音環境知能技術を活用した聴覚支援システムのプロトタイプの開発

Developing a prototype of hearing support system using sound environment intelligence

石井カルロス[1]，劉超然[1]，Jani Even[1]
Carlos ISHI, Chaoran LIU, Jani EVEN
国際電気通信基礎技術研究所
[1]石黒浩特別研究所
[1]ATR/HIL

carlos@atr.jp, chaoran.liu@irl.sys.es.osaka-u.ac.jp, even@atr.jp

Abstract

難聴者に対して従来の補聴器が持つ問題点を解決するため、提案者らがこれまで培ってきた音環境知能（音の時空間的構造化）技術を発展させ、利用者と利用環境に適応して、聞き取るべき音（対話相手の声、呼びかけ、アラームなど）とその妨げとなる不要・不快な音（ドア、エアコン、対話相手以外の声など）を取捨選択でき、更に選択された音に対する空間的感覚を再構築できる聴覚支援システムの実現を目的とする。本稿では、聴覚支援システムのプロトタイプの開発について進捗を報告する。

1 はじめに

世界各国で共通して、その国における人口の 1 割～2 割程度が難聴・聴覚障害を持っているといわれている。2009 年の日本補聴器販売店協会による「補聴器供給システムの在り方に関する研究」報告書の中で、日本の難聴者人口は 15.7% (1944 万人)と報告されている。そのうち、自覚のない難聴者(7.2%)、自覚がある難聴者(4.5%)、ほとんど使用しない補聴器所有者(1.0%)、常時または随時使用の補聴器所有者(2.7%)に分かれる。高齢者の難聴は、神経細胞などの老化現象としての老人性難聴で、65 歳以上では 25～40%、75 歳以上では 40～66%の割合で見られる。高齢化に伴い、難聴者数は更に増加すると予想される。

日本で補聴器を使っている人は 400 万人程度であり、難聴者のうち 5 人に 1 人しか補聴器を使っていないことになる。補聴器を途中で使わなくなる難聴者も多く、その理由として以下が記載されている：

「会話中、周りの音も大きくて、肝心な言葉が聞き取れない。」
「テレビのセリフが聞こえない。」
「コップをテーブルに置いた音、ドアの音などが大きくてびっくりする。」
「水音、新聞をめくる音などが気になる。」
「ピーピー音（ハウリング）が鳴る。」
「玄関チャイムが聞こえない。」
「自分の声が最も大きく聞こえる。」
「自分の声が変に聞こえて気持ち悪い。」
「声や音が聞こえても、どこから鳴ったのかが分からない。」

一般の補聴器は、マイクが補聴器に埋め込まれているため、周囲の雑音も増幅されてしまうという根本的な問題がある。ハウリング（ピーピー音）も起きやすく利用者に苦痛を感じさせる。最近の補聴器は、デジタル処理の導入により、周波数帯域ごとの音量調整や騒音抑制などの機能が埋め込まれ、性能は上がっている。ハウリング防止の信号処理も施しているものがあるが、その分、音量を抑える必要があり、重度難聴には十分な音量が出力できない。

補聴器コンサルタントによると、補聴器を止める原因は多くの場合、利用者に合った補聴器を選べていない、または設定が難しく誤った設定で使用しているためとされているが、それらが適切であっても補聴器単体による快適さ（聞こえやすさ）には限界がある。

ピンマイクやペン型などの遠隔マイクにより、FM 経由で遠隔の声を送受信する機能を持つ補聴器もあるが、遠隔のマイク周辺の雑音も増幅する問題や、音の方向を感知するための空間的情報も保たれない問題が残る。

空間的情報の伝達においては、マイク埋め込みの補聴器を両耳にかけることにより、ある程度解決されるが、自分の声も大きく聞こえる問題は残る。

遠隔センサによる空間的情報の伝達における問題点は、センサと音源の相対的角度が利用者と音源の相対的角度と異なることが原因で、音の方向情報を取得できる多チャンネルの場合でも生じる。聴覚支援を目的に多チャンネルのマイクロホンアレイ技術を活用した研究は国内外多数あるが、ほとんどが一つの音源を強調させ、モノラル信号を出力する仕組みで、空間的情報が失われる。

以上、従来の補聴器の問題点は、次の(1)～(3)にまとめられる。

(1) 利用者に必要な音と不要な音を選択することができない。
(2) 音の空間的情報が失われる。
(3) 設定が複雑で使いにくい。

提案者らは、これまで環境内に設置した複数のマイクロホンアレイと人位置検出システムを組み合わせて、いつ誰がどこで発話したのかを検出できる音環境知能の基盤技術の研究開発を進めてきた。本提案では、環境センサネットワークによる音環境知能技術を発展させ、上述の従来の補聴器の問題点を解

決することにより、利用者が快適な日常生活を可能とする聴覚支援システムの実現を目的とする。

まず問題点(1)に対し、環境内の個々の音を分離することにより、これまで補聴器単体では出来なかった、利用者に対して必要な音と不要な音を取捨選択的に制御可能な聴覚支援システムを提案する。環境センサの利用により、対象音の強調と不要音の抑圧に加え、ハウリングの問題および自分の声が大きく聞こえる問題も解決できる。これにより、従来の補聴器より音量を上げることができ、対象となる音や声が聞きやすくなる。

問題点(2)に対処するために、環境センサにより分解された個々の音源に対し、センサと利用者の相対的な位置や向きに応じた音像（音の空間的情報の感覚）の再構築手法を提案する。これにより、どの方向から音が鳴ったのか、といった空間的情報の知覚を可能にする。

問題点(3)に対して、時と場と利用者の好みに合わせて、環境センサにより、利用者の注意対象および利用者向けの発話対象をシステムが自動的に学習する手法を提案し、利用者の負担を最小限にする対象音選択インタフェースを追究する。スマートホンやタブレットを用いたものや利用者の頭部動作を用いたジェスチャ入力など、複数の利用者層を想定した数種類のインタフェースを提案する。

図１に提案する聴覚支援システムの利用場面のイメージ図を示す。老人ホームや介護施設などの供用空間で複数の利用者が環境センサを共用して、ドアの音や足音、食器の音など、不要・不快な音を抑圧し、利用者が注意している対話相手の声やテレビの音（利用者指向の注意対象）と利用者に背後から話しかけられた声（利用者向けの発話対象）を強調し、利用者に応じてその場で聞くべき音のみを提供するようなシステムの実現を目指す。

本論文では、上記の問題点(1)と(2)を解決するための基本的機能を備えた聴覚支援システムの概要を紹介し、プロトタイプの実現に向けた進捗を報告する。

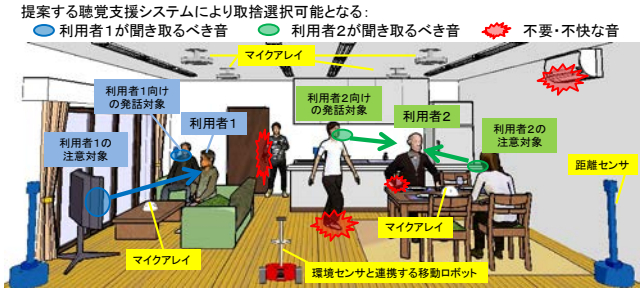


図１．提案する聴覚支援システムの利用場面の例。

2 関連研究

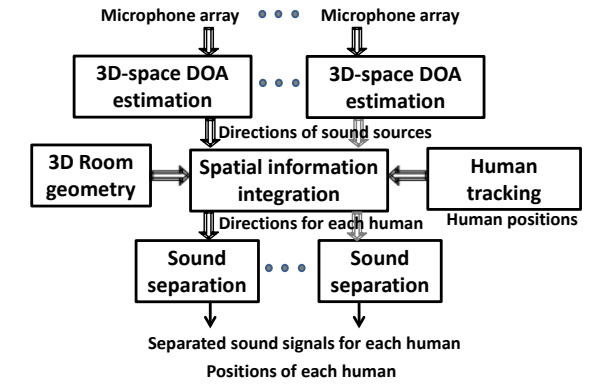
補聴器への応用においては、バイノーラル処理（両耳に装着した補聴器のマイクを利用した信号処理）が、国内外で多く研究されている。例えば、猿渡らは，バイノーラル信号を用いてブラインド信号処理とポストフィルタリングを中心に，両耳補聴器に適用した研究を進めてきた[高藤 2008]. 鵜木らは、「聞き耳」型補聴システムの研究開発が実施し[鵜木 2013]、中藤らも、高齢者の聴覚機能の低下に向けた聴覚支援システムに関する研究を進めている[中藤 2014]。

海外でも、補聴器への応用として、アレイ処理や多チャンネル Wiener フィルタなどの信号処理を導入した研究が多い（[Desloge 1997],[Bogaert 2008], [Cornelis 2012]など）。しかし、その殆どは利用者が装着した補聴器のバイノーラル処理を施したものであり、本研究のように環境センサを利用したものはあまり存在しない。

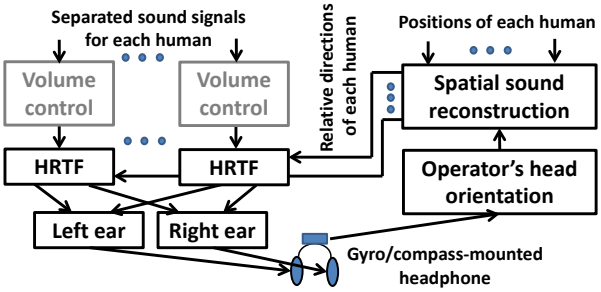
3 提案する聴覚支援システム

図２に提案システムのブロック図を示す。提案システムは二つの部分から構成される。一つは環境センサネットワーク側の音源位置推定・トラッキングと複数人の音源分離であり，もう一つは利用者側の頭部回転トラッキングと空間的感覚の合成である。

本システムの構成は、著者らが先行研究[Liu 2015]で提案した遠隔操作ロボットシステムにおいて音響臨場感を操作者に伝達する手法と類似している。その違いとして、遠隔操作システムでは操作者は遠隔地にいるが、本研究で提案する聴覚支援システムの場合は、利用者は環境センサと同じ場にいる。また、先行研究で報告したシステムに対し、本研究では主に音源分離のリアルタイム実装およびアルゴリズムの改善を進めた。



(a) 環境センサ側の処理



(b) 利用者側の処理

図２．提案する聴覚支援システムの概要

環境センサネットワーク側の処理では，まず，各マイクロホンアレイによって音の３次元到来方向（DOA）を推定する。環境とアレイの位置関係と各音源の DOA を統合することにより，３次元上での人位置（厳密には口元の位置）情報が得られる。人位置情報は，ヒューマントラッキングシステムにより，非発声時にも常時追跡されている。次に，推定した人位置情報に基づいて各人の音声を分離し，位置情報と合わせて利用者側のシステムに送信する。

利用者側の処理では，まず，人位置情報と利用者の顔の向きによって，左右のチャンネルに対応した最適な頭部伝達関数（HRTF: Head-Related Transfer Functions [Cheng 2001]）をデータベースから選択する。次に，分離した音声に畳み込み演算を行い，ステレオヘッドフォンに再生する。利用者の頭部回転トラッキングには，ヘッドホンの上部に取り付けたジャイロセンサーとコンパスを用いた。また，分離した各音源のボリュームは，独立して調節することができるユーザインタフェースを開発した。

3.1 ３次元音源定位

音源定位に関して，まず，各マイクロホンアレイで DOA 推定を行う。複数のアレイによる DOA 情報と人位置情報を統合することで，音源の３次元空間内の位置を推定する。

実環境での音の DOA 推定は広く研究されてきた。MUSIC 法は，複数のソースを高い分解能で定位できる最も有効な手法の一つである。この手法を使うには事前に音源数が必要であるため，本研究では[Ishi 2009]で提案した解決法を用いる。音源数を固定した数値に仮定し，閾値を超えた MUSIC スペクトルのピークを音源として認識する。この研究で使用した MUSIC 法の実装は 100 ms ごとに 1 度の分解能を有しており，2 GHz のシングルコア CPU でリアルタイムに探索することができる。

聴覚支援システムにおいて，利用者にとって最も重要な音源は人の音声である。本研究では人の声を抽出するために，複数の 2D-LRF（Laser Range Finder）で構成したヒューマントラッキングシステムを使用した[Glas 2007]。複数のマイクロホンアレイからの DOA 推定出力と LRF のトラッキング結果が同じ位置で交差すれば，そこに音源がある可能性が高い[Ishi 2013]。本システムでは 2D の LRF を用いているため、人位置情報は 2D に限られる。ここでは、検出された音源の位置が口元の高さの範囲内にあるかの制限もかけている（z = 1 ~ 1.6m）[石井 2014] [Ishi 2015]。無音区間や音源方向推定が不十分な区間では、最後に推定された口元の高さと最新の 2D 位置情報を用いて，音源分離を行う。

3.2 音源分離

音源分離では，選択された複数の人物を並列に分離する。図３に処理の流れを示す。

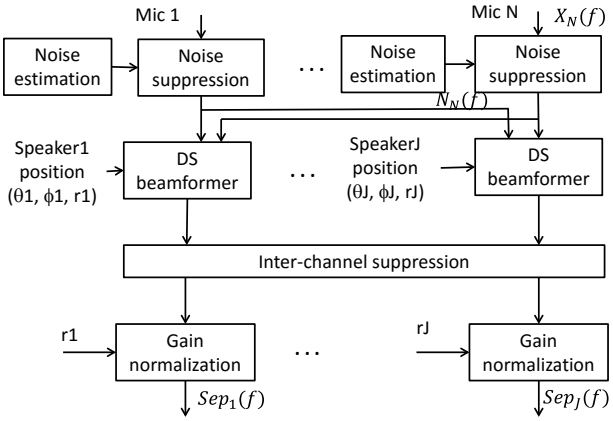


図 3. 音源分離の処理の流れ

まず、分離の第 1 ステップとして、エアコンなどの定常雑音抑圧（noise suppression）をチャンネル毎に行う。定常雑音抑圧手法として式(1)に示すように Wiener filter を用いる。

$$H_{WFi}(f) = \frac{1}{1+\frac{N_i(f)}{X_i(f)}} \quad (1)$$

定常雑音（$N_i(f)$）は、対象となる人の声が存在しない区間での平均スペクトルとして推定する。

定常雑音抑圧処理は、ポストフィルタとして、ビームフォーマを施した後に行うことも可能であるが、ここでは、musical ノイズの発生を抑えるため、ビームフォーマの前に施す。

次に、音源定位部から得られる方向（方位角、仰角）と距離情報を基に、ビームフォーマを施す。ここでは計算量が少なく且つロバストな DS ビームフォーマ（Delay-Sum Beamformer）を用いて、対象となる人の声を強調する。フレーム長は 32 ms で，シフト長は 10 ms である。

本研究で使用した 16 チャンネルのマイクロホンアレイ（半球 30cm にマイクを配置した形状）の DS ビームフォーマのレスポンスの特徴として，低周波領域の分解能が低いことが挙げられる。そのため，無指向性雑音の低周波成分が分離音に多く混在してしまい、臨場感の伝達に悪影響を与える可能性がある。

空間に指向性音源 S と無指向性雑音源 N が存在すると仮定した場合，DS ビームフォーマの出力は以下の形になる：

$$Y_{DS}(f) = \boldsymbol{w}_{Sdir}(f) \cdot S(f) + \int_0^{2\pi} (\boldsymbol{w}_{\theta}(f) \cdot N(f)) d\theta \quad (2)$$

$Y_{DS}(f)$は周波数 f に対応したビームフォーマの出力で，S_{dir} は信号の方向，$\boldsymbol{w}_{Sdir}$ は S_{dir} 方向のビームフォーマレスポンスを指す。式の二つ目の項目は，分離音声に混在する雑音を表している。この雑音成分を低減させるために，各周波数に以下のようなウェイトを掛けた。

$$w_{norm}(f) = \frac{1}{\int_0^{2\pi} \boldsymbol{w}_\theta(f)\, d\theta} \quad (3)$$

$$Y_i = \sum_f w_{norm}(f) \cdot Y_{DS}(f) \quad (4)$$

Y_i はウェイト掛けした後のビームフォーマ出力である。

また、DS ビームフォーマのみでは、十分な音源分離が出来ず、チャンネル間の信号（妨害音）の漏れを抑えるための処理（inter-channel suppression）を行う。妨害音抑圧処理には、式(5)に示すように Wiener filtering を用いる。

$$H_{WFi}(f) = \frac{1}{1+\frac{I_i(f)}{Y_i(f)}} \quad (5)$$

$$I_i(f) = \max_{j \neq i}\{Y_j(f)\} \quad (6)$$

$I_i(f)$ は式(6)に示すように、分離された対象音以外の音源の中で、最も強い周波数成分を表す。上述の妨害音抑圧処理の一つの問題点として、同じ方向に対象音と妨害音が存在する場合、対象音に歪みが生じる可能性が高い。そこで、ここでは対象音と妨害音の差が 5 度以内であれば、抑圧処理を行わない制約を設けた。

$$I_i(f) = \frac{|dir_1 - dir_2|}{5} I_i(f), \text{ if } |dir_1 - dir_2| < 5 \quad (7)$$

最後に、音源とマイクロホンアレイの距離によって、観測される音圧が異なるため、距離による振幅の正規化（gain normalization）を施す。

$$g_j = \frac{1}{r_i} \quad (8)$$

3.3 音の空間的情報の再構築

環境センサ側から提供される分離音を受信し、利用者と対象音源の相対的位置関係を考慮して、音の空間的感覚を再構築する。処理としては、複数音源に対する音量調整と、頭部伝達関数（HRTF）を用いた音像の合成となる。

まず、音量調整に関しては、各音源とアレイの間の距離による違いを補正するため，分離された各音源に対して、それぞれの距離によって以下のように正規化を行う。

$$g_i = \frac{\sum_{n=1}^{N} dist_n - dist_i}{(N-1)\cdot\sum_n^N dist_n} \quad (9)$$

$$Y_i = g_i \cdot Y_{PF,i} \quad (10)$$

ここで，N は音源の数で，$dist_n$ は n 番目の音源とアレイの距離を表す。g_i は i 番目の音源に掛ける正規化ファクタで，Y_i は i 番目の音源の分離結果を示している。

音像の合成においては、一つの音源を特定の方向から聞こえるようにするため，その方向に対応した HRTF によってフィルタリングするステレオ化方法が一般的である。本研究では，一般公開されている KEMAR (Knowles Electronics Manikin for Acoustic Research) ダミーヘッドの HRTF データベースを利用した[Gardner 1995]。KEMAR は HRTF 研究のために一般的な頭部サイズを使って作られたダミーヘッドで，データベースには空間からのインパルス信号に対するダミーヘッドの左右耳のレスポンスとして，仰角-40 度から 90 度までの総計 710 方向のインパルス応答が含まれている。各インパルス応答の長さは 512 サンプルで，サンプリング周波数は 44.1 kHz である。

前述のように, HRTF を用いて動的に音像を合成するには，頭部の向きのリアルタイム検出が必要である。このため，本研究ではヘッドホンの上部にジャイロセンサーとコンパスを取り付け，頭部回転のトラッキングを行った。角度情報はシリアルおよびブルートゥース経由のいずれかでシステムに送られる。音場の合成に使う方向は音源方向から頭部角度を引いたもので，この方向に対応した左右チャンネルのインパルス応答がデータベースから選出され，分離音と畳み込み演算を行った音声が利用者の両耳に再生される。

4 予備的評価

現段階では、開発したシステムの定性的な評価に留まっている。まず、研究室内での予備的評価により、wiener filter のパラメータは、$\alpha = 1$, $\beta = 0.001$ とした。式(8)の振幅の正規化に関しては、距離が大きくなり過ぎると、背景雑音も増幅されてしまうため、距離による正規化は 2m までと制限した。

著者らの研究所のオープンハウス（2015 年 10 月）で開発したシステムのデモを行った。デモシステムとして、LRF 2 個で人位置推定を行い、ポスター前のテーブル上にマイクアレイ 1 個を設置して、訪問者にヘッドホンをかけてもらい、ポスターの周りにいる人のうち、強調したい人をマウスの左クリックで選択し、抑圧したい人を右マウスで選択する機能を設けたインタフェースを開発した。取捨選択型の機能を体験していただいた方々には、高評価の感想をいただいた。一つの大きな課題として、処理後の音声が再生される遅延が大きすぎることが挙げられる。現在は遅延が 300ms 程度で、対話相手が目の前で発話している状況では、口の動きや頷きなどのタイミングが音声とずれて見えるため、違和感があるという意見が多かった。この遅延は、処理時間に加え、再生用のバファリングも大きな原因となっているが、ハードウェアの開発により、短くすることは可能である。その他、訪問した一般の高齢者の方も数人体験していただき、使いたいので早く実用化していただけないかとの意見もいただいた。

分離音の音質においては、研究室で予備評価を行った際、図３に表示したすべての処理を用いるのが最も聞きやすかった。しかし、オープンハウス会場では、入力の noise suppression を用いない方が分離音の音質が良かった。研究室では空調音が最も強い背景雑音源であるが、ポスター会場の雑音はカクテルパーティ効果のようなバブル雑音が大きかったため、システムを起動した際に推定した背景雑音のレベルが大きく、定常雑音の wiener filter 処理を施すと強い歪みが生じてしまうことが原因と考えられる。定常雑音の推定については、今後改善する予定である。また、システム全体の詳細な評価についても今後進める予定である。

謝辞

本研究は、総務省 SCOPE の委託研究によるものである。

参考文献

[高藤 2008] 高藤、森、猿渡、鹿野 (2008). SIMO モデルに基づく ICA と頭部伝達関数の影響を受けないバイナリマスク処理を組み合わせた両耳聴覚補助システム、電子情報通信学会技術研究報告. EA, 応用音響 108(143), 25-30, 2008.

[鵜木 2013] 鵜木祐史. 「聞き耳」 型補聴システムの研究開発. 「戦略的情報通信研究開発推進事業 SCOPE ）」 平成 25 年度新規採択課題 http://www.soumu.go.jp/main_content/000242634.pdf

[中藤 2014] 高齢者の聴覚機能の低下に向けた聴覚支援システムに関する研究、文部科学省科学研究費基盤研究(C) 、2014 年 04 月 ～ 2017 年 03 月

[Desloge 1997] J.G. Desloge, W.M. Rabinowitz, and P.M. Zurek, Microphone-Array Hearing Aids with Binaural Output- Part I: Fixed-Processing Systems, IEEE Trans. Speech Audio Processing, vol. 5, no. 6, pp. 529□542, Nov. 1997.

[Bogaert 2008] Bogaert, T.V., Doclo, S., Wouters, J., Moonen, M. The effect of multimicrophone noise reduction systems on sound source localization by users of binaural hearing aids, J. Acoust. Soc. Am. 124 (1), 484-497, July 2008

[Cornelis 2012] Cornelis B., Moonen, M., Wouters, J. Speech intelligibility improvements with hearing aids using bilateral and binaural adaptive multichannel Wiener filtering based noise reduction. J Acoust Soc Am. 2012 Jun;131(6):4743-4755.

[Liu 2015] Liu, C., Ishi, C., Ishiguro, H., Bringing the Scene Back to the Tele-operator: Auditory Scene Manipulation for Tele-presence Systems, Proc. ACM/IEEE International Conference on Human Robot Interaction (HRI 2015), USA. 279-286, March, 2015.

[Cheng 2001] Cheng, C. I., Wakefield, G. H. Introduction to head-related transfer functions (hrtfs): Representations of hrtfs in time, frequency, and space. *J. Acoust. Soc. Am*, 49(4):231-249, April 2001.

[Ishi 2009] Ishi, C. T., Chatot, O., Ishiguro, H., Hagita, N. Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments. Proceedings of the *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 09)*. 2027‑2032. 2009.

[Glas 2007] Glas, D.F. et al, 2007. Laser tracking of human body motion using adaptive shape modeling. In Proceedings of the *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2007)*, 602-608. 2007.

[Ishi 2013] Ishi, C., Even, J., Hagita, N. (2013). Using multiple microphone arrays and reflections for 3D localization of sound sources. In Proc. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2013)*, 3937-3942, Nov., 2013.

[石井 2014] 石井カルロス寿憲, Jani EVEN, 萩田紀博, (2014) "複数のマイクロホンアレイと人位置情報を組み合わせた音声アクティビティの記録システムの改善", 第３２回日本ロボット学会学術講演会, Sep. 2014.

[Ishi 2015] Ishi, C., Even, J., Hagita, N. (2015). “Speech activity detection and face orientation estimation using multiple microphone arrays and human position information,” *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2015)*, pp. 5574-5579, Sep., 2015.

[Gardner 1995] Gardner, W. G., Martin, K. D. HRTF measurements of a KEMAR. *J. Acoust. Soc. Am*. 97(6):3907-3908, Jun. 1995.