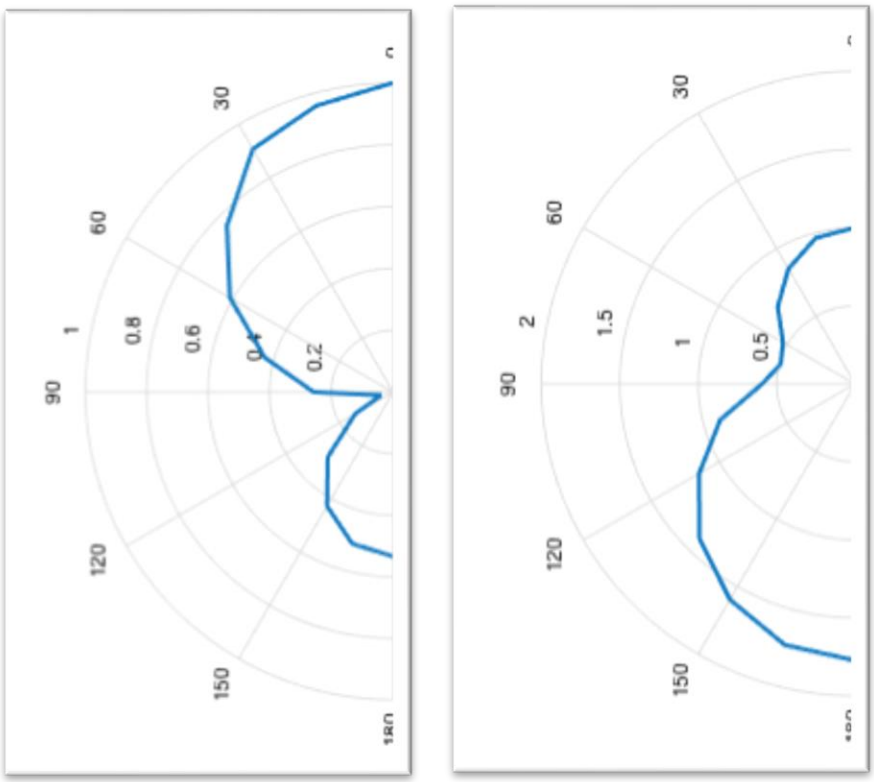

Figure 7: 4-microphone array with excellent directionality (Left), 3 dB shift in one of the microphones leading to backwards directionality (Right)

Figure 8 and Figure 9 show the sensitivity difference between the two microphones for ECM Vs Vesper Array. ECM microphones show a sensitivity difference of 3 dB between

the microphones, whereas Vesper's microphones have a sensitivity difference within only 1 dB.

Figure 8: Sensitivity matching between EM6027 Mics in Synaptics Reference kit

Figure 9: Sensitivity matching between VM1001 Mics in Synaptics Reference kit

Sensitivity shifts in an array can occur for 2 reasons –

1. Manufacturing tolerances: Electret microphones are often susceptible to manufacturing tolerances due to their inherent sensitivity to temperature and incompatibility to solder reflow process. A typical ECM microphone can have a sensitivity shift of +/-4 dB over its operating temperature range. Capacitive MEMS microphones are compatible to solder reflow, but are prone to flux vapors getting deposited on the diaphragm, thereby resulting in sensitivity shifts.

2. Degradation over time: Sensitivity shifts can also happen in real-world applications when the products are exposed to environmental contaminants such as kitchen oil, dust, water, moisture etc. Capacitive MEMS microphones can have these contaminants trapped between the diaphragm and back plate, thereby degrading the performance of the array over time.

Piezoelectric microphones are the best solution to both the problems above with their solder reflow compatibility and a single layer MEMS design which makes it immune to sensitivity shifts from accumulation of contaminants. Vesper microphones, therefore, provide tight sensitivity matching during manufacturing process as well as remain stable and reliable through the product's lifetime. These characteristics ensure long-term acoustic performance of far-field arrays built with Vesper microphones. Results from stress test experiments on Vesper's piezoelectric microphones are covered in detail in [1]

Signal to Noise Ratio (SNR)

Signal to Noise Ratio is the difference between the sensitivity of the microphone, and the inherent self-noise level of the microphone. A high SNR value indicates a quiet microphone. SNR has been a significant metric for the characterization of MEMS microphones in single and multi-mic arrays in mobile devices, particularly for far-field situations where the ambient noise and room reverberation add to the acoustic path loss of the speech signal. Our test data shows that Vesper array with lower SNR provides a better far-field performance for all noise conditions tested, compared to ECM array. 8 dB higher SNR in EM6027 microphone does not translate into a better far-field performance.

SNR metric in a datasheet is calculated as a ratio of reference level of 94 dB and the noise level at the output of the microphone, measured over 20 kHz bandwidth. However, given that the speech signal has its spectral content within the 8-kHz bandwidth, voice

processing algorithms, typically, filter out the captured signal above this bandwidth for computational efficiency. Therefore, SNR of a microphone measured within 8kHz bandwidth is a better indication of how it performs in a far-field voice processing algorithm than 20kHz bandwidth since 8kHz is the actual bandwidth used by the audio subsystem. Typically, a higher SNR microphone (> 60 dB) is recommended for far-field scenarios. However, array configuration and signal processing down the audio path can significantly alter the overall SNR of the system which is more critical for far-field performance instead of the SNR of the discrete microphone. Our data also shows that beyond a certain value, SNR alone would provide a diminishing return, especially if high performance is not maintained in other parameters such as sensitivity matching, voice band performance etc. Even the highest SNR mic will not make the best array, if the other specifications are sub-par. Vesper's VM1001 microphone optimizes performance metrics in the voice band to make better far-field arrays.

Far-field performance with Music Barge-in

Far-field performance of Vesper array compared to ECM array in music barge-in (device playback) scenario is shown in
Figure 10. Vesper array improves the wake word detection by 3% compared to ECM array and 6% compared to 7-mic capacitive MEMS microphone array on Echo Dot.

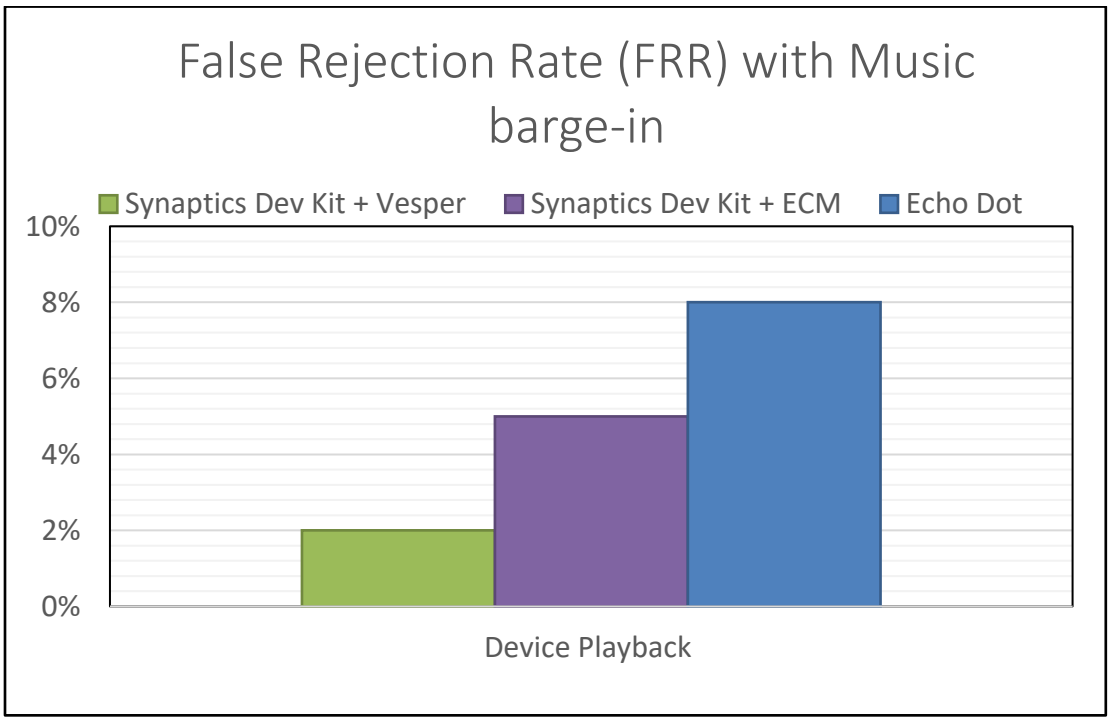

Figure 10: Sensitivity matching between VM1001 Mics in Synaptics Reference kit

Acoustic Overload Point (AOP)

Acoustic Overload point is the maximum acoustic level that a microphone can hear without significant distortion, typically specified at 10% Total harmonic distortion. Higher AOP microphone can withstand higher acoustic input levels before introducing non-linear distortions into the audio signal chain. This is particularly useful for Acoustic Echo Canceler to improve wake word detection accuracy in a low SNR scenario such as voice/music barge-in. In the music barge-in test condition, music file is played at a level 15 dB higher than the speaker issuing the voice prompt. The 3% improvement in wake word detection can be attributed to the higher AOP on Vesper microphones. It is also interesting to see that Vesper 2-mic array also improves the wake word detection compared to the 7-mic capacitive MEMS microphone array on Echo Dot. For a 10% total

harmonic distortion level, ECM microphones have 115 dBSPL AOP whereas Vesper microphones have an AOP of 127 dBSPL, which means that the Vesper microphone is linear until it reaches a sound pressure level that is 12 dB more, when compared to an ECM microphone. Piezoelectric MEMS microphones tend to exhibit better linearity characteristics with increasing sound pressure than capacitive MEMS microphone. This is because linearity of piezoelectric microphones is only limited by the ASIC voltage rail, since the MEMS design itself does not hit 10% distortion until 160 dBSPL. A recent article in Acoustical Society of America by Martin Ring from Bose [2] highlights the linearity advantages of Piezoelectric microphones compared to capacitive microphones.

False Acceptance Rate (FAR)

FAR indicates how many times the device under test wakes up when there is no wake word spoken, within a 24-hour timeframe. Vesper array provides comparable results to Echo Dot in terms of false positives as shown in Figure 11. ECM array, on the other hand has more false alarms, meaning the device wakes up without the actual wake word spoken in the news recording.

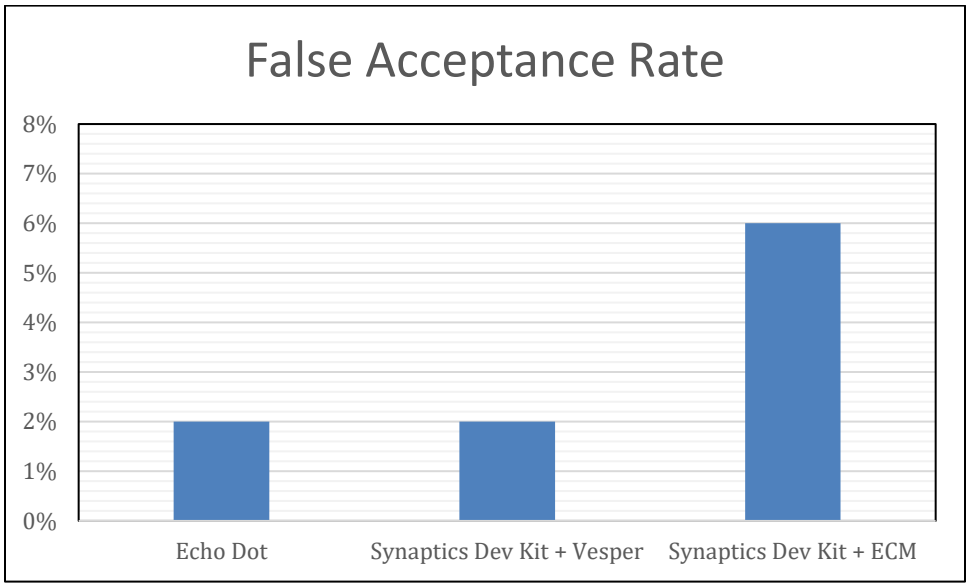

Figure 11: False Acceptance Rate comparison

Effect of number of microphones

The more the number of microphones in an array, the better the far-field system performance. Number and geometry of microphone placement depends on the expected use case and system cost. For example, a Smart TV mounted on a wall will have the user issuing commands only from the front of the system where a linear array of 2-4 microphones would suffice. For a smart speaker system at the center of a room, 360-degree field of the user's speech might require a circular array of 6 microphones.

Theoretically, a 7-microphone array should improve the system performance significantly compared to a 2-mic array. Test results from Echo Dot indicate that the 7-microphone capacitive MEMS array improves the overall wake-word detection and response accuracy compared to 2-mic Vesper array. However, it is worth noting that the performance of 2-mic Vesper array is close to that of the 7-mic capacitive MEMS array in Echo Dot compared to ECM array. A major constraint for system designers is the increase in system costs that come with the increase in number of microphones in array. Given the marginal

performance improvement achieved with 7-capacitive MEMS microphones compared to say, a 4 or 6 microphone array, Vesper arrays with lower number of microphones could be a better tradeoff for low cost products. Alternatively, large arrays of Vesper MEMS microphones could be used to achieve the highest levels of performance, accuracy, longevity and reliability for premium products. Acoustic mesh/membrane required to protect capacitive MEMS microphones from environmental contaminants is another constraint that adds up to the overall system cost, while reducing performance and longevity. Acoustic mesh typically degrades the sensitivity of the microphone by 3 dB, thereby decreasing the overall SNR available for signal processing while adding cost and manufacturing complexity. Vesper's piezoelectric microphones offer a compelling alternative, with native resistance to environmental factors such as dust, water and kitchen oil etc. at the MEMS itself, thereby eliminating a mesh or membrane and preserving the sensitivity, SNR and long-term speech recognition accuracy of the microphone array.

Conclusion

The microphone is the key component in audio signal chain. Based on our test data, it is evident that Vesper MEMS microphones have superior far-field performance combined with savings in overall system cost. Due to this premium performance of Vesper microphones, Synaptics has recently selected Vesper microphones for its two- and four-microphone Amazon AVS development kits [3]. Besides the performance advantages, dust and moisture resistance of the piezoelectric technology creates durable microphones dramatically increasing the long-term stability of the arrays and hence providing a superior value when compared to ECM or capacitive MEMS technologies. Our analysis also concludes that metrics such as tight sensitivity matching, and Acoustic Overload Point are all important criteria for microphone selection apart from Signal to Noise Ratio. The appropriate metrics for proper microphone selection depends on the use case as well as

the system constraints such as cost, design and selection of DSP hardware/software voice processing solutions.

As the demand for voice as the user interface increases, products with more diverse and complex use cases will require a strong emphasis on microphone selection. This white paper is intended to provide initial guidance on the microphone selection. To get a better understanding of the importance of each metric in far-field performance, further investigation on piezoelectric microphones in comparison with existing microphone technologies and speech algorithms is required. Vesper's Zero Power Listening™ is another incredible technology that provides significant power savings with its almost zero-power draw and a fast startup - metrics that are critical to provide long batter life and accurate wake word detection. These topics will be covered in the subsequent white papers. In the meantime, any inquiries on Vesper product's and roadmap should contact us at info@vespermems.com

Citations

[1] "SmartEverything and the Rise of the Microphone Array," [Online]. Available: https://www.edn.com/design/analog/4457396/3/SmartEverything-and-the-Rise-of-the-Microphone-Array.

[2] M. D. Ring, "When good Mics go bad," *The Journal of Acoustical Society of America,* pp. 141, 3676, 2017.

[3] "Vesper partners with Synaptics for Advancing Far field voice interfaces for Amazon Alexa Voice service," [Online]. Available: https://www.broadwayworld.com/bwwgeeks/article/Vesper-Partners-with-Synaptics-Advancing-Far-field-Voice-Interfaces-for-the-Amazon-Alexa-Voice-Service-2.