Listener-Position Adaptive Crosstalk Cancelation Using A Parameterized Superdirective Beamformer

Xiaohui Ma[†]
Dynaudio A/S
8660 Skanderborg, Denmark
xma@dynaudio.com

Christoph Hohnerlein
Berlin Institute of Technology
Ernst-Reuter-Platz 7, 10587 Berlin, Germany
christoph.hohnerlein@qu.tu-berlin.de

Jens Ahrens
Chalmers University of Technology
412 96 Gothenburg, Sweden
jens.ahrens@chalmers.se

***Abstract*—We present a method for cancelation of the acoustic crosstalk between the ears of a listener using a linear loudspeaker array. This allows for controlling the signals that arise at listener's ear independently of each other so that binaural cues can be imposed. We employ superdirective near-field beamforming (SDB) over the vast part of the frequency range. Allowing the listener to move in front of the array requires the computation of a new SDB solution for each possible position of the listener, which is a computationally expensive procedure. We show that the beamformer weights exhibit a very smooth evolution for listening positions along a line parallel to the array. This makes it straightforward to parameterize pre-computed beamformer weights with a few parameters for each frequency. Upon realtime execution, the beamformer weights can be determined efficiently for any arbitrary position along the listener contour from the parameters with negligible error. Simulation results show that the proposed beamformer provides a higher channel separation than previously published solutions while maintaining robustness.**

I. Introduction

Binaural audio reproduction provides the users an immerse listening experience, when the signals are encoded with head-related transfer functions. A fundamental requirement for such rendering systems is high channel separation between the two ears. Ideally, the signal intended for the ipsilateral ear will not be received by the contralateral ear. Binaural audio playback through headphones can be found in many applications, such as VR games, due to the negligible crosstalk between the ears. However, headphone based reproduction is in many ways unfavourable due to reasons like social separation, comfortability, and head internalization of sound [1], etc. Loudspeakers in this sense could be a good alternative. To make the binaural audio work for loudspeaker systems, the crosstalk between the two ear channels should be carefully eliminated.

In a two-loudspeaker setup, crosstalk cancelation (CTC) can be achieved by system inversion [1], [2], which can easily break down in the presence of even only small deviations from the assumptions. The obtained inverse filters are also subject to large errors around ill-conditioned frequencies [3]. A lot of efforts have been applied to improve the system robustness, such as using frequency-dependent regularization [1] and optimizing the loudspeaker positions [4]. Alternatively, recursive ambiophonic crosstalk elimination (RACE) proposed by Glasgal [5] provides simple means of CTC for two loudspeaker symmetric setup that is surprisingly robust with respect to head movement. However, the performance can strongly depend on loudspeaker position and even loudspeaker model.

Though the robustness with respect to the aforementioned limitations can be improved, the listeners are constrained in a narrow sweet spot, and CTC can easily break down outside the optimized position. It is therefore favourable to achieve CTC for a larger area or for moving listeners. Recent research using multiple (more than two) loudspeakers has shown promising results. Bauck [6] uses a multi-way loudspeaker array system to enlarge the sweet spot. Takeuchi and Nelson [3] proposed the optimal source distribution, which greatly improves the robustness in terms of room reflections and misalignment of the system. Hohnerlein and Ahrens [7] employ least-squares frequency-invariant beamforming to achieve CTC that is robust with respect to small head movements. This approach is the basis for the work presented in this paper.

Adaptive CTC updates the CTC filters according to the instantaneous listener position, which is being tracked. Cecchi *et al.* [8] proposed an extension of RACE applied to asymmetric two loudspeaker setups, where the delays and attenuations for each channel are updated based on the real-time position. Gálvez *et al.* [9] implement dynamic CTC by combining a fixed CTC filter with a delay-and-sum beamformer that steers towards the listener. The listener can only move along a circular trajectory, and the achievable crosstalk cancelation is limited.

The present paper introduces a dynamic CTC using a linear loudspeaker array. We employ superdirective beamforming to achieve high amounts crosstalk cancelation. The pre-computed beamformer weights are parameterized, and the CTC filters are then computed from the parameters in real-time.

II. Methods

A. Least-squares frequency-invariant beamforming

For a linear array with N equispaced loudspeakers, the directional response of a filter-and-sum beamformer is [10]

$$B(\omega, \vec{r}) = \sum_{n=0}^{N-1} W_n(\omega) \frac{1}{r_n} e^{-j\omega \frac{r_n}{c}}, \qquad (1)$$

[†] Also at Department of Engineering, Aarhus University, Finlandsgade 22, 8200 Aarhus N, Denmark.

$$r_n = ||\vec{r} - \vec{x}_n||, \tag{2}$$

where $\vec{r}$ is a position at which the beamformer response $B(\omega, \vec{r})$ is prescribed, $\vec{x}_n$ is the position of the n-th driver, $W_n(\omega)$ is the frequency response of the beamforming filter for the n-th driver, and c is the sound speed in air.

Least-squares (LS) beamforming optimally approximates a target response $\hat{B}(\omega, \vec{r})$ by $B(\omega, \vec{r})$ in the LS sense. If the target directional response is frequency independent, i.e. $\hat{B}(\omega, \theta) = \hat{B}(\theta)$, the beamformer is called least-squares frequency-invariant beamforming (LSFIB) [11].

Combining all prescribed $\vec{r}_m$ $(m = 1, \cdots, M)$, the beamformer response in Eq. (1) can be reformed as

$$\boldsymbol{b}(\omega) = \boldsymbol{G}(\omega)\boldsymbol{w}_f(\omega), \tag{3}$$

where $[\boldsymbol{G}(\omega)]_{mn} = e^{-j\omega\frac{r_{mn}}{c}}/r_{mn}$, $r_{mn} = ||\vec{r}_m - \vec{x}_n||$, and $[\boldsymbol{w}_f(\omega)]_n = W_n(\omega)$.

The filter responses $\boldsymbol{w}_f(\omega)$ can be determined by minimizing the squared error between the predicted and desired directional responses.

$$\min_{\boldsymbol{w}_f(\omega)} ||\boldsymbol{G}(\omega)\boldsymbol{w}_f(\omega) - \hat{\boldsymbol{b}}||_2^2. \tag{4}$$

To achieve CTC, the directional response of a beamformer should have the mainlobe in the direction of the illuminated ear, and the energy in the shadowed ear direction is minimized [7]. This can be accomplished by adding an energy constraint in the shadowed ear direction,

$$||\boldsymbol{G}_s(\omega)\boldsymbol{w}_f(\omega)||_2^2 \leq e, \tag{5}$$

where $\boldsymbol{G}_s(\omega)$ is a subset of $\boldsymbol{G}(\omega)$ containing the directions around the exact crosstalk path to allow for a smooth pressure transition, e determines the energy attenuation. The separation between the illuminated and shadowed ears with respect to the array center is set to be 10° at a distance of 1 m. This optimization problem can be solved using the CVX toolbox [12]. This approach is identical to the one employed in [7]. Results of the user study presented *ibidem* show localization accuracy comparable to headphone auralization of binaural signals.

B. Adaptive crosstalk cancelation

Crosstalk cancelation implemented using LSFIB has been shown to be robust in terms of small head movements up to 5 cm [7]. It will obviously collapse for larger head movements due to the relatively small distance between the ears. It is therefore desirable to update the beamformer in real-time according to the listener's position, i.e. to perform adaptive crosstalk cancelation. We assume for convenience in the remainder of this paper that the listener moves along a straight line parallel to the linear loudspeaker array in use as depicted in Fig. 1.

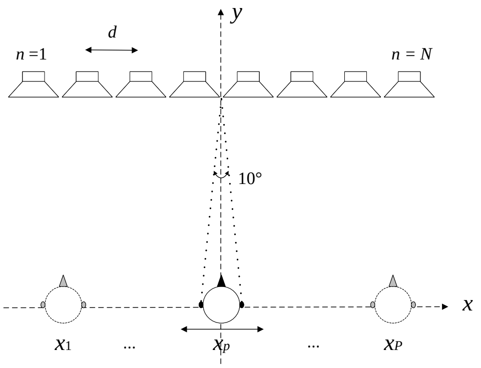

Fig. 1. System geometry. The listener moves along a straight line 1 m from the array. The angle between left and right ear to the array center is 10° and assumed to be constant along the moving track. The central listening position is at the coordinate origin.

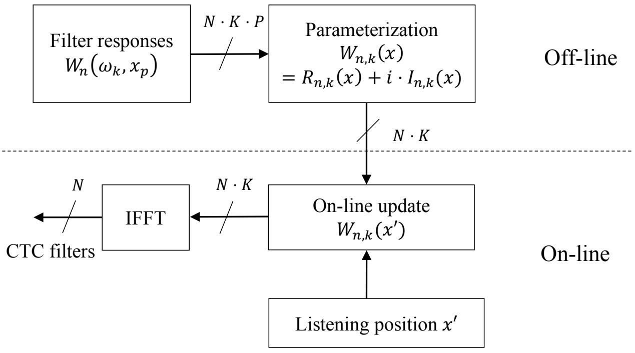

Fig. 2. Signal flow of the proposed approach.

As LSFIB is computationally very expensive, we propose adaptive CTC that employs an off-line modelling phase together with an on-line updating phase. The algorithm is illustrated in Fig. 2. In the off-line modelling phase, the designated contour of possible listener positions is discretized into P positions with a resolution of 5 mm. For each frequency ω_k, $k = 1, \cdots, K$, the beamformer weights are calculated for all positions $W_n(\omega_k, x_p)$, $p = 1, \cdots, P$. A typical evolution of the beamformer weights as a function of listening position is depicted in Fig. 3. We observe that the evolution is smooth and fairly periodic. We therefore employ a sum of sine functions to parameterize the beamformer weight. We perform this separately for the real and imaginary parts as functions of the listener position:

$$\begin{aligned} R(\omega_k, x) &= \Re\{W_n(\omega_k, x)\} \\ &= \sum_{m=1}^{M} A_m(\omega_k)\sin[B_m(\omega_k)x + C_m(\omega_k)]. \end{aligned} \tag{6}$$

$$\begin{aligned} I(\omega_k, x) &= \Im\{W_n(\omega_k, x)\} \\ &= \sum_{m=1}^{M} D_m(\omega_k)\sin[E_m(\omega_k)x + F_m(\omega_k)]. \end{aligned} \tag{7}$$

The coefficient set $\{A_m(\omega), \cdots, F_m(\omega)\}$, $m = 1, \cdots, M$, is then used to calculate the actual CTC filters in real-time. In total there are $6M \cdot K \cdot N$ coefficients need to be stored.

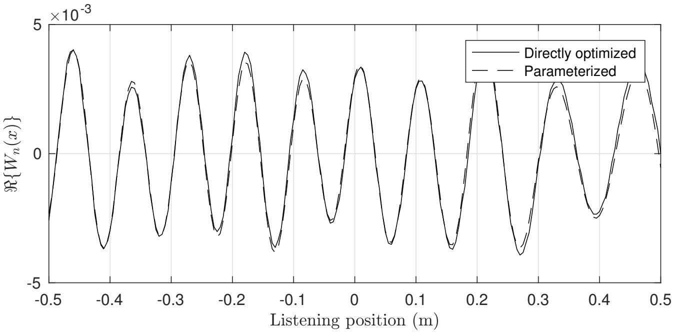

Fig. 3. Sine parameterization of the beamformer weights. The real part of the beamformer weights of loudspeaker 1 at 7.4 kHz is shown by the solid curve; the parameterized beamformer weights are given by the dashed curve.

Crosstalk cancelation based on LSFIB is limited at low frequencies because the beams tend to broaden up causing more crosstalk. Spatial aliasing is the limiting mechanism at high frequencies. A hybrid solution is proposed in [7], where high frequency components, e.g. above 8 kHz, are rendered as stereo through the two loudspeakers at the ends of the array to evoke natural shadowing due to the user's head. The mid and low frequencies (250 Hz – 1000 Hz) are rendered through RACE using a pair of loudspeakers [5], and a single subwoofer is used to render the lowest frequency band.

III. Results

Simulations are performed on a linear equispaced loudspeaker array with 8 elements as depicted in Fig. 1. Each loudspeaker is modelled by a point source. The listener moves along a straight line 1 m from the array. The separation angle between the ears with respect to the array center is 10°, and assumed to be constant while the listener moves around.

The applicable frequency bandwidth is investigated for the central listening position (the origin of the coordinate system) with two array geometries: 10.0 cm and 15.2 cm loudspeaker spacing. The main lobe of the beamformer is steered towards the left ear, while a null is steered towards the right ear. To incorporate the physical uncertainties in reality, e.g. mismatch between the loudspeakers, variations in the loudspeaker placement, Gaussian noise

$$\Delta A \sim \mathcal{N}(0, 0.3) \quad \text{and} \quad \Delta\Phi \sim \mathcal{N}(0, 0.001\omega/c),$$

is added to the beamformer weights at each frequency for each loudspeaker, i.e.

$$W_n(\omega_n, x) = (|W_n| + \Delta A_n)e^{j(\angle W_n + \Delta\Phi_n)}. \tag{8}$$

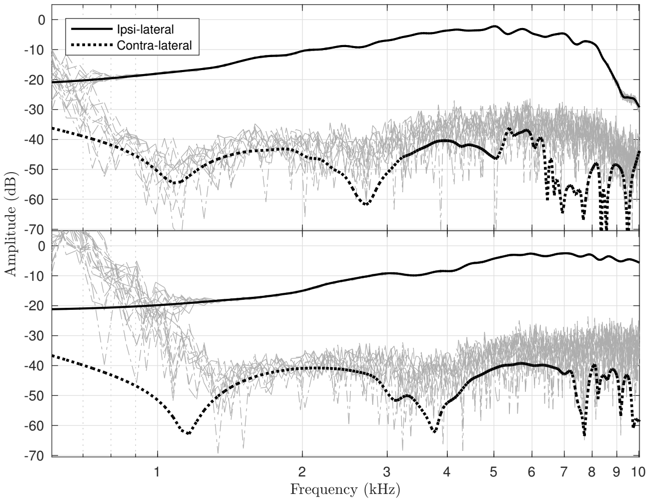

Fig. 4. Transfer function of the system to the user's ears. Thick black lines represent ideal results; thin lines show the results for 10 simulations with random noise applied to the beamformer weights. Top: array with 15.2 cm spacing; bottom: array with 10.0 cm spacing.

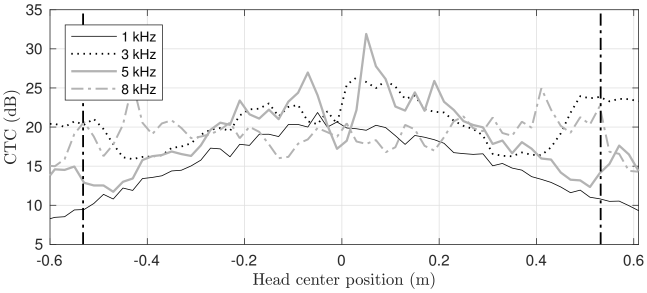

Fig. 5. Crosstalk cancelation at different listening positions at 1 m distance from the loudspeaker array. Results are obtained by averaging 20 random positions around the exact ear position; the head is modelled as a rigid sphere. Four frequencies within the applicable frequency range are shown. The vertical dash-dotted lines indicate the usable array aperture.

Simulated transfer functions from the array input to the ears of the user are presented in Fig. 4. Although both array geometries show similar channel separation over a wide frequency range for the ideal setup, mismatch reduces the usable frequency band differently for the different element spacings. The array with 15.2 cm spacing has a usable frequency range from 900 Hz to 8 kHz, where the channel separation is larger than 15 dB, which constitutes the lower boundary for binaural audio systems [13]. The array with 10.0 cm spacing has a usable frequency range from 1.5 kHz to 10 kHz. Large spacing can extend the low working frequency while small spacing can extend the high working frequency, as the spatial aliasing frequency is accordingly higher. No spatial aliasing is evident in the setups under consideration. We will assume the 15.2 cm spacing in the remainder of the paper.

The achievable performance of the position-adaptive CTC using LSFIB is simulated at discrete positions along the listening contour. The listener's head is modelled as a rigid sphere with a radius of 9 cm to take the scattering into consideration. Fig. 5 shows the results for the obtained CTC. At each listening position, the simulation averages the sound pressure level for 20 random positions around the exact ear position with distances up to 4 cm. It is noticeable the listener should not be located at $|x_P| \leq 0.35$ m, so that a channel separation of more than 15 dB is maintained over the entire beamforming frequency range.

Above presented results hold for the case that the beamformer weights $W_n(\omega_k, x)$ were obtained directly from the

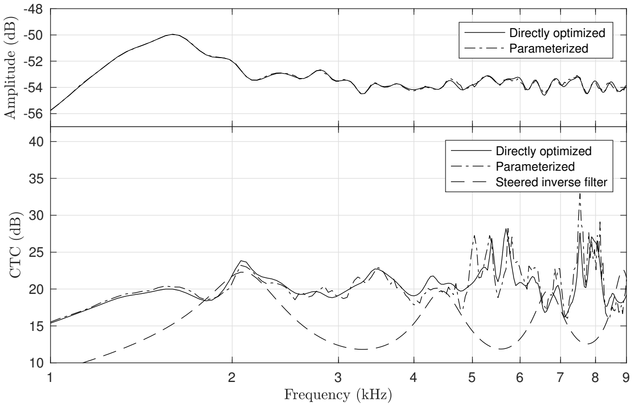

Fig. 6. Frequency response at the ipsilateral ear (top) and crosstalk cancelation (bottom) at the listener position at 0.2 m from the center, solid lines are results from direct optimization dash-dot lines are results from parameterization. As a comparison, CTC obtained from the steered inverse filter from [9] is shown by the dashed line.

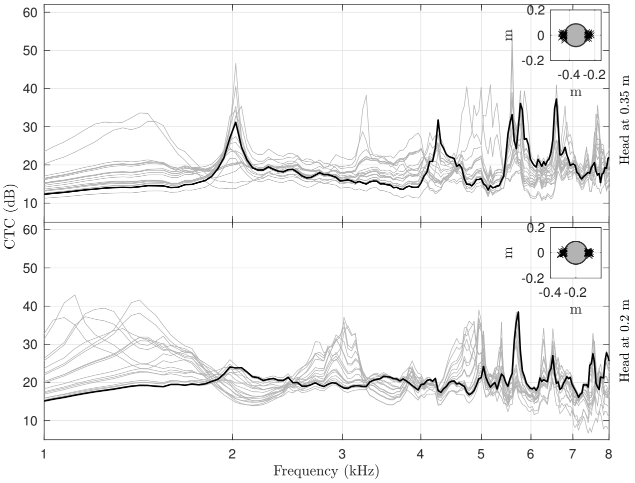

Fig. 7. Crosstalk cancelation at random ear positions up to 4 cm away from the exact ear positions. The thick black line shows the results at the ears; the thin gray lines show the results at 20 random position. The inset figure shows the head and the random evaluation position. Top: Listening position 35 cm from the central one; bottom: 20 cm from the central one.

optimization. Fig. 3 depicts exemplarily the real part of $W_n(\omega_k, x)$ for a given loudspeaker at a given frequency as a function of the listener location. The deviation of the curves sum-of-sines parameterization represented by Eq. (6) and Eq. (7) from the optimal data is small and is similar across loudspeakers and across different frequencies.

The performance of the proposed parameterized adaptive CTC is compared to CTC through direct optimization in Fig. 6, which shows the frequency response and obtained CTC for the listening position at 20 cm from the center. It can be observed that the proposed CTC gives approximately the same performance as the directly optimized CTC. Similar behaviors are found for other listening positions, however overall level differences are observed for different listening positions. This requires equalization to assure consistent perception when the listener moves along the listening contour. Similarly, the propagation delay for the two ears should also be carefully aligned.

Fig. 6 depicts a comparison of the performance of the proposed CTC to the approach presented in [9]. The obtained CTC for the depicted listening position at 20 cm from the center is ~5dB in average higher for the proposed approach. This can be attributed to the fact that the presented approach applies superdirective beamforming contrary to [9]. Further simulations show that the approach from [9] has an asymmetric CTC performance about the y-axis, i.e., CTC is significantly higher if the ipsilateral ear is facing the array center. Our proposed parameterization makes the computational cost of the presented approach similar to the one from [9].

System robustness with respect to accuracy of the positioning of the listener's head is investigated based on the listening position at 20 cm from the center. The head is modelled as a rigid sphere with radius of 9 cm; the two ears are on the ends the diameter parallel to the array. CTC for 20 random positions with deviations conforming to $\mathcal{N}(0, 0.015)$ around the exact ear position are depicted in Fig. 7. It can be seen that the loss in channel suppression is moderate so that we can conclude that the proposed CTC is robust in terms of small head movements. This finding is supported by the observations in the user study presented in [7], which investigated the non-parameterized beamformer for the central listening position. No fixation of the subjects' heads was required.

Always assuming an ear separation angle of 10° might not be a good estimate when the head approaches the array ends. When analysis CTC at random positions around the exact ear position, some ear position pairs might match the 10° separation angle better than the exact ones and therefore give better channel separation. This can explain the observation that CTC are sometimes higher for the random positions than that for the assumed ear positions.

Adaptive crosstalk cancelation based on polynomial beamforming [11], [14] is also implemented and simulated. Polynomial beamforming is designed with prototype look directions (PLD) distributed over the whole steering range. Large number of PLDs enhances the performance of polynomial beamforming but also increases the order of the polynomial post filters [11], which can also cause overfitting of the problem. Our results, which are not presented here, show that the obtained CTC is fragile with respect to small array geometry changes. We therefore assume the sine parameterization is a better choice in this use case.

IV. Conclusions

Adaptive crosstalk cancelation based on sine parameterization is proposed, which involves an off-line modelling phase and an on-line updating phase. Sum-of-sines is used to model the beamformer weights as a function the listening position, which is assumed to be a straight line parallel to the array. Our simulations show that the system is relatively robust and outperforms non-superdirective solutions while maintaining a comparable computational cost during runtime.

REFERENCES

[1] E. Choueiri, "Optimal crosstalk cancellation for binaural audio with two loudspeakers," in *Princeton University*, 2010.

[2] O. Kirkeby and P. A. Nelson, "Digital filter design for inversion problems in sound reproduction," *J. Audio Eng. Soc*, vol. 47, no. 7/8, pp. 583–595, 1999.

[3] T. Takeuchi and P. A. Nelson, "Optimal source distribution for binaural synthesis over loudspeakers," *The Journal of the Acoustical Society of America*, vol. 112, no. 6, pp. 2786–2797, 2002.

[4] D. B. Ward and G. W. Elko, "Effect of loudspeaker position on the robustness of acoustic crosstalk cancellation," *IEEE Signal Processing Letters*, vol. 6, no. 5, pp. 106–108, 1999.

[5] R. Glasgal, "360 localization via 4.x race processing," in *Audio Engineering Society Convention 123*, Oct 2007.

[6] J. Bauck, "A simple loudspeaker array and associated crosstalk canceler for improved 3d audio," *J. Audio Eng. Soc*, vol. 49, no. 1/2, pp. 3–13, 2001.

[7] C. Hohnerlein and J. Ahrens, "Perceptual evaluation of a multiband acoustic crosstalk canceler using a linear loudspeaker array," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 96–100.

[8] S. Cecchi, A. Primavera, M. Virgulti, F. Bettarelli, J. Li, and F. Piazza, "An efficient implementation of acoustic crosstalk cancellation for 3d audio rendering," in *2014 IEEE China Summit International Conference on Signal and Information Processing (ChinaSIP)*, July 2014, pp. 212–216.

[9] M. F. Simón Gálvez, T. Takeuchi, and F. M. Fazi, "Low-complexity, listener's position-adaptive binaural reproduction over a loudspeaker array," *Acta Acustica united with Acustica*, vol. 103, no. 5, pp. 847–857, 2017.

[10] H. L. V. Trees, *Optimum array processing: Part IV of detection, estimation, and modulation*. New York: Wiley, 2002.

[11] E. Mabande, M. Buerger, and W. Kellermann, "Design of robust polynomial beamformers for symmetric arrays," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 1–4.

[12] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," Website, 2014, http://cvxr.com/cv.

[13] Y. L. Parodi and P. Rubak, "A subjective evaluation of the minimum channel separation for reproducing binaural signals over loudspeakers," *J. Audio Eng. Soc*, vol. 59, no. 7/8, pp. 487–497, 2011.

[14] M. Kajala and M. Hamalainen, "Filter-and-sum beamformer with adjustable filter characteristics," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 5, 2001, pp. 2917–2920 vol.5.