Spectral Delay Filters*

Miguel Aguilar, *AES Member* **AND María Inmaculada Mohíno,** *AES Fellow*
(m.aguilar@edu.uah.es) (inmaculada.mohino@edu.uah.es)

University of Alcalá, Alcalá de Henares, Spain

In a real scenario, the audio processing can be used in several applications, since it contains useful infomation. This information is composed of audio events which are grouped according to its nature. Each group shares different features. In the present study, different events have been classified, such as: microwave, hair dryer, plate sorting, vacuum cleaner, stirring cup, speech and water tap. In order to classify and detect audio, Toni Heittola's algorithm published in 2016 is used, which was designed to detect and classify acoustic events in real environments. Different position microphone configurations were tested in order to establish if the classification is improved using a metaclassifier by majority. The variable parameters of the acoustic scene were coefficient of reflection and signal-to-noise ratio. The results of the simulations predict a better classification when more than one microphone is used, that is, more than one classifier.

0 INTRODUCTION

Detection and classification of acoustic events is necessary to analize the scenario's acoustic scene. The goal is develop a system able to recognise different classes of acoustic events. It is the same task that speech recognition systems achives when it distinguishes between phonemes. In general, there are two main steps: feature extraction and classification. The first one is to find features which are useful to distinguish one acoustic event class from another. For example, Mel-Frequency Cepstrum Coefficients (MFCC) are features that emulate the non-linearity response of the human sense of hearing [1]. The features extracted are used to train a model that is able to characterise the presence of subpopulations within an over population.

The aplications of this kind of systems are numerous. It is proposed in [2] to apply this thecniques to develop systems able to attend people with hearing disabilitys. It is necessary to use the information that contains the acoustic events. So first of all, microphones are used to pick up the sound of the house/room. After that, classification methods are applied to this sound.

In the present work, seven acoustic events are classified and two cases are compared. This events are: microwave, hair dryer, plate sorting, vacuum cleaner, stirring cup, speech and water tap. The first case uses one microphone and one classifier. The second case works with four microphones, four classifiers and a metaclassifier by majority. The goal is to check if increasing the number of microphones and classifiers will improve the results. The room conditions are simulated by modified version of RIR algorithm (McGovern [3]) with microphone directivity. The classifier applied to analize the sounds is the algorithm develop by Toni Heittola to carry out a challenge about sound event detection in real life audio proposed by the conference DCASE2016 [4] (Detection and classification of acoustic scenes and events).

1 CLASSIFICATION PROCEDURE

As previously said, Toni Heitola's algorithm is used to classify the audio picked up by each microphone. It obtains MFCCs and applies a classifier based on Gaussian Mixture Models. In this case that there are more than one microphone, fusion approches using majority voting is employed.

1.1 Feature extraction

MFCC's provide a compact representation of the signal spectra, concentrating the grater part of its energy in the first coefficients. The non-linearity response of the human sense of hearing is emulated by applying mel filterbank to the power spectrum. In order to characterize dynamic behavior of the signal, Delta (ΔMFCC) and Delta-Delta ($\Delta(\Delta$MFCC)) are obtained.

The steps to calculate the coefficients are shown in Figure 1. First, the signal is splitted in short time periods, in wich it can be considered stationary. Discrete Fourier Transform is applied to each period to obtain the energy. After that, the energy of each period is multiplied by the

*To whom correspondence should be addressed Tel: +1-240-381-2383; Fax: +1-202-508-3799; e-mail: info@schtm.org triangular filter bank. The process ends applying the logarithm and the discrete cosine transform, in order to transform the coefficients to the frequency domain. MFCC are widely explained in [5].

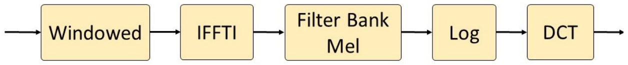


Fig. 1. Mel-Frequency Cepstral Coefficients

Once MFCCs are calculated, statistics are determined from differential values. Those are ΔMFCC (velocity) and Δ(ΔMFCC) (acceleration).

In this algorithm, 20 coefficients are calculated, so 20 velocitys and 20 accelerations are obtained. MFCC0 is omitted because, effectively, 0th coefficient is a accumulation of average energies of all frequency bands [6]. To carry out the first step, 40 ms window length is employed with 50% hop size).

1.2 Classifier based on Gaussian Mixture Model

Gaussian Mixture Model (GMM) is used to classify the acoustic signals. In order to initialize weights of system information, k-means algorithm is employed. For the purpose of optimizing the estimation of system parameters, Expectation-Maximization (EM) algorithm is applied.

K-means. It's a unsupervised algorithm that classifies an amount of data separating it in Z subgroups. Each subgroup is associated to one centroid. After that, the centroids are recalculated as baricenters until the error satisfies the condition imposed.

GMM. Gaussian Mixture Model [7] is a parametric probability density function represented as sum of gaussian components with associated weights, Equation [1].

$$p(u|\Theta) = \sum_{i=1}^{Z} v_i \mathcal{N}(u|\mu_i, \xi_i) \tag{1}$$

where u is the ρ-dimension data vector, $\mathcal{N}(u|\mu_i, \xi_i)$ are the Gaussian components and v_i, $i = [1,...,Z]$, the associated weights. μ_i is the mean vector and ξ_i the covariance matrix. In Equation [2], the parameters of each Gaussian Mixture Model are shown.

$$\Theta = \{v_i, \mu_i, \xi_i\},\ i = [1,...,Z] \tag{2}$$

EM. It's a maximum likelihood parameter estimation of distribution from a known data set when this data is incomplete or has missing values [8]. Firt of all, initial parameters $\Theta^{[r=0]}$ are chosen in order to start the r-th iterations, $(r = 1,2,...)$. Then non-observed parameters are estimated from current parameters $\Theta^{[r]}$ (E-step). After that, parameters $\Theta^{[r+1]}$ are estimated using maximum likelihood estimation (M-step). The r^{th} iteration ends checking if $\Theta^{[r]} \simeq \Theta^{[r+1]}$. If that condition is not met, continues from E-step $([r+1]^{th}$ iteration).

In the case of Toni Heitolla's algorithm, one GMM binary classifier is trained for each acoustic event class using real recordings. Sixteen Gaussian distributions are used, so $Z = 16$. Cross validation methods [9] are applied to train the classifier, in order to obtain realistic results and avoid overtrainning.

1.3 Metaclassifier by majority voting

Metaclassifier by majority voting [10] is a system that evaluates the results obtained for the same task in different systems. That is, it compares and combines the outputs to merge them. The criterion chosen is majority voting. The final result depend on the number of systems that obtains the same output.

In this case, the outputs of several classifiers are fusioned as shown in Figure 2. The condition imposed to determine the output of this metaclassifier is that more than two classifiers obtain the same result at the same segment w. In the example is shown with four microphones, so four classifiers.

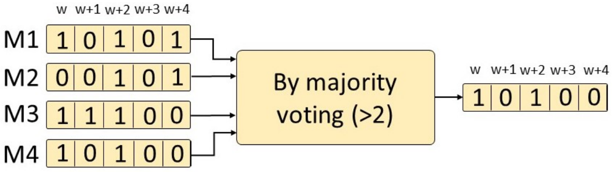


Fig. 2. Metaclassifier by majority

1.4 Evaluation metrics

Two different metrics are used [11]. F1-score evaluates the accuracy of the classifier and Error Rate tests system's precision. To avoid over-optimistic results, cross-validation (k-fold) methods are applied.

F1-score. The results are divided into groups as its defined in Table 1. This groups are: True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN).

Table 1. Confusion matrix for binary classifiers

Event class	**Recognize as positive**	**Recognize as negative**
Positive	TP	FN
Negative	FP	TN

Then, precision (Equation 3) and recall (Equation 4) are calculated to obtain F1-score, Equation 5.

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

$$F = \frac{2PR}{P + R} \tag{5}$$

Error Rate. The result is splited in substitutions, deletions, and insertions. Being the substitutions (S) when the system obtains FP and FN for the same segment. The rest of FN are deletions (D) and the rest of FP are insertions (I). Error rate is calculated with Equiation 6.

$$ER = \frac{\sum_{k=1}^{N} S(k) + \sum_{k=1}^{N} D(k) + \sum_{k=1}^{N} I(k)}{\sum_{k=1}^{N} N(k)} \qquad (6)$$

N is the number of segments.

2 SIMULATION

The goal of this study is check if applying fusion tecniques and using four microphones will improve the results. In this section is explained the process followed to build the database. It is necessary to create one database with the conditions imposed (Signal-to-Noise Ratio and Reflexion coefficient) for each microphone. The acoustic scene is designed, chosing the position of the sources (events) and the microphones. The variables and his range are defined, so are the experiments.

2.1 Database

The isolated acoustic events chosen previously were extracted from two databases. Speech event from TIMIT [13] and the rest of the events from [14]. In Table 2, the number of each event and the average length is shown. In order to use the same sample rate of 44.1 kHz for all events, speech were resampled from 16 kHz and its power were decreased 5 dB.

Table 2. Acoustic events

Event	Number sounds	Average length (s)
Hair dryer	66	2.0057
Microwave	92	2.012
Plates sorting	135	1.4678
Speech	200	4.0
Stirring cup	59	2.014
Vacuum cleaner	79	2.0009
Water tap	114	2.0063

2.1.1 Acoustic scene design

The room taken as a template for this study, is the laboratory of department of theory signal of Alcalá 's University. The measurements obtained are $[x\ y\ z] = [8\ m\ 7\ m\ 2.5\ m]$. After that, furniture that can be found in any residence are distributed, like a couch, a carpet or a table. In Figure 3, the acoustic scene obtained is shown. Fixed positions are allocated to the events consistently with the situation of the elements previously defined, Table 3.

2.1.2 Microphone positions

In this paper three microphone configurations are proposed, shown in Figure 4. The first one has just one microphone situated on the mid and uppersection of the room. This position is taken as the best with only one microphone. The other two configurations have four microphones situated in the room. One of them in the upper corners and the other in the mid of upper edges. These configurations were selected to compare the classification between using one and four microphones (metaclassifier by majority voting).

Table 3. Event's positions

Event	Position [x y z] (m)
Hair dryer	[6 1 1.7]
Microwave	[7 7 0.9]
Plates sorting	[7.7 3.5 1]
Speech	[1.5 1.5 1.6]
	[4 1 1.6]
	[2 6 1.2]
Stirring cup	[3 3 1]
Vacuum cleaner	[2 5.5 0.3]
Water tap	[7.7 3 1.2]

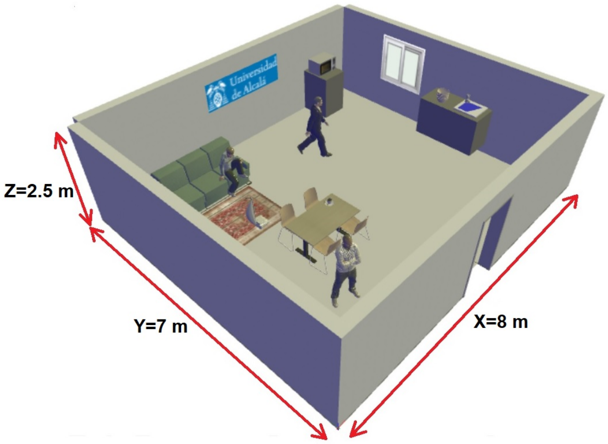


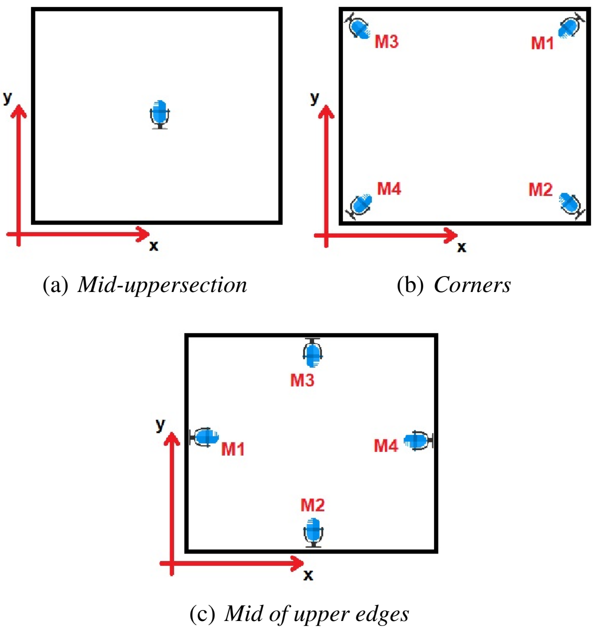
Fig. 3. Acoustic scene

2.2 Generate synthetic audio frames

At this point, it had been defined microphone positions, sources positions and room dimension. DCASE algorithm uses real recordings to train and test the classifier, that is, three-minute frames with multiple acoustic events. Our task is to create synthetic audio frames containing background noise and our acoustic events. In order to generate different room conditions, reflexion coeficient and signal-to-noise ratio are changed in each experiment.

2.2.1 Room impulse response calculation

The asociated position of each event class is simulated by modified version of RIR algorithm (McGovern [3]), directivity microphone parameter is added, with inputs: sample rate, microphone position, reflexion coefficient, room dimension, source position, main lobe to back lobe ratio and steering direction of the microphone. The microphones are always steered to the very center of the room. The impulse response h[n] is calculated with this algorithm. After that, every event class is filtered with its impulse response. This is how simulated audio of each microphone is obtained.



(a) *Mid-uppersection*

(b) *Corners*

(c) *Mid of upper edges*

Fig. 4. Microphone configurations

2.2.2 Procedure

One minute audio frames are generated. Summary of the procedure is shown in Figure 5. First of all, power of additive white Gaussian noise (AWGN) is defined with repect to average power of the whole isolated events. Then, one minute frames of AWGN with the power previously defined is generated. The sample rate used is 44100 Hz. The next step is to add a random number of filtered acoustic events to this frame. The whole process is randomize, so between four and eight events are added to each frame. This process ends when the whole events are added and the pertienent documentation is generated. The onset of the events in a noise frame depends on the number of events that are added. For example, in Equation 7 the onset in seconds of the event a of the frame x, it's known that E events are added and frame's length L_{wav} is sixty seconds, is shown.

$$Onset_{ax} = \frac{L_{wav}}{E}(a-1), \quad 1 \leq a \leq E \tag{7}$$

The same random vectors are employed for the experiments in order to compare the results.

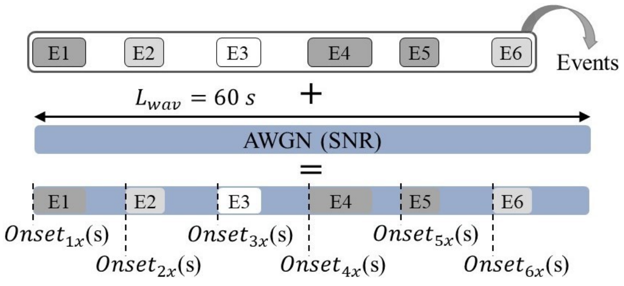


Fig. 5. Generate synthetic audio frames procedure

2.2.3 Experiments

The range of the reflexion coefficient and signal-to-noise ratio is defined as follows in order to test how affect its changes to the systems. $SNR = [20\,dB\;30\,dB\;40\,dB\;50\,dB]$ and CR=[0.1 0.3 0.6]. Each experiment is composed of nine microphones so nine classifiers. Metaclassifier by majority voting is applied in both four-microphone configurations. As previously said, seven acoustic events are classified, that is: microwave, hair dryer, plate sorting, vacuum cleaner, stirring cup, speech and water tap. A new database is generated every experiment, with a pair of values of CR and SNR.

3 RESULTS

The output of the algorithm is composed of two different evaluation modes. Those are segments results and event results. In this paper, segments results are employed. The main evaluation metric of the challenge is Error Rate, so it is chosen to evaluate the results. F1-score brings almost the same information.

In Table 4, Total Error Rate of the three configurations is shown for noisy situation (SNR=20 dB). The worst results are reached by the one-microphone configuration for all possible values of reflexion coefficient. It is observed how results are improved since CR is increased. The results of four-microhone configurations are very similar.

Table 4. Total Error Rate for configurations with SNR=20 dB

Configurations	CR		
	0.1	**0.3**	**0.6**
Mid-uppersection	1.42	1.30	1.15
Corners	1.41	1.19	1.00
Mid of upper edges	1.46	1.19	0.91

Respect to Figure 6, Total Error Rate for all configurations is shown. The results get worse while the noisy level increases. The worst results for all values of noise that have been studied is obtained by *mid-uppersection* configuration. On the other hand, four-microphone configurations obtain almost the same results. In general, corners configuration is the best.

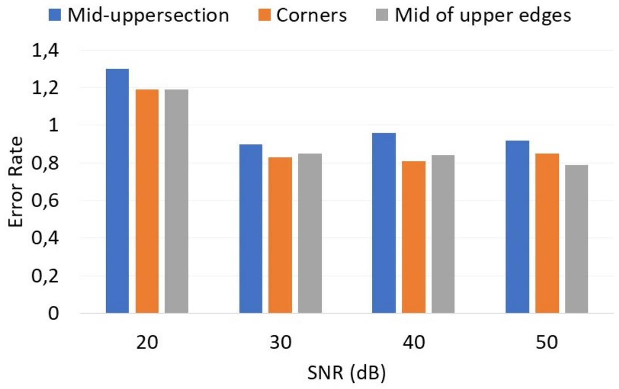


Fig. 6. Error rate for CR=0.3

In Figure 7, results per event for all configurations are shown. First, the results of event water tap is much the same for all configurations. Furthermore, for the other event with the lowest volume (microwave), the result obtained when using *mid-uppersection* configuration is by far the worst one. In general, the results of *mid-uppersection* configuration is acceptable only for those events located in mid section of the room (stirring cup) or for this events with high volume (hair dryer). It can be seen that for those events that are fixed near the wall, the four-microphone configurations obtains the best results with respect to *mid-uppersection* configuration. In addition, events placed in the midle near the wall (microwave) are better clasified by *mid of upper edges* configuration.

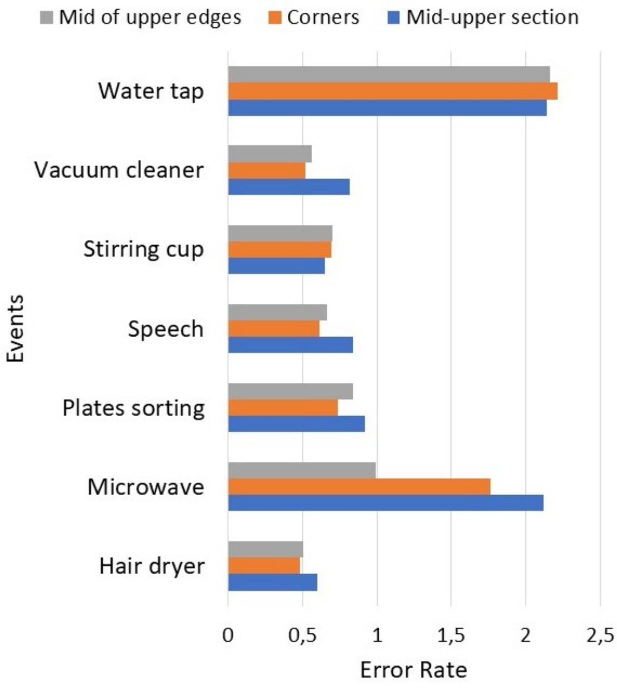


Fig. 7. Error Rate per event with SNR=20 dB and CR=0.6

4 CONCLUSIONS

Various conclusions are obtained from this results. This project was started believing that worst room condition were high reflexion coefficient. It has proven that results are improving by increasing CR's value. This is becouse power of reflected acoustic rays increase so microphones perceive more power of sounds in this cases that high value of CR is used. This situation does not affect noise. Furthermore, the impact of changing CR decrease when room conditions are idealized, that is, low noise.

Admittedly, it has been noted that SNR has large impact in the results. Let us remind that power of AWGN is defined with repect to average power of the whole isolated events, so every acoustic event is affected the same way. This produces that those events with high power becouse of their nature (vacuum cleaner or hair dryer) obtains better results than others (water tap or microwave). It is because SNR has higher values for this events with large volume. In any case, it has proven that large noise produces worst classification.

It has been shown that proximity and placement of the event respect to the microphone has impact in the results. It is becouse of microphone's directivity. Microphones are steered to the center of the room so those events wich position is near the wall obtains worst results. For example, the results obtained for stirring cup for configuration 1 are over-optimistic because of its position. The microphone position configuration 3, Figure 4 (b), is the best in this aspect since it covers big space of the room.

The goal of this project is to check if increasing the number of microphones will improve results. It has been shown that results obtained for those microphone position configurations with more than one microphone are better. First of all, applying fusion by majority voting techniques does the system more robust againts reflexion coefficient changes. As previously said, increasing reflexion coefficient improves results for noisy cases. However, this increase further enhance four-microphones configurations results. With respect to the position of the sources, four-microphones configurations covers big part of space including wall sides. Corner configuration covers most part of the room. At this point, it can be said that the best configuration to detect and classify the acoustic scene of the room that has been studied is four-microphones corners configuration.

5 ACKNOWLEDGMENT

6 REFERENCES

[1] W. Han, C.-F. Chan, C.-S. Choy, K.-P. Pun, "An efficient MFCC extraction method in speech recognition," presented at the *Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on*, pp. 4–pp (2006).

[2] H. L. Peiteado, *Detección e identificación de señales sonoras en entornos asistivos*, Ph.D. thesis, Universidad del País Vasco-Euskal Herriko Unibertsitatea (2016).

[3] S. McGovern, "A model for room acoustics, 2003," *URL: http://www. 2pi. us/rir. html* (2013).

[4] A. Mesaros, T. Heittola, *Detection and classification of acoustic scenes and events 2016, task 3* (2016 (accessed October 5, 2018)), URL `http://www.cs.tut.fi/sgn/arg/dcase2016/task-so`

[5] I. Mohino-Herranz, R. Gil-Pita, S. Alonso-Diaz, M. Rosa-Zurera, "Synthetical enlargement of mfcc based training sets for emotion recognition," *Int. J. Comput. Sci. Inf. Technol*, vol. 6, no. 1, pp. 249–259 (2014).

[6] F. Zheng, G. Zhang, Z. Song, "Comparison of different implementations of MFCC," *Journal of Computer science and Technology*, vol. 16, no. 6, pp. 582–589 (2001).

[7] G. J. Zapata-Zapata, J. D. Arias-Londoño, J. F. Vargas-Bonilla, J. R. Orozco-Arroyave, "On-line signature verification using Gaussian Mixture Models and small-sample learning strategies," *Revista Facultad de Ingeniería Universidad de Antioquia*, , no. 79, pp. 86–97 (2016).

[8] J. A. Bilmes, *et al.*, “A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models,” *International Computer Science Institute*, vol. 4, no. 510, p. 126 (1998).

[9] M. I. M. Herranz, *Emotion Analysis through Biological Signal Processing*, Ph.D. thesis, University of Alcalá (2017 4).

[10] D. Ruta, B. Gabrys, “Classifier selection for majority voting,” *Information fusion*, vol. 6, no. 1, pp. 63–81 (2005).

[11] A. Mesaros, T. Heittola, T. Virtanen, “TUT database for acoustic scene classification and sound event detection,” presented at the *Signal Processing Conference (EUSIPCO), 2016 24th European*, pp. 1128–1132 (2016).

[12] B. Efron, “Estimating the error rate of a prediction rule: improvement on cross-validation,” *Journal of the American statistical association*, vol. 78, no. 382, pp. 316–331 (1983).

[13] V. Zue, S. Seneff, J. Glass, “Speech database development at MIT: TIMIT and beyond,” *Speech communication*, vol. 9, no. 4, pp. 351–356 (1990).

[14] J. S. L. Stork, J. A.; Silva, K. O. Arras, “Audio Data Set for Human Activity Recognition,” (2010).

APPENDIX

NOMENCLATURE

MFCC = Mel Frequency Cepstral Coefficients
GMM = Gaussian Mixture Model
EM = Expectation-Maximization
CR = Reflexion Coefficient
SNR = Signal-to-noise Ratio
AWGN = Additive White Gaussian Noise

THE AUTHORS

A1firstname A1lastname

A2firstname A2lastname

A1firstname A1lastname is professor of audio signal processing at Helsinki University of Technology (TKK), Espoo, Finland. He received his Master of Science in Technology, Licentiate of Science in Technology, and Doctor of Science in Technology degrees in electrical engineering from TKK in 1992, 1994, and 1995, respectively. His doctoral dissertation dealt with fractional delay filters and physical modeling of musical wind instruments. Since 1990, he has worked mostly at TKK with the exception of a few periods. In 1996 he spent six months as a postdoctoral research fellow at the University of Westminster, London, UK. In 2001-2002 he was professor of signal processing at the Pori School of Technology and Economics, Tampere University of Technology, Pori, Finland. During the academic year 2008-2009 he has been on sabbatical and has spent several months as a visiting scholar at the Center for Computer Research in Music and Acoustics (CCRMA), Stanford University, Stanford, CA. His research interests include musical signal processing, digital filter design, and acoustics of musical instruments. Prof. Välimäki is a senior member of the IEEE Signal Processing Society and is a member of the AES, the Acoustical Society of Finland, and the Finnish Musicological Society. He was the chairman of the 11th International Conference on Digital Audio Effects (DAFx-08), which was held in Espoo, Finland, in 2008.

●

A2firstname A2lastname is a consulting professor at the Center for Computer Research in Music and Acoustics (CCRMA) in the Music Department at Stanford University where his research interests include audio and music applications of signal and array processing, parameter estimation, and acoustics. From 1999 to 2007, Abel was a co-founder and chief technology officer of the Grammy Award-winning Universal Audio, Inc. He was a researcher at NASA/Ames Research Center, exploring topics in room acoustics and spatial hearing on a grant through the San Jose State University Foundation. Abel was also chief scientist of Crystal River Engineering, Inc., where he developed their positional audio technology, and a lecturer in the Department of Electrical Engineering at Yale University. As an industry consultant, Abel has worked with Apple, FDNY, LSI Logic, NRL, SAIC and Sennheiser, on projects in professional audio, GPS, medical imaging, passive sonar and fire department resource allocation. He holds Ph.D.

and M.S. degrees from Stanford University, and an S.B. from MIT, all in electrical engineering. Abel is a Fellow of the Audio Engineering Society.