Program Digest, 2019 IEEE International Ultrasonics Symposium (IUS)
Glasgow, Scotland, October 6-9, 2019

Fast GPU-Assisted FEM Simulations of 3D Periodic TCSAW, IHP, and XBAR Devices

J. Koskela[1,2], V. P. Plessky[1,2], B. A. Willemsen[2], P. J. Turner[2], B. Garcia[2], R. B. Hammond[2], and N. O. Fenzi[2]
[1]*GVR Trade SA*, Gorgier, Switzerland, [2]*Resonant, Inc.*, Santa Barbara, California, USA
vplessky@resonant.com

***Abstract*—Recently, hierarchical cascading accelerated with graphics processing units (GPUs) has proven efficient for accelerating 3D periodic finite-element method (FEM) analysis of acoustic wave devices. However, the limited memory available in GPUs severely restricts the accuracy of the simulations; moreover, numerical libraries are not fully matured yet. This paper considers two techniques to circumvent these limitations. Firstly, to extend the size of the problems that can be handled, the use of GPUs can be limited to isolated cascading operations. While the related data transfer between RAM and GPUs hinders the efficiency of GPUs, significant acceleration is still achieved as compared to CPU-based implementation. Secondly, a GPU-friendly alternative is presented for the reduction of FEM system matrices to B-matrices. Examples of simulations of 3D periodic TCSAW, IHP, and XBAR arrays are provided.**

***Index Terms*—SAW, FEM, SAW simulation**

I. Introduction

The stringent requirements on modern RF filters have led to the introduction of new families of acoustic devices based on thin-film technologies. These have demonstrated high temperature stability (TCSAW [1]), extremely low loss (IHP [2], [3]), and operation frequencies up to 5 GHz (XBAR [4]). However, they tend to exhibit complicated acoustic behavior, exhibiting spurious transversal modes and acoustic radiation. The understanding and suppression of these mechanisms calls for fast and accurate physics-based simulation tools.

The hierarchical cascading FEM [5] has proven an efficient tool for 2D FEM simulation of SAW devices. Dramatic acceleration in 3D periodic simulation was recently achieved by accelerating hierarchical cascading with GPUs [6]. However, the limited memory available in GPUs significantly restricts the size and accuracy of the models that can be analyzed. This restriction can be relaxed somewhat by limiting the GPUs only to perform cascading operations: B-matrices are transferred from RAM to the GPU, cascaded, and the result is retrieved from the GPU to RAM. The large amount of slow data transfer results in rather inefficient use of the GPU computational power. Nevertheless, it is still considerably faster than purely CPU-based computation.

With the bulk of the cascading work moved to GPU, the reduction of the FEM system matrices to B-matrices becomes a limiting factor. This can be addressed by direct synthesis of the B-matrices from the FEM element equations.

GVR Trade SA is a wholly owned subsidiary of Resonant Inc., Santa Barbara, CA, USA 93117.

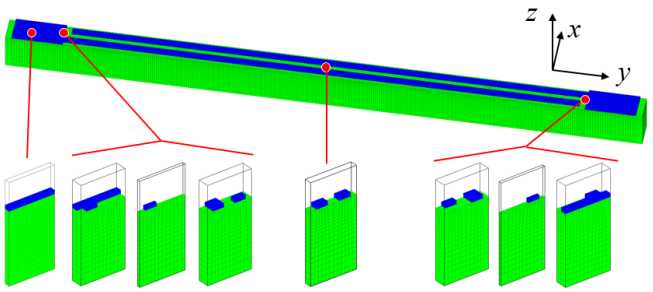

Fig. 1. Unit cell of a 3D periodic electrode array and its decomposition into unique unit blocks. The unit cell consists of an electrode pair, gaps, and busbars, the underlying piezoelectric substrate mesh, and the vacuum above. In y- and z-directions the mesh is surrounded by perfectly matched layers (PML, not shown). Unit blocks from left to right: busbar, transition between busbar and gap, left gap, transition between gap and IDT, IDT block, transition between IDT and gap, right gap, and transition between gap and busbar. Blocks associated with side PMLs are not shown.

This paper is organized as follows. Sec. II provides a brief overview of the unit block modeling with FEM. The principles of hierarchical cascading algorithm are discussed in the preceeding work [5] and not repeated here. Instead, the paper focuses on the options for evaluating the B-matrix. Examples of 3D periodic simulations and the recorded performance metrics are presented in Sec. III. Discussion and conclusions are provided in Sec. IV.

II. Hierarchical Cascading Method

A. FEM Modeling of Unit Blocks

Figure 1 shows a computational mesh for a unit cell of a 3D periodic electrode array, decomposed into unique unit blocks for evaluation with hierarchical cascading. Consider harmonic excitation at angular frequency $\omega = 2\pi f$. Modeling any of the unit blocks with FEM yields a linear system of equations

$$\left[\mathbf{K} + i\omega\mathbf{D} - \omega^2\mathbf{M}\right](\mathbf{x}) = (\mathbf{F}). \tag{1}$$

Here, the expression in the brackets is the system matrix, consisting of the stiffness matrix $\mathbf{K}$, damping matrix $\mathbf{D}$, and mass matrix $\mathbf{M}$. They are inherently symmetric. The vector $\mathbf{x}$ contains the degrees-of-freedom (DOFs) of the model: the nodal values of mechanical displacement and electric potential at the nodes. The vector $\mathbf{F}$ contains the external sources–the charge density and the boundary stresses.

The DOFs and the external sources can be classified into those associated with the boundaries in the left-right aperture direction y, periodic front-back direction x, those associated with the interior, and the electric potentials connected to electrodes (V). Periodic boundary conditions can be applied to eliminate the DOFs on the back surface. With the front DOFs subsumed into those for left (L), right (R), and interior (I), the system of equations (1) can be reordered as

$$\begin{bmatrix} \mathbf{A}_{\mathrm{LL}} & \mathbf{A}_{\mathrm{LI}} & 0 & \mathbf{A}_{\mathrm{LV}} \\ \mathbf{A}_{\mathrm{IL}} & \mathbf{A}_{\mathrm{II}} & \mathbf{A}_{\mathrm{IR}} & \mathbf{A}_{\mathrm{IV}} \\ 0 & \mathbf{A}_{\mathrm{RI}} & \mathbf{A}_{\mathrm{RR}} & \mathbf{A}_{\mathrm{RV}} \\ \mathbf{A}_{\mathrm{VL}} & \mathbf{A}_{\mathrm{VI}} & \mathbf{A}_{\mathrm{VR}} & \mathbf{A}_{\mathrm{VV}} \end{bmatrix} \begin{pmatrix} \mathbf{x}_{\mathrm{L}} \\ \mathbf{x}_{\mathrm{I}} \\ \mathbf{x}_{\mathrm{R}} \\ V \end{pmatrix} = \begin{pmatrix} \tau_{\mathrm{L}} \\ 0 \\ \tau_{\mathrm{R}} \\ -Q \end{pmatrix}. \quad (2)$$

Here, $\mathbf{A}$ is the reordered system matrix. It is symmetric and very sparse. We have also summed over the equations associated with the electric DOF on the electrode; this can be interpreted as integration over charge density. On the right-hand side, τ_{L} and τ_{R} are integrals over surface stresses at the left and right edge. These will cancel out later in the cascading process, in the same way as stresses between finite elements cancel out in the assembly of the system matrix. The vector Q denotes the net surface charges connected to different busbars; the currents flowing into the electrodes are $I = i\omega Q$.

B. Reduction to B-matrix

In the hierarchical cascading method, the system matrix is reduced to a B-matrix, which relates the DOFs at the boundaries of the unit blocks and the electric DOFs:

$$\begin{bmatrix} \mathbf{B}_{\mathrm{LL}} & \mathbf{B}_{\mathrm{LR}} & \mathbf{B}_{\mathrm{LV}} \\ \mathbf{B}_{\mathrm{RL}} & \mathbf{B}_{\mathrm{RR}} & \mathbf{B}_{\mathrm{RV}} \\ \mathbf{B}_{\mathrm{VL}} & \mathbf{B}_{\mathrm{VR}} & \mathbf{B}_{\mathrm{VV}} \end{bmatrix} \begin{pmatrix} \mathbf{x}_{\mathrm{L}} \\ \mathbf{x}_{\mathrm{R}} \\ V \end{pmatrix} = \begin{pmatrix} \tau_{\mathrm{L}} \\ \tau_{\mathrm{R}} \\ -Q \end{pmatrix}. \quad (3)$$

The B-matrix is symmetric and dense.

1) Schur complement: The most straightforward way to evaluate the B-matrix (3) is to compute the Schur complement $\mathbf{A}/\mathbf{A}_{\mathrm{II}}$: the formal solution for the internal DOFs $\mathbf{x}_{\mathrm{I}}$ from Eq. (2) is substituted back to Eq. (2), for example

$$\mathbf{B}_{\mathrm{LL}} = \mathbf{A}_{\mathrm{LL}} - \mathbf{A}_{\mathrm{LI}} \left[\mathbf{A}_{\mathrm{II}}\right]^{-1} \mathbf{A}_{\mathrm{IL}}, \quad (4)$$

and likewise for the other components. A numerically efficient implementation uses LU- or LDL-decomposition of $\mathbf{A}_{\mathrm{II}}$.

The Schur complement approach is sufficiently efficient for 2D simulation. However, it becomes a bottleneck in GPU-assisted 3D cascading, mainly due to the high sparsity of the system matrix and the larger number of DOFs in 3D. Computing a dense Schur complement from a sparse matrix, in particular on GPUs, seems to be a weak point in current numerical libraries. This is likely to change in future.

2) Direct Synthesis of the B-matrix: Alternatively, the B-matrix can be constructed directly from the element contributions, bypassing the building of the system matrix, see the visualization in Fig. 2. Within this paper, this approach it referred to as Direct Synthesis (DS). In practice, the main difference is that the subsystem matrices in DS are dense, allowing more efficient implementation on the GPU.

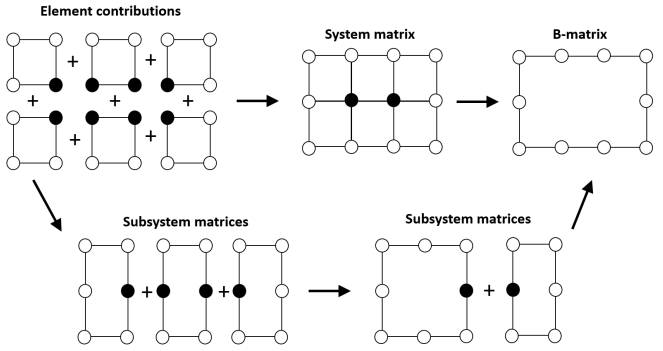

Fig. 2. Construction of the B-matrix via Schur complement (top row) and Direct Synthesis (bottom row). In the former approach, one first builds a system matrix from the contributions of finite elements, followed by elimination of the internal degrees-of-freedom (black circles). In Direct Synthesis, element contributions are combined gradually into subsystem matrices, eliminating internal DOFs at the earliest convenience.

III. Examples

A 3D periodic hierarchical cascading algorithm was implemented on the commercial Matlab platform[1]. Four test cases of varying complexity are considered: an LSAW array, an IHP array, a TCSAW array, and an XBAR array, all modeled using a structured mesh with quadratic elements. Tradeoffs between accuracy, memory consumption, and simulation time are unavoidable. The applied lateral mesh density varies from 12 (LSAW, IHP, TCSAW) to 24 (XBAR) elements per unit cell. The LSAW case is essentially the same as in Ref. [5] and it won't be further analyzed here. The other cases are discussed in the following subsections.

Tests were run on two high performance platforms equipped with computation-oriented GPUs: a workstation[2] with two nVidia Tesla Quadro GV100 GPUs (32 GB HBM2 memory each), and an Amazon cloud p3.8xlarge instance[3] with 4 nVIDIA Tesla V100 GPUs (16 GB memory each). The LSAW and XBAR test cases were run on both systems.

The measured simulation times are summarized in Table I, for a single frequency point and using only one GPU, and excluding the time required to construct the FEM models (typically 3-4 minutes). The implementation of the direct synthesis method on the CPU doesn't provide much benefit against the Schur complement method, but the implementation on the GPU is found about 4 times faster on the workstation and 6-10 times faster on p3.8xlarge. Cascading on the GPU is found 8 times faster than on the CPU on the workstation and 5-10 times faster on p3.8xlarge. Without direct synthesis, the CPU-based Schur complement method would be the dominant bottleneck on both platforms.

[1]MathWorks Inc., Natick, Massachusetts, USA.

[2]Intel Xeon Gold 5120T CPU @ 2.20GHz, 512 GB RAM, running on 64-bit Windows operating system.

[3]32 vCPUs Intel Xeon E5-2686v4 @ 2.7 GHz, 244 GB RAM, running on Ubuntu Linux operating system.

TABLE I
ACHIEVED SIMULATION SPEED PER FREQUENCY POINT AND PER GPU

Test case	Platform	Unit Blocks	DOFs [$\times 10^6$]	B-matrix construction [s]			Cascading [s]		Total time [s]		Speed-up
				Schur (CPU)	DS (CPU)	DS (GPU)	CPU	GPU	CPU	GPU	CPU / GPU
LSAW	p3.8xlarge	11	3.8	313	158	33	752	133	919	175	5.3
TCSAW	p3.8xlarge	17	17.2	567	398	74	1652	269	2049	354	5.8
XBAR	p3.8xlarge	12	12.9	498	378	89	4914	469	5329	595	8.9
LSAW	workstation	11	3.8	319	263	74	1218	171	1550	258	6.0
IHP	workstation	11	10.8	442	447	113	2335	293	2806	430	6.5
XBAR	workstation	12	12.9	573	563	144	4024	534	4622	712	6.5

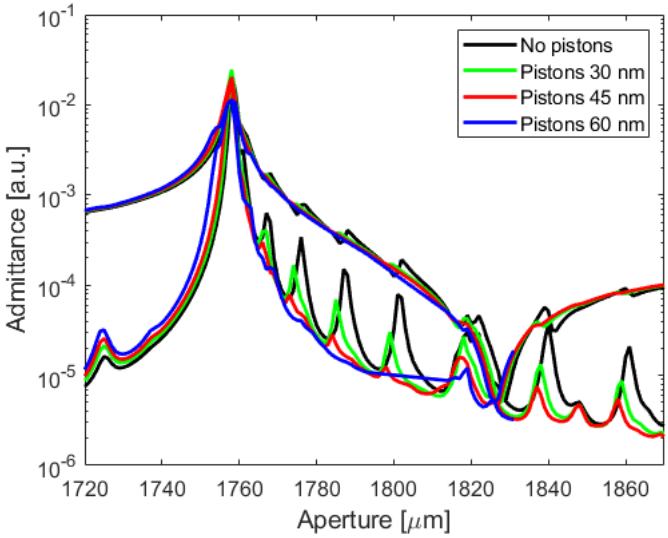

Fig. 3. Harmonic admittance of the TCSAW array vs piston thickness.

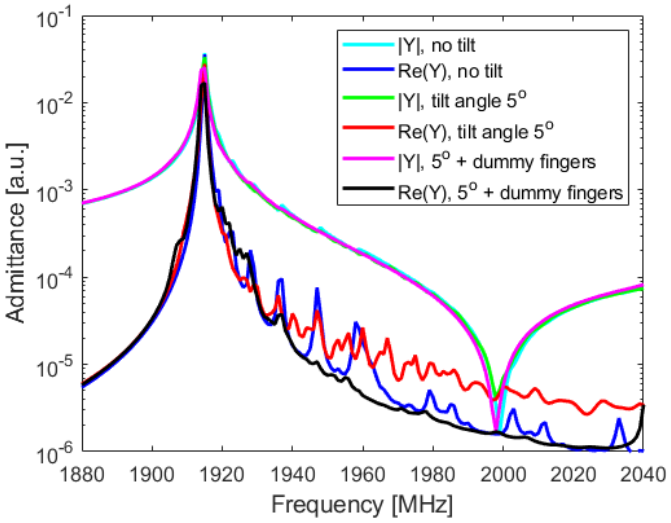

Fig. 4. Harmonic admittance in a straight and a tilted IHP unit cell, and in a tilted IHP unit cell with 2 μm dummy fingers.

A. Transverse Mode Suppression in TCSAW

As a TCSAW test case, suppression of transversal modes with piston masses [1] is considered. A TCSAW array was modeled on 128^{o}YX-cut $LiNbO_3$, with pitch 1 μm, mark-to-pitch ratio 0.5, aperture 40 μm, 150 nm thick Cu electrodes and 650 nm thick SiO_2 coating. Copper piston masses with length 1.1 μm and thickness varying from 0 to 60 nm were simulated. The results are shown in Fig. 3. The strong spurious modes present in the pistonless structure are gradually suppressed with an increasing piston mass (30 nm and 45 nm). Too large piston mass (60 nm) spoils the resonance Q-factor.

B. Transverse Mode Suppression in IHP Structure

Last year, Iwamoto *et al.* [7] demonstrated suppression of transversal modes in IHP devices on a multilayered 50YX-cut $LiTaO_3$ (0.3λ) / SiO_2 (0.3λ) / Si substrate. They showed that a conventional straight resonator structure exhibits transverse modes, which can be suppressed using a 5^o tilted structure. Here, similar conventional and a tilted arrays are considered with wavelength $\lambda = 2$ μm, $h/\lambda = 8\%$ Al thickness, metallization ratio 50%, and aperture $W = 40$ μm. Moreover, intriqued by the dummy electrodes visible in Fig. 8 of Ref. [7], a third variant is also modeled, where 2 μm dummy electrodes were added to the tilted resonator.

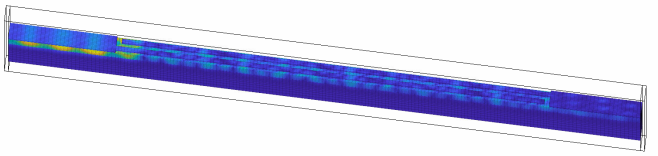

Fig. 5. Visualization of power flow in the tilted IHP unit cell at 1970 MHz, revealing acoustic leakage to the busbar in the direction of the tilting.

The results are compared in Fig. 4. While the presence of transversal modes in the straight resonator and their suppression in the tilted resonator are very clear, the conductance in the tilted resonator is substantially higher, implying increased losses. Visualization of the acoustic power flow in Fig. 5 identifies the losses as strong radiation to the busbars. The dummy fingers effectively solve the problem.

C. Transverse Modes in XBAR Structure

An XBAR unit cell on 500 nm thick YX-cut $LiNbO_3$ platelet [4] was chosen to demonstrate simulation of a varying geometry with hierarchical cascading. The unit cell has

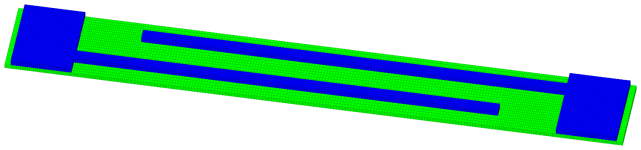

Fig. 6. Computational mesh for simulating an XBAR unit cell.

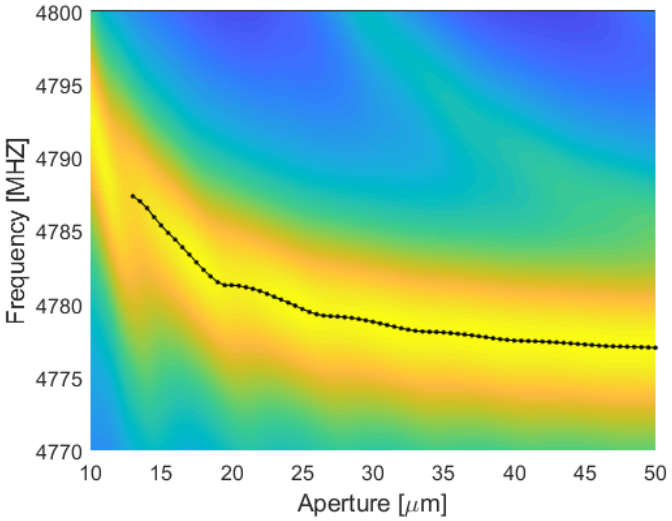

Fig. 7. The simulated conductance of the XBAR unit cell as functions of aperture and frequency (color map), and the extracted resonance frequency (black dots). Logarithmic color mapping has been applied to improve the contrast of the conductance variations.

periodicity λ = 6.5 μm, 500 nm thick and 0.89 μm wide Al electrodes, and an aperture varying in the range 10...50 μm, see Fig. 6. The computational meshes for all different apertures are composed of the same unit blocks. Consequently, many cascading operations are common to all variations, and it is more efficient to solve all of them in the same simulation run, instead of doing a separate simulation run for each aperture value. While the solution time for a single geometry was about 12 minutes per frequency point, each geometry variation added about 50 seconds to the simulation time.

The simulated conductances around the main resonance are shown in Fig. 7. The fitted resonance frequencies as a function of aperture are also shown in the figure as an overlay. The harmonic admittance for a single aperture 30 μm is shown in Fig. 8. Also, a comparison to 2D simulation with high-density computational mesh is shown, pronouncing the spurious resonances of transversal origin.

IV. Discussion

In this paper, GPU-assisted hierarchical cascading has been used to accelerate 3D FEM analysis of periodic acoustic arrays, using four test cases and two high-performance computational platforms. GPU-assisted cascading is a powerful tool but the limited memory available on GPUs significantly restricts the size and accuracy of the FEM models. This limitation can be relaxed somewhat by using GPUs only for

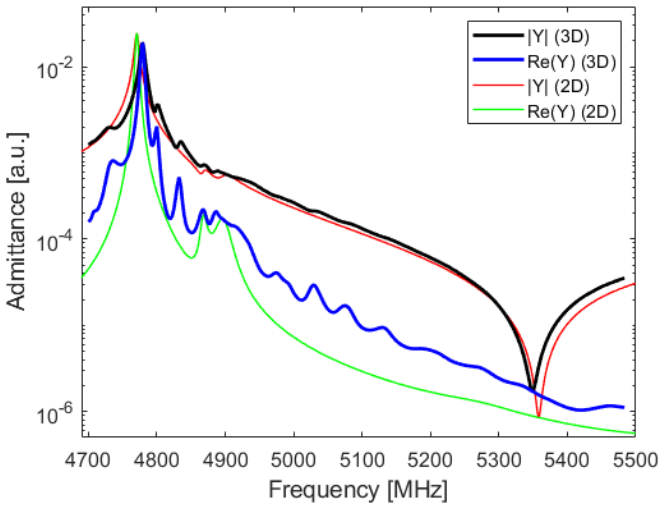

Fig. 8. Harmonic admittance of the XBAR array for aperture 30 μm.

isolated cascading operations. Although suboptimal from a GPU efficiency point-of-view, 5-10 times acceleration was found as compared to conventional CPU-based approach. A GPU-friendly alternative to the Schur complement method was presented. This new method could be further accelerated by taking advantage of the repeated element substructure typical to FEM models, but this was not attempted here. In overall, 5-9 times faster simulation times were achieved with GPU as compared to CPU. The numbers depend on the hardware and the applied parallel-processing configuration; much larger acceleration factors were found in prior experiments with Amazon cloud p3.2xlarge instance. Despite the impressing acceleration factors achieved, simulation times remain very long and limit the practical usefulness of the method.

References

[1] B. Abbott, R. Aigner, A. Chen, K. Gamble, T. Kook, J. Kuypers, M. Solal, and K. Steiner, "High Q TCSAW," in Sixth International Symposium on Acoustic Wave Devices for Future Mobile Communication Systems, Chiba University, Japan, 2015.

[2] M. Kadota and S. Tanaka, Solidly mounted ladder filter using shear horizontal wave in $LiNbO_3$, Proc. IEEE Freq. Cont. Symp., pp. 361-364, 2016.

[3] T. Takai, H. Iwamoto, Y. Takamine, H. Yamazaki, T. Fuyutsume, H. Kyoya, T. Nakao, H. Kando, M. Hiramoto, T. Toi, M. Koshino, N. Nakajima, Incredible high performance SAW resonator on novel multi-layered substrate," in 2016 IEEE Ultrasonics Symposium, 2016.

[4] V. P. Plessky, S. Yandrapalli, P. J. Turner, L. G. Villanueva, J. Koskela, and R. B. Hammond, "5 GHz laterally-excited bulk-wave resonators (XBARs) based on thin platelets of lithium niobate," Electronics Letters 55(2), November 2018, DOI : 10.1049/el.2018.7297.

[5] J. Koskela and V. P. Plessky, "Hierarchical cascading in FEM simulations of SAW devices, in 2018 IEEE International Ultrasonics Symposium, 2018.

[6] X. Li, J. Bao, Y. Huang, B. Zhang, T. Omori, and K. Hashimoto, "Use of hierarchical cascading technique for FEM analysis of transverse mode behaviours in surface acoustic wave devices, in 2018 IEEE International Ultrasonics Symposium, 2018.

[7] H. Iwamoto, T. Takai, Y. Takamine, T. Nakao, T. Fuyutsume, and M. Koshimo, , "Transverse modes in I.H.P. SAW resonator and their suppression method," in 2018 IEEE International Ultrasonics Symposium, 2018.