Estimation of Source Panning Parameters and Segmentation of Stereophonic Mixtures

Hjerrild, Jacob Møller; Christensen, Mads Græsbøll

Published in:
IEEE International Conference on Acoustics, Speech and Signal Processing

DOI (link to publication from Publisher):
10.1109/ICASSP.2018.8462522

Publication date:
2018

Document Version
Accepted author manuscript, peer reviewed version

ESTIMATION OF SOURCE PANNING PARAMETERS AND SEGMENTATION OF STEREOPHONIC MIXTURES.

Jacob Møller Hjerrild and Mads Græsbøll Christensen

Audio Analysis Lab, CREATE, Aalborg University, Denmark
{jmhh, mgc}@create.aau.dk

ABSTRACT

In this paper, we propose a method for finding the number of sources and their parameters from stereophonic mixtures. The method is based on clustering of narrowband interaural level and time differences for an unknown number of sources and uses an optimal segmentation on which the clustering is based. The parameter distribution, for both individual segments and across segments that comprise the entire signal, is modelled as a Gaussian mixture. For each segment parameters are estimated using a minimum description length algorithm for mixtures based on the expectation-maximization algorithm. The generalized variance and degree of membership of the Gaussian components across segments is used as a basis for the proposed selection of clusters amongst candidates. Simulations on synthetic and real audio shows promising results for source parameter estimation and number of sources estimated across segments. The optimal segmentation shows an improvement for parameter estimation success rate, compared to the uniform segmentation.

Index Terms— Source localisation, signal segmentation, multi-channel processing, audio clustering, audio analysis

1. INTRODUCTION

Broadcast audio streams, movie soundtracks, records for consumer playback and many handheld recorders are mainly available in stereo. Exploiting this allows for improved performance in tasks such as enhancement [1], pitch estimation [2] and source separation [3]. In [4] a stereophonic parametric multi-pitch estimator is proposed, based on a stereophonic signal model, assuming known panning parameters. By using knowledge about panning parameters, improvements have been shown for estimating multiple fundamental frequencies, as shown in [5] where the panning parameter estimation method requires a preceeding pitch estimate. The explicit estimation of panning parameters can be considered as a special case of the time direction of arrival (TDOA) estimation problem, which has applications such as blind source separation (BSS) and stereo to multichannel upmix, as proposed for stereo mixtures in [6] and [7]. In frequency domain BSS methods, it is often assumed that each frequency bin is dominated by one source only [8], meanwhile using uniform segments of data, with a short time Fourier transform (STFT) for binary mask estimation, often with a known number of sources [9, 10].

In this paper, we propose a blind stereophonic source panning estimation algorithm that determines the number of sources in the observed mixture, across segments, and segments the observed mixture in the process. The number of sources can be greater than the number of channels. The proposed method naturally engenders the W-disjointness of sources [8] by applying an adaptive frame size to improve local time segmentation, with a segmentation scheme that is optimal in some sense (e.g., 2-norm, posterior probability) [11]. We here propose to use the cost function of the mixture minimum description length (MMDL) method [12] for the segmentation, which is optimal in an approximate maximum a posteriori sense. The MMDL method is based on the Dirichlet prior in a Gaussian mixture model (GMM). This implicitly estimates the parameters and number of sources for each segment. Ideally, every mixture component is interpreted as a cluster, but each of the underlying components do not necessarily correspond to a source. This is also complicated because no unique definition of a "true cluster" necessarily exists, and the Gaussian assumption does not hold exactly, see e.g. [13]. We, therefore, based on the distribution of interaural time and level differences as derived in [8], propose to interpret diagonal Gaussian component candidates as true "clusters" if they have low generalized variance (GV) and are well separated from other candidates. This leads to an estimate of the number of unique sources and their panning parameters across all segments.

2. SIGNAL MODEL

In the following the signal model and assumptions are introduced. The mixture $y_m(n)$ in channel m at time instance n is modelled as a sum of K attenuated and delayed sources $s_k(n)$

$$y_m(n) = \sum_{k=1}^{K} \gamma_{m,k} s_k(n - \delta_{m,k}). \quad (1)$$

The kth source signal $s_k(n)$ mixed in channel m, is attenuated by gain coefficient $\gamma_{m,k}$ and delayed by $\delta_{m,k}$ samples. By expressing the kth source on the frequency grid ω as $S_{m,k}(\omega) = \sum_{n=0}^{N-1} s_k(n)\gamma_{m,k}e^{-j\omega(n+\delta_{m,k})}$, the assumption of W-disjoint orthogonality [14] can be expressed as

$$S_{1,k}(\omega)S_{1,i}(\omega) \approx 0 \quad \forall \omega, k \neq i. \tag{2}$$

Since this assumption is violated even for speech signals [8], we propose an optimal adaptive time segmentation to support this underlying assumption of frequency sparsity, as described in Section 3.1. In typical music productions, the gain coefficients are based on the tangent law, consequently inducing a constant perceived distance between listener and virtual source position. The gains $\gamma_{m,k}$ for channel m are given by [15]

$$\gamma_{m,k} = \begin{cases} \cos\Phi_k, & \text{for } m = 1 \\ \sin\Phi_k, & \text{for } m = 2 \end{cases} \tag{3}$$

where $\Phi_k = \phi_k + \phi_0$ is a sum of the perceived angle ϕ_k of the kth source and the speaker base angles $\pm\phi_0 = 45°$. The conditions $0° < \phi_0 < 90°$, $-\phi_0 \leq \phi \leq \phi_0$ and $\gamma_1, \gamma_2 \in [0, 1]$ are met [16]. Delays below 600 μs between stereo channels makes the virtual source position migrate toward the earlier speaker [17]. A two channel signal $y_m(n) \in \mathbb{R}^N$ is defined from N consecutive samples, defined for discrete time indeces going forward from n to $n+N-1$, where we propose that N is the adaptive segment size. We estimate a frequency domain measurement vector $\boldsymbol{x}$ for each segment, as a collection of panning parameters and relative channel delays

$$\boldsymbol{x} = \left[\hat{\gamma}(\omega), \hat{\tau}(\omega)\right]^T = \left[\arctan\left(\left|\frac{Y_1(\omega)}{Y_2(\omega)}\right|\right), \frac{1}{\omega}\angle\frac{Y_2(\omega)}{Y_1(\omega)}\right]^T, \tag{4}$$

where $\hat{\tau}(\omega) = \hat{\delta}_1(\omega) - \hat{\delta}_2(\omega)$, $Y_m(\omega)$ is the discrete Fourier transform of the mixture $y_m(n)$ in channel m and $\angle$ denotes phase. (4) is only true under the assumption of W-disjoint orthogonality and under the narrowband assumption where the maximum frequency ω_{max} and delay δ_{max} are restricted to $|\omega_{\text{max}}\delta_{\text{max}}| < \pi$. As modelled in [8], the marginal distributions are non-correlated and the non-W-disjoint error does not produce parameter dependency. Therefore, we model $\boldsymbol{x}$ as a K-component Gaussian mixture with diagonal covariance matrices expressed as

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \alpha_k \frac{\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{C}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\}}{\sqrt{(2\pi)^d \det(\boldsymbol{C}_k)}}, \tag{5}$$

where $\{\alpha_k, \boldsymbol{\mu}_k, \boldsymbol{C}_k\}$ is the mixing probability, mean and covariance of the kth Gaussian. Thus, $\boldsymbol{\mu}_k$ is the source parameter expressing the virtual positioning of the kth source and $\boldsymbol{\theta} \triangleq \{\alpha_1, \ldots, \alpha_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{C}_1, \ldots, \boldsymbol{C}_K\}$ specifies the full mixture as the complete set of parameters. Generally, $\alpha_k \geq 0, \quad \sum_{k=1}^{K} \alpha_k = 1, \text{ for } \quad k = 1, \ldots, K$.

3. PROPOSED METHOD

By observing a set of I i.i.d. samples $\mathcal{X} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(I)}\}$, where each element $\boldsymbol{x}^{(i)}$ is the measurement in (4) of the ith segment, the log-likelihood function corresponding to a K-source mixture is

$$\ln p(\mathcal{X}|\boldsymbol{\theta}) = \ln \prod_{i=1}^{I} p(\boldsymbol{x}^{(i)}|\boldsymbol{\theta}) = \sum_{i=1}^{I} \ln \sum_{k=1}^{K} \alpha_k p(\boldsymbol{x}^{(i)}|\boldsymbol{\theta}_k). \tag{6}$$

The maximization of the expression in (6) has no closed form solution and we find the maximum likelihood solution in an iterative manner, using the EM algorithm [18]. Since $\boldsymbol{\theta}$ has the number of sources K as dimensionality, the estimation of source parameters $\boldsymbol{\theta}$ is implicitly a joint estimation of the number of sources K and can be done by maximizing the MAP criterion

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} \{\ln p(\mathcal{X}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})\}, \tag{7}$$

where $\ln p(\boldsymbol{\theta})$ is the log of the prior on the parameters, which for the mixture model can be derived as [12]

$$\ln p(\boldsymbol{\theta}) = \frac{K(N_p+1)}{2} \ln I + \frac{N_p}{2} \sum_{k=1}^{K} \ln(I\alpha_k), \tag{8}$$

where N_p is the number of parameters specifying each component. In (8) one addend is dependent on K and one addend is dependent on the mixing probability α_k. When K is fixed, this prior has a simple Bayesian interpretation:

$$p(\{\alpha_1, \ldots, \alpha_K\}) \propto \prod_{k=1}^{K} \alpha_k^{-\frac{N_p}{2}}, \tag{9}$$

which is a Dirichlet-type prior that can be used for driving irrelevant components to extinction [12, 19, 20]. We use the EM-method of [12] with a Dirichlet prior in a modified M-step. The minimization criterion is based on an asymptotic assumption on the MAP cost function $\mathcal{J}(\boldsymbol{\theta}, \mathcal{X})$. The fully derived MMDL cost function is [12]

$$\mathcal{J}(\boldsymbol{\theta}, \mathcal{X}) = -\ln p(\mathcal{X}|\boldsymbol{\theta}) + \frac{N_p}{2} \sum_{k=1}^{K} \ln \frac{I\alpha_k}{12} + \frac{K}{2} + \ln\frac{I}{12} + \frac{K(N_p+1)}{2}, \tag{10}$$

where K is the number of components with non-zero weight in the mixture ($\alpha_k > 0$). The Gaussian source parameters and number of sources can be jointly estimated by minimizing the cost function

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathcal{J}(\boldsymbol{\theta}, \mathcal{X}). \tag{11}$$

The EM algorithm is initialized with a much higher K than expected, and the M-step applies Dirichlet prior in (9) to annihilate irrelevant components. When initial K is high, it is ensured that the true clusters are among the estimated candidates, with the probability shown in Fig. 1.

Algorithm 1 Optimal segmentation

while $m \times N_{\text{MIN}} \leq$**length(signal) do**
  Initialize $B = \min([m, B_{\text{max}}])$.
  for $b = 1$ **to** B **do**
    block of signal to use is $m - b + 1, \ldots, m$
    estimate $(\hat{\gamma}(\omega), \hat{\delta}(\omega))$ from (4)
    compute $\mathcal{J}_{(m-b+1)m}$ from (10)
    if $m > b$ **then**
      $\mathcal{J}(b) = \mathcal{J}_{(m-b+1)m} + \mathcal{J}_{1(m-b)}$
    else
      $\mathcal{J}(b) = \mathcal{J}_{(m-b+1)m}$
    end if
  end for
  $b_{\text{opt}} = \arg\min \mathcal{J}(b)$
  $m = m + 1$
end while
$m = M$
while $m > 0$ **do**
  number of blocks in segment is $b_{\text{opt}}(m)$
  $m = m - b_{\text{opt}}(m)$
end while

3.1. Signal segmentation

The characteristics of each dominating source in the mixture is varying over time, meaning that a uniform segment length N is not optimal. The optimal segmentation of $y(n)$ requires that the cost is additive over distinctive segments, which is true for the MMDL criterion of (11). The cost associated with the different outcomes from the set of segment lengths can be compared and the optimal can be chosen as the one that minimizes (11). The segmentation is based on the principle in [11] which has also been applied in [21, 22] and is outlined in Algorithm 1. A minimal segment sample size, N_{min} is defined, generating a block of N_{min} samples and dividing the signal into M blocks. This gives 2^{M-1} ways of segmenting the signal into M blocks. A maximum number of blocks B_{max} is defined to ease on computational complexity. A dynamic programming algorithm computes the optimal segment length b_{opt} for all blocks, $m = 1, \ldots, M$, starting at $m = 1$ moving continously to $m = M$. For every block, the cost of all new block combinations are reused from earlier blocks. When the end of the signal is reached, the optimal segmentation of the signal is found, starting with backtracking the last block and continuing through the signal to the beginning. Starting at $m = M$, setting the number of blocks in the last segment to $b_{\text{opt}}(M)$. The next segment ends at block $m = M - b_{\text{opt}}(M)$ and includes $b_{\text{opt}}(M - b_{\text{opt}}(M))$ blocks. This is continued until $m = 0$.

General test setup	
Sampling rate	44.1 kHz
Mixture duration	SQAM: 15 sec. IOWA: 60 sec.
Uniform seg.	600 ms
N_{min}	150 ms
B_{max}	20
Synthetic signals	
Non-unit amplitudes	20 complex and inharmonic
Note durations	$[300, 600, \cdots, 3000]$ ms
$f_0 \in [80, 1700]$ Hz	In semitone steps
Cluster is correct if:	$\lvert\phi_k - \hat{\phi}_k\rvert < .5^\circ \wedge \lvert\delta_k - \hat{\delta}_k\rvert < .5$

Table 1. Test and signal specifications.

3.2. Selecting clusters amongst candidates

We apply a ranking of the candidates components, such that the best candidate has the lowest GV, the second best candidate has the second lowest GV and so forth. We define GV as β and for each candidate we compute $\beta = \det(\hat{\boldsymbol{C}})$ and define s as the ordered index, such that it reflects that $\text{GV}_1 < \text{GV}_2 < \cdots < \text{GV}_S$ and rank all estimated parameters in $\hat{\boldsymbol{\theta}}$ by $s = \{1, 2, \cdots, S\}$.

Next, we select the set of β-ranked components that is well separated, meaning that their measurements are only assigned to one component. The degree of membership ζ_{is}, which is the a posteriori probability that $\boldsymbol{x}_i$ was generated by mixture component s is

$$\zeta_{is} = \frac{\hat{\alpha}_s \mathcal{N}(\boldsymbol{x}_i | \hat{\boldsymbol{\mu}}_s, \hat{\boldsymbol{C}}_s)}{\sum_{j=1}^{S} \hat{\alpha}_j \mathcal{N}(\boldsymbol{x}_i | \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{C}}_j)}. \tag{12}$$

Going forward in s from 2 to S, we check for shared measurement between the first s columns of ζ_{is}. A shared measurement is found if $0 < \zeta_{i\sigma} < 1 \wedge 0 < \zeta_{is} < 1$, where $\sigma = \{1, 2, \cdots, s-1\} \forall\, \boldsymbol{x}_i$.

4. EXPERIMENTS

4.1. Signal segmentation

The segmentation is tested on a synthetic mixture of two sources with a ground truth to when each source is active. The signal is segmented according to the MMDL criterion of (11). A representative example of the chosen segment length as a function of time is shown in Fig. 2 with white vertical lines. In the top the two channel signal is shown in time domain, with a black horizontal line that indicates which of the two sources is active. A spectrogram of the right channel is shown to detail the view. The chosen segments are long if the content is not changing. The two sources played from 0 to 4 sec, will consistenty not overlap, and we would expect the segments to be long but random. When the silence begins a shorter segment length is chosen in all three

silent periods, at [3.6, 5.4, 8] sec. The overlapping notes at 5 sec. are chosen as expected, and the next three notes have an overlap that is segmented in to two parts, where the first part has only one source active. The following notes after the silence at 8 sec. are chosen precisely in 300 ms segments. Lastly, the long note from 12-15 sec. is chosen in segments of 600 ms, and the two notes in the end are also chosen, even at an overlap with the other source. This indicates that the MMDL-criterion decribes the dominating sources well, based on the underlying clustering.

4.2. Selection of clusters amongst candidates

The ranking and selection of component candidates is tested using the University of Iowa musical instruments database [23] in 2500 iterations for uniform segmentation only. The procedure is to fit a mixture of 35 Gaussians components to the parameter distribution, β-rank the candidates and then select the well separated clusters. We measure the precision and recall as

$$\text{precision} = \frac{\sum_{i=1}^{I} \text{TP}_i}{\sum_{i=1}^{I} \text{TP}_i + \text{FP}_i}, \tag{13}$$

$$\text{recall} = \frac{\sum_{i=1}^{I} \text{TP}_i}{\sum_{i=1}^{I} \text{TP}_i + \text{FN}_i}, \tag{14}$$

where TP_i, FP_i and FN_i denote the number of true positive, false positive and false negative cluster estimates. For 2-5 randomly picked sources we obtain a recall rate of 96.9% with a precision of 95.6%. Fig. 1 compares β-ranking to ranking by the component mixing probability α. It is clear that by β-ranking we can recall the true K clusters with 97% certainty within the first K candidates, and the first $K-1$ candidates is true with more than 99% certainty for true $K = 4$, as long as the number of component candidates is high.

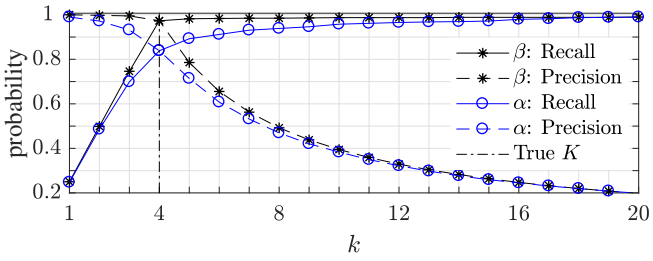

Fig. 1. Precision and recall comparison between ranking by α or β, in 1000 iterations with true $K = 4$, on the IOWA database.

4.3. Source parameter estimation

For further evaluation of the segmentation and panning parameter estimation performance, we compare the adaptive to the uniform segmentation in 100 iterations. For each iteration,

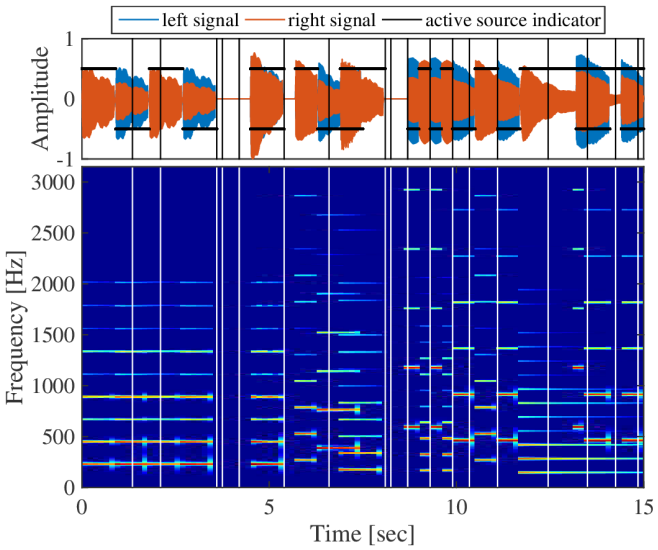

Fig. 2. Optimal segmentation on two synthetic sources.

Estimation methods	MMDL seg.	uniform seg.
Precision rate	94.6%	93.5%
Recall rate	89.1%	88.6%

Table 2. Estimation performance across segments on audio from the SQAM database.

a stereophonic mixture consists of minimum 2 and maximum 5 randomly picked sources from the sound quality assessment material recordings (SQAM) [24]. The files containing pink noise has been removed from the test set. The results in Table 2 shows a precision of 93.5% and a recall rate of 88.6% for the uniform segmentation with a small improvement with adaptive segmentation.

5. CONCLUSION

In this paper, a blind source panning estimation method has been proposed that determines the number of sources and their parameters in the observed stereophonic mixture. The method is based on clustering of narrowband interaural level and time differences and uses an optimal segmentation. The parameter distribution, for both individual segments and across segments is modelled as a Gaussian mixture. The generalized variance and degree of membership of the Gaussian components are used for the proposed selection of clusters across segments. Simulations on synthetic and real audio show promising results for the source parameter estimates, for both the uniform and the optimal segmentation. Experiments on both the IOWA and SQAM data show robust precision and recall rate for anechoic and various instrument and music ensemble samples.

6. REFERENCES

[1] J. R. Jensen, J. Benesty, and M. G. Christensen, "Joint filtering scheme for nonstationary noise reduction," *Proc. European Signal Processing Conf.*, pp. 2323–2327, 2012.

[2] M. W. Hansen, J. R. Jensen, and M. G. Christensen, "Estimation of multiple pitches in stereophonic mixtures using a codebook-based approach," *IEEE ICASSP. Proc.*, 3 2017.

[3] M. I. Mandell, R. J. Weiss, and D. P. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 18, no. 2, pp. 384–394, November 2010.

[4] M. W. Hansen, J. R. Jensen, and M. G. Christensen, "Pitch estimation of stereophonic mixtures of delay and amplitude panned signals," in *European Signal Processing Conference*, 2015, pp. 36–40.

[5] T. Kronvall, A. Jakobsson, M. Hansen, J. Jensen, and M. Christensen, "Sparse multi-pitch and panning estimation of stereophonic signals," in *11th IMA International Conference on Mathematics in Signal Processing*, 2016.

[6] C. Avendano and J.-M. Jot, "Frequency domain techniques for stereo to multichannel upmix," *AES 22nd international Conference on Virtual, Synthetic and Entertainment Audio*, June 2002.

[7] S. Kraft and U. Zolzer, "Stereo signal separation and upmixing by mid-side decomposition in the frequency-domain," *Int. Conference on Digital Audio Effects (DAFx-15)*, 2015.

[8] S. Rickard and O. Yilmaz, "On the appriximate w-disjoint orthogonality of speech," *IEEE Acoustics, Speech, and Signal Processing*, 2002.

[9] S. Rickard and O. Yilmaz, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[10] N. Ito, S. Araki, and T. Nakatani, "Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors," *IEEE ICASSP.*, pp. 3238–3242, 2013.

[11] P. Prandoni and M. Vetterli, "R/d optimal linear prediction," *IEEE Trans. Speech and Audio Proc*, vol. 8, pp. 646–655, 2000.

[12] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 24, pp. 381–396, 2002.

[13] C. Hennig, M. Meila, F. Murtagh, and R. Rocci, *Handbook of Cluster Analysis*, Chapman and Hall/CRC, 2015.

[14] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures," *IEEE International Conference on Acoustics, Speech, and Signal Processing.*, vol. 5, pp. 2985–2988, 2000.

[15] B. Bernfeld, "Attempts for better understanding of the directional stereophonic listening mechanism," in *Audio Engineering Society Convention 44*, March 1973.

[16] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc*, vol. 45, no. 6, pp. 456–466, June 1997.

[17] J. Blauert, *The Psycophysics of Human Sound Localization*, MIT Press, 2009.

[18] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[19] D. Ormoneit and V. Tresp, "Averaging, maximum penalized likelihood and bayesian estimation for improving gaussian mixture probability density estimates," *IEEE Trans. Neural Networks*, vol. 9, no. 4, pp. 639–650, 1998.

[20] Z. Zivkovic and F. van der Heijden, "Recursive unsupervised learning of finite mixture models.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 5, pp. 651–656, 2004.

[21] P. Prandoni, M. M. Goodwin, and M. Vetterli, "Optimal time segmentation for signal modeling and compression," *IEEE Acoustics, Speech, and Signal Processing*, pp. 2029–2032, 1997.

[22] S. M. Nørholm, J. R. Jensen, and M. G. Christensen, "Instantaneous fundamental frequency estimation with optimal segmentation for nonstationary voiced speech," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 24, no. 12, pp. 2354–2367, 2016.

[23] L. Fritts, "University of iowa musical instruments database," `http://theremin.music.uiowa.edu/MIS.html`, pre and post 2012.

[24] E. B. Union, *Sound quality assessment material recordings for subjective tests: Users handbook for the EBU SQAM CD*, 2008, Tech. Rep. EBU - TECH 3253.