Interspeech 2018
2-6 September 2018, Hyderabad

Subband Weighting for Binaural Speech Source Localization

Girija Ramesan Karthik[1], *Parth Suresh*[2], *Prasanta Kumar Ghosh*[1]

[1]Electrical Engineering, Indian Institute of Science (IISc), Bengaluru-560012, India
[2]Computer Science and Engineering, TKM College of Engineering, Kollam-691005, India

karthikgr@iisc.ac.in, parthsuresh96@gmail.com, prasantg@iisc.ac.in

Abstract

We consider the task of speech source localization from a binaural recording using interaural time difference (ITD). A typical approach is to process binaural speech using gammatone filters and calculate frame-level ITD in each subband. The ITDs in each gammatone subband are statistically modelled using Gaussian mixture models (GMMs) for every direction during training. Given a binaural test-speech, the source is localized using maximum likelihood (ML) criterion. In this work, we propose a subband weighting scheme where subband likelihoods are weighted based on their reliability. We measure the reliability of a subband using the average frame level localization error obtained for the respective subbands. These reliability values are used as the weights for each subband likelihood prior to combining the likelihoods for ML estimation. We also introduce non-linear warping of these weights to accommodate and analyse a larger space of possible subband weights. Experiments on Subject_003 from the CIPIC database reveal that weighting the subbands is better than the unweighted scheme of combining likelihoods.

Index Terms: gammatone filters, interaural time difference, warping function

1. Introduction

Machine localization of speech sources is essential for a wide range of applications, including human-robot interaction, surveillance and hearing aids. Robot sound localization algorithms have been proposed using microphone arrays with varied number of microphones [1–6]. However, humans have an incredible ability to localize sounds with just two ears. The major cues that help in localization are interaural time difference (ITD) and interaural level difference (ILD). These cues can be captured by the head-related transfer function (HRTF) [7]. Its time domain equivalent is the head-related impulse response (HRIR). Algorithms inspired by binaural localization of humans would extract these features from the input signals [8–22]. To account for the time and frequency variability of these cues, time-frequency (TF) representations of the binaural signals are used. One of the most common time-frequency representations is the Short-Time Fourier Transform (STFT) [10, 12, 15, 17, 19]. Another approach is to use gammatone filters [23] where, unlike STFT, the subband width and spacing are not uniform [8, 11, 13, 14, 16, 22, 24]. The use of gammatone filters is inspired by the filter structure of the cochlea in human ears. In this work, we use gammatone filters to preprocess the binaural signals.

May et al. [13, 14] and Woodruff et al. [16] use Gaussian mixture models (GMMs) to model ITD, ILD and their joint distribution (ITLD) for each gammatone subband in each direction. Then, for a test-speech, log-likelihood integration is performed across the TF plane and the direction with the maximum likelihood is picked as the direction of arrival (DoA). Ma et al. [22] use DNNs to map the cross-correlation function, instead of ITDs, of each TF bin to the posterior distribution (PD) of DoAs. PDs across the TF plane are averaged similar to the likelihood integration in GMM based methods and the direction with the maximum posterior probability is chosen as the DoA. In this work, we consider only ITD and extend the GMM based method of May et al. [13] by investigating the localization accuracy of each subband. DoA estimation can be treated as a classification problem where, given a feature (ITD), the latent class (DoA) needs to be inferred. For a subband, each direction will have its own distribution of ITD. Higher discrimination among distributions of different directions results in a better classification accuracy. Hence, subbands with a high level of discrimination are more reliable than the ones with a low level of discrimination. Addition of noise could decrease this discrimination and lead to a decrease in localization accuracy. We hypothesize that weighting the subband likelihoods based on their reliabilty can improve localization accuracy. Using localization error as a measure of discrimination, we find a measure of reliability for each of these subbands. These reliability values are then used as the subband weights. To account for a larger space of weights and possible changes in the reliability of subbands with SNR, we propose to warp the weights using a non-linear warping function. From the localization errors obtained for all possible weights at each SNR, we select one set of weights that performs best on all SNRs combined. Experiments with Subject_003 from the CIPIC database [25] reveal that the weighted subband localization scheme works better than the unweighted scheme. It also turns out that the best weights obtained for each SNR performs marginally better than the SNR independent weights under the simulated additive white gaussian noise (AWGN) conditions.

2. Binaural Cue Extraction and Localization

DoA estimation consists of the steps shown in figure 1. We provide the details of these steps in the following subsections.

2.1. Gammatone Filters

The binaural signals are processed through N=32 fourth order gammatone filters. Their center frequencies are equally distributed with respect to the equivalent rectangular bandwidth (ERB) scale between 80Hz and 5kHz, starting with 80Hz and ending with 4.6kHz. This range primarily covers the entire speech spectrum. To approximate the neural transduction process of the inner hair cells, the outputs of the gammatone filters are halfwave rectified and square-root compressed [13]. The resulting outputs of the left and right channels of the i^{th} subband are denoted by l_i and r_i.

Authors thank Pratiksha Trust for their support.

10.21437/Interspeech.2018-2173

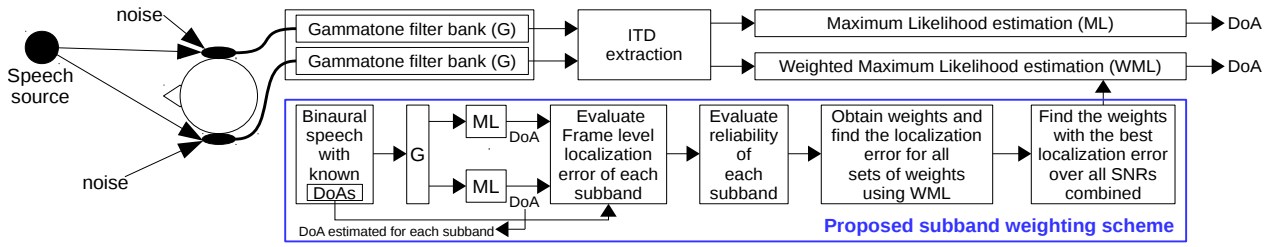

Figure 1: *Proposed subband weighting scheme to obtain the best weights for weighted Maximum Likelihood localization*

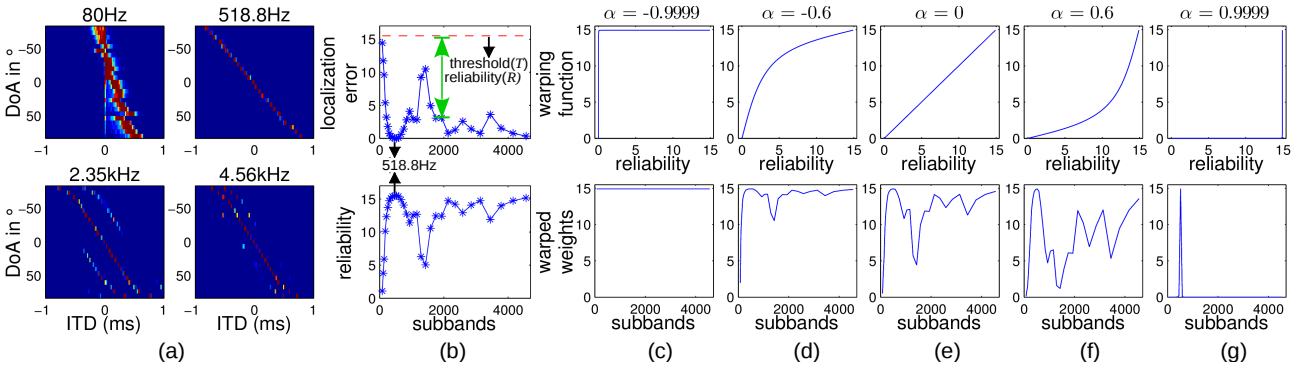

Figure 2: *(a) ITD distributions of all directions in gammatone subbands with center frequencies 80 Hz, 518.8Hz, 2.35kHz & 4.56 kHz. (b) Average frame-level localization errors of subbands (top) and their respective reliabilities (bottom) (c-g) Warping functions (top) and their respective warped subband weights (bottom) for α = -0.9999, -0.6, 0, 0.6, 0.9999 and T=14.92. Choosing α = -0.9999 is equivalent to weighing all subbands equally whereas α = 0.9999 is almost equivalent to choosing just one subband.*

2.2. ITD Extraction

Frame-level ITD in each subband is then calculated using normalized cross correlation (NCC) [8, 13] between l_i and r_i with a rectangular window of length W and shift of length W_s. $\tau_{i,j}$ is the ITD of i^{th} subband in the j^{th} frame and is given by

$$\tau_{i,j} = \underset{\tau}{\operatorname{argmax}}\, C_{i,j}(\tau), \tag{1}$$

where $C_{i,j}$ is the NCC function. In addition to this, exponential interpolation is used to obtain fractional delays [13].

2.3. GMM Parameter Estimation

GMMs are trained on clean speech ITDs of each subband in each direction. As shown in [13], ITD distributions of high-frequency subbands have multiple modes . Hence, the need for GMMs. Information theoretic criteria such as Akaike Information Criterion (AIC) [26] and Bayesian Information Criterion (BIC) [27] are used to evaluate the optimum number of components in each GMM. The optimal number of components is obtained using AIC as well as BIC. The lower number between the two is chosen as the optimal number of components.

2.4. Likelihood and Localization

Suppose we consider D directions. Then $\lambda_{i,d}$ is the set of GMM parameters for the i^{th} subband in the d^{th} direction where d ranges from 1 to D. $\tau_{i,j}$ denotes the ITD of the i^{th} subband in the j^{th} frame. Then, $p(\tau_{i,j}|\lambda_{i,d})$ is calculated $\forall i, d$. Given a test binaural speech, May et al. [13] combined the likelihoods of all the subbands, for each d to obtain a single likelihood for each direction. And then, the direction with the maximum likelihood (ML) is chosen as the DoA estimate in the j^{th} frame.

$$DoA_j = \underset{d\in\{1,..,D\}}{\operatorname{argmax}} \sum_{i=1}^{N} \log p(\tau_{i,j}|\lambda_{i,d}) \tag{2}$$

The DoAs from multiple number of frames (nf) are pooled to obtain the DoA with the maximum frequency of occurrence.

3. Proposed Subband Weighting

We propose a weighted maximum likelihood (WML) method in which the DoA estimate in the j^{th} frame is given by

$$DoA_j = \underset{d\in\{1,..,D\}}{\operatorname{argmax}} \sum_{i=1}^{N} w_i \log p(\tau_{i,j}|\lambda_{i,d}), \tag{3}$$

where w_i is the weight corresponding to the i^{th} subband. With the motivation to weight subband likelihoods based on their reliabilities, the method to obtain these weights is described in the following subsections.

3.1. Subband Reliability

Reliability of a subband is measured as a quantity that is inversely proportional to the average localization error of that subband. To calculate the localization error of each subband we need the frame level direction estimates of each subband.

$$DoA_{i,j} = \underset{d\in\{1,..,D\}}{\operatorname{argmax}} \log p(\tau_{i,j}|\lambda_{i,d}), \tag{4}$$

where $DoA_{i,j}$ is the DoA estimate of the i^{th} subband in the j^{th} frame. The frame level estimates are then pooled over nf frames to obtain the final DoA. Figure 2(a) shows the ITD distributions of 25 azimuthal directions in 4 different subbands. It can be seen that subband at 80 Hz has a lot of overlap among the distributions unlike 518.8 Hz whose distributions are well separated. This means that subband at 518.8 Hz is more reliable than 80 Hz. In a previous work [24], we evaluated the subband localization errors for clean speech using speech segments of duration nf = 100. However, here we use nf = 1. This is equivalent to sampling ITDs from each of the trained GMMs and finding the localization error due to the samples that are incorrectly localized. Hence, choosing nf = 1 is a better way of measuring discrimination amongst GMMs in a subband. The average localization errors of subbands obtained using nf = 1 is shown in figure 2(b) (top row). As hypothesized, subband at 518.8 Hz has a localization error lower than that of 80 Hz.

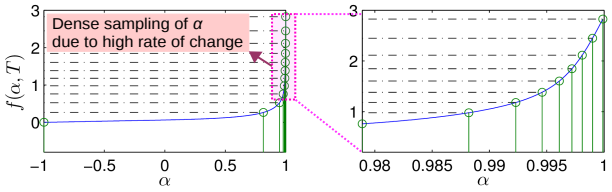

Figure 3: *An illustration of the α values obtained by sampling $f(\alpha, T)$ for 12 samples of α using T = 14.42.*

Now, we propose a reliability measure which is obtained using the localization error of each subband. As seen in the top row of figure 2(b) each subband localization error (S_i) is subtracted from a chosen threshold (T) to obtain the reliability (R_i) of the i^{th} subband shown in the bottom row. Now, these reliabilities can be used as the subband weights.

3.2. Non-Linear Warping

In the previous section, reliabilities are obtained using the subband localization errors of clean speech. The relative reliabilities of different subbands might vary in the presence of noise and it can also be SNR dependent. Also, $R_i = T - S_i$ is one way, amongst many other ways, of obtaining reliability. So, to account for these possibilities, we introduce non-linear warping of the obtained reliabilities so as to span over a wide range of possible reliabilities which are used to compute the subband weights. The weight computation and warping function used are defined as follows.

$$w_{\alpha,T}(i) = \frac{2T}{\pi} \arctan\left(\frac{(1-\alpha)}{(1+\alpha)} \tan\left(\frac{R_i \pi}{2T}\right)\right), \quad (5)$$

where α is the warping parameter and $w_{\alpha,T}(i)$ is the weight of the i^{th} subband. Top row of figure 2(c-g) shows the warping functions for different values of α. The corresponding warped weights are shown respectively in the bottom row. It should be noted that choosing α = -0.9999 is equivalent to weighting all the subbands equally. However, α = 0.9999 almost corresponds to choosing just one subband corresponding to the lowest localization error, i.e., the highest reliability. So, given a pair of parameters (α, T) and the subband errors (S_i), the corresponding weights are calculated as per the following steps.

Step1 : $S_i \leftarrow S_i - min(S_i)$,
Step2 : $R_i \leftarrow T - S_i$,
Step3 : $w_{\alpha,T}(i) \leftarrow$ Substitute R_i, α and T in eqn. (5)

Step1 is essential to ensure that the subband with the minimum error (maximum reliability) is always given the maximum weight equal to T.

3.3. α Sampling

α has a range of $(-1, 1)$. However, uniformly choosing alphas in this range is not reasonable as the rate of change of weights is not linear with α. We define the rate of change of weights as

$$v(\alpha, T) = \lim_{\delta \to 0} \frac{1}{2\delta} \sum_{i=1}^{N} (w_{\alpha+\delta,T}(i) - w_{\alpha-\delta,T}(i))^2, \quad (6)$$

where $v(\alpha, T)$ is the rate of change of weights at (α, T). We then integrate this function with respect to α to get the cummulative change, $f(\alpha, T)$. Figure 3 shows $f(\alpha, T)$ for T = 14.42. $f(\alpha, T)$ is then used to select different values of α. $f(1, T)$ is equal to the total cummulative change. To obtain n samples of α we find the α such that $f(\alpha, T) = \frac{f(1,T)}{n}, \frac{2f(1,T)}{n}, ..., f(1, T)$. The motivation is to sample more α values in the regions with higher rate of change of weights. In figure 3 it can be seen that upto $\alpha \sim 0.8$ there is hardly any increase in $f(\alpha, T)$. However, it then increases steeply. As it can be seen in figure 3, there is a higher density of samples close to α=1.

3.4. Best SNR independent weights

For each SNR, we calculate the average localization error of WML using all possible (α, T) pairs. Let $E_{\alpha,T}(k)$ be the average localization error of WML at the k^{th} SNR obtained using (α, T). Then, the minimum localization error for the k^{th} SNR is given by $E_{min}(k) = \min_{\alpha,T} E_{\alpha,T}(k)$. Now, to obtain the best SNR independent weights (BSIW), we need to obtain the (α, T) pair which performs best over all SNRs combined. The best (α, T) pair is given by

$$(\alpha_{best}, T_{best}) \leftarrow \underset{\alpha,T}{\operatorname{argmin}} \sum_{k=1}^{nSNR} (E_{\alpha,T}(k) - E_{min}(k))^2, \quad (7)$$

where $nSNR$ is the total number of SNRs considered.

4. Experiments and Results

4.1. Database

Speech from TIMIT database [28] is used for all experiments. Binaural speech is simulated using HRIRs of Subject_003 from the CIPIC database [25].

4.2. Experimental Setup

***Binaural speech data preparation*:** Localization experiments are performed only in the frontal horizontal plane. The CIPIC database consists of HRTFs of 25 directions in the frontal horizontal plane. Speech from the TIMIT database has a sampling frequency of 16kHz, whereas CIPIC HRIRs are sampled at 44.1kHz. Therefore, speech is upsampled to 44.1kHz and then filtered through the HRIRs to obtain binaural speech corresponding to each direction.

***ITD extraction* & *GMM parameter estimation*:** ITDs are calculated using eqn. (1) with a frame duration of 20msec (W = 882) and a frame shift of 10msec (W_s = 441). GMMs are trained using frame-level ITDs that are computed from a training binaural speech of duration 10sec. This provides 1000 frames to train each of the 800 (25 directions $\times$ 32 subbands) GMMs. EM algorithm [29] with random initialization is used for parameter estimation. As described in Section 2.3, AIC and BIC are used to compute the optimal number of Gaussian components. However, the maximum number of components is restricted to 20.

***Localization error*:** Let ϕ be the actual azimuthal angle and $\hat{\phi}$ be the estimated angle. Then the localization error is $e = |\phi - \hat{\phi}|$. Average localization error is obtained by taking the mean of the localization errors in ns different binaural speech segments. We use ns = 1000 for calculating subband reliability. In all other experiments we consider ns = 180. The localization errors are calculated in degrees.

***Subband reliabilities*:** Subband reliabilities are calculated using average localization error of each subband as per the procedure outlined in section 3.1. Bottom row of Figure 2(b) shows the obtained reliabilities. Subband with center frequency 518.8 Hz has the highest reliabilty as it has the least average localization error equal to 0.028°.

***α Sampling* & T:** Thresholds are considered from 14.42 to 40 in steps of 0.5. It has been observed that the sampled α values do not change considerably with change in T. Hence, a set of 100 α values common to all T has been sampled using T = 14.42. A dense set of α starting from -0.9999 to 0.9999 with a step of 0.0001 have been considered to obtain $f(\alpha, T)$. δ = 0.00005 has been used to obtain the rate of change function $v(\alpha, T)$. 1/2δ factor in eqn. (6) has been omitted while calculating the rate of change as it is a common factor and does not affect the sampling process. The 100 sampled α values are shown in figure 4(a). They start with -0.9998 and end with 0.9999.

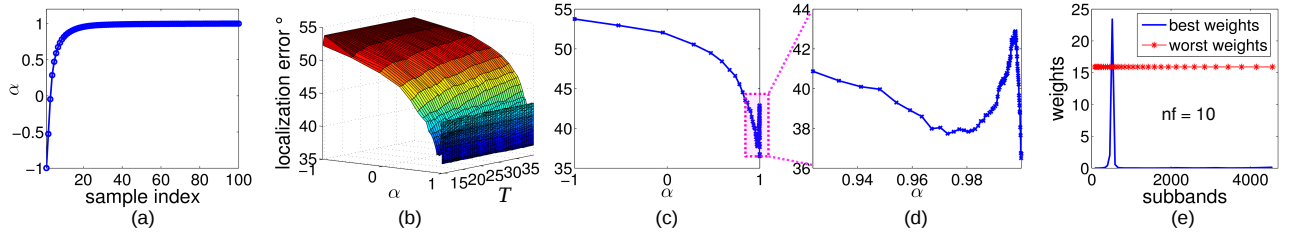

Figure 4: *(a) 100 sampled α values using T=14.42. (b) Surface plot of the localization error (°) vs (α, T) for nf = 10 at SNR = -12dB. (c) Cross-section of the surface plot with fixed T = 39.92. (d) A zoom of the region in (c) with the highest density of α samples. (e) Best (blue-) and the worst (red-$*$) SNR independent weights corresponding to nf = 10.*

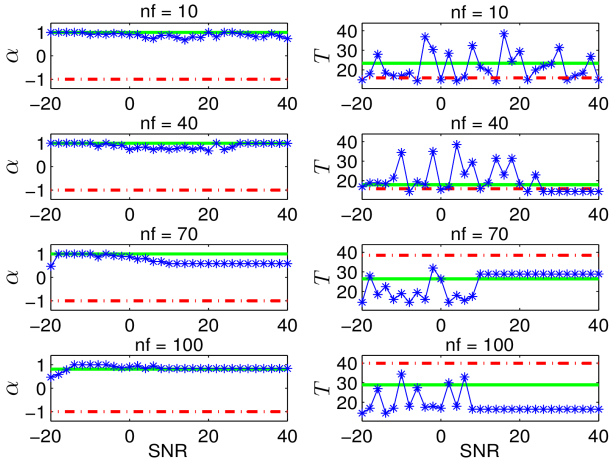

Figure 5: *The best SNR specific (blue-$*$) and SNR independent weights (green-). The worst SNR independent weights (red-·).*

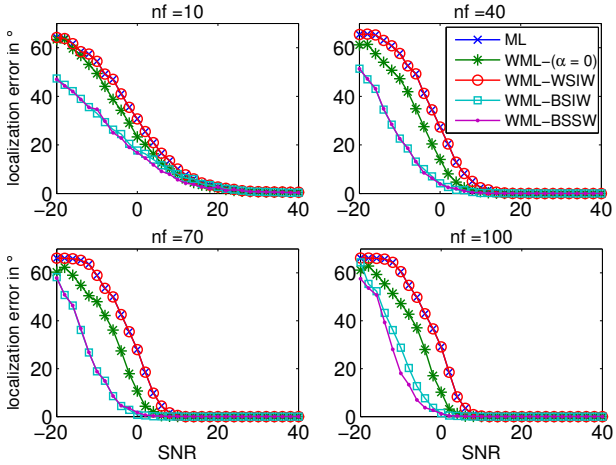

Figure 6: *Localization error vs SNR for different durations (nf) of test speech.*

4.3. Results and Discussion

***Best weights*:** We first evaluate the localization accuracy of WML for all pairs of (α, T) at different SNRs. We consider SNRs varying from -20dB to 40dB in steps of 2 by adding AWGN. The SNR is relative to each channel. We also consider different durations of test speech i.e. nf = 10,40,70 & 100. Localization errors at every pair of (α, T) for nf = 10 at SNR = -12 dB is shown in the 3D surface plot in figure 4(b). Figure 4(c) shows the surface profile for a fixed T = 39.92. Figure 4(d) zooms into the region with maximum density of α samples. It can be seen that this region has the lowest localization errors.

The best (α, T) pairs for each SNR for each nf are then obtained as shown in figure 5. Now we obtain the best SNR independent (α, T) pair using the procedure outlined in Section 3.4. Similarly, we also find the worst SNR independent (α, T) pair i.e., the (α, T) pairs that maximize the quantity on the right hand side of eqn. (7). The best and the worst SNR independent (α, T) pairs obtained for each nf are shown in figure 5 and in the table below.

	α_{best}	T_{best}	α_{worst}	T_{worst}
nf = 10	0.9996	23.42	-0.9998	15.92
nf = 40	0.9997	17.92	-0.9998	15.92
nf = 70	0.9998	26.42	-0.9998	38.42
nf = 100	0.8152	28.92	-0.9998	38.92

The weights corresponding to $(\alpha_{best}, T_{best})$ and $(\alpha_{worst}, T_{worst})$ of nf=10 are shown in figure 4(e). As seen in the figure, using the best weights is almost equivalent to selecting just the most reliable subband i.e. subband with center frequency 518.8 Hz. On the other hand, using the worst set of weights uniformly weighs all the subbands. This is equivalent to ML.

***Performance Evaluation*:**The following schemes are evaluated: WML with reliabilities in Section 3.1 as weights (WML(α=0)), WML with the best SNR independent weights (WML-BSIW), WML with the worst SNR independent weights (WML-WSIW) and WML with the best SNR specific weights (WML-BSSW). As seen in figure 6, WML-BSIW performs much better than ML especially at very low SNRs with an average localization accuracy improvement of upto 30° to 40°. The performance of WML-WSIW is close to ML. α_{worst} for all values of nf is equal to -0.9998. Choosing α = -0.9998 is almost equivalent to uniformly weighing all subbands which is the same as ML. This shows that uniformly weighing all subbands has the lowest performance among all the weights considered. Another important observation is that WML-BSIW performs as good as WML-BSSW at all SNRs. This suggests that in the presence of AWGN, choosing a fixed set of weights over all SNRs yields a performance comparable to using SNR specific weights.

5. Conclusions

We propose a weighted Maximum Likelihood (WML) method for binaural speech source localization. We present a measure of reliability of each subband obtained by using frame level localization error which measures the discrimination of the trained GMMs in each subband. This reliability is, in turn, used as the weights with the inclusion of non-linear warping to account for changes in reliabilities with SNR and also to span a larger space of possible weights. Experimental results with the best set of weights show that WML performs better than ML. It has also been observed that WML-BSIW performs as good as WML-BSSW suggesting that in the presence of AWGN, selection of SNR specific weights is not necessary. It would be interesting to extend this weighting scheme to other kinds of noisy conditions including diffuse noise and reverberation.

6. References

[1] S. Argentieri, P. Danes, and P. Souères, "A survey on sound source localization in robotics: From binaural to array processing methods," *Computer Speech & Language*, vol. 34, no. 1, pp. 87–112, 2015.

[2] J.-M. Valin, F. Michaud, J. Rouat, and D. Létourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 2, 2003, pp. 1228–1233.

[3] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 2007, pp. 121–124.

[4] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *The Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 384–391, 2000.

[5] Y. Tamai, S. Kagami, H. Mizoguchi, Y. Amemiya, K. Nagashima, and T. Takano, "Real-time 2 dimensional sound source localization by 128-channel huge microphone array," in *13th IEEE International Workshop on Robot and Human Interactive Communication*, 2004, pp. 65–70.

[6] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2193–2206, 2013.

[7] C. P. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *IEEE Transactions on Speech and Audio processing*, vol. 6, no. 5, pp. 476–488, 1998.

[8] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *The Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[9] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *The Journal of the Acoustical Society of America*, vol. 116, no. 5, pp. 3075–3089, 2004.

[10] H. Viste and G. Evangelista, "Binaural source localization," in *Proc. 7th International Conference on Digital Audio Effects (DAFx-04), invited paper*, no. LCAV-CONF-2004-029, 2004, pp. 145–150.

[11] V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Korner, "A probabilistic model for binaural sound localization," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 5, pp. 982–994, 2006.

[12] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 68–77, 2010.

[13] T. May, S. van de Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Transactions on Audio, Speech, and Language processing*, vol. 19, no. 1, pp. 1–13, 2011.

[14] T. May, N. Ma, and G. J. Brown, "Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 2679–2683.

[15] A. Deleforge and R. Horaud, "2D sound-source localization on the binaural manifold," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2012, pp. 1–6.

[16] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1503–1512, 2012.

[17] F. Keyrouz, "Advanced binaural sound localization in 3-D for humanoid robots," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 9, pp. 2098–2107, 2014.

[18] D. S. Talagala, W. Zhang, T. D. Abhayapala, and A. Kamineni, "Binaural sound source localization using the frequency diversity of the head-related transfer function," *The Journal of the Acoustical Society of America*, vol. 135, no. 3, pp. 1207–1217, 2014.

[19] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of the direct-path relative transfer function for supervised sound-source localization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2171–2186, 2016.

[20] X. Zhong, L. Sun, and W. Yost, "Active binaural localization of multiple sound sources," *Robotics and Autonomous Systems*, vol. 85, pp. 83–92, 2016.

[21] M. Zohourian and R. Martin, "Binaural speaker localization and separation based on a joint ITD/ILD model and head movement tracking," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 430–434.

[22] N. Ma, T. May, and G. J. Brown, "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2444–2453, 2017.

[23] D. Wang and G. J. Brown, "Computational auditory scene analysis: Principles, algorithms, and applications," 2006.

[24] G. R. Karthik and P. K. Ghosh, "Subband selection for binaural speech source localization," *Proc. Interspeech 2017*, pp. 1929–1933.

[25] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 99–102.

[26] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[27] G. Schwarz *et al.*, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[28] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[29] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (methodological)*, pp. 1–38, 1977.