Received : Agustus 2019	Accepted : September 2019	Published : Oktober 2019

Pemisahan Suara Tercampur Berdasarkan Karakteristik Binaural

Mifta Nur Farid[1]
[1]Institut Teknologi Kalimantan

[*]miftanurfarid@lecturer.itk.ac.id

Abstract

The human auditory system is able to focus its hearing on an interlocutor and ignore other sources of sound. In a preliminary study it was explained that the location of the sound source to both ears (binaural) made an important contribution in the human hearing system. Research on sound separation based on sound localization has been done before. In that study, sound separation was mixed based on only one binaural characteristic, namely interaural time difference (ITD). So in this study, the separation of mixed sounds will be based on two binaural characteristics, namely interaural time difference (ITD) and interaural level difference (ILD). In this study the training process and the separation process were carried out. The training process was carried out 75 times and the separation process was 1440 times with 4 angular variations (30 °, 20 °, 10 ° and 5 °) and 3 variations of SIR (10 dB, 5 dB, and 0 dB). The training process aims to obtain the opportunity density function between ITD and ILD against the relative strength (RS) of the target voice. ITD and ILD values are obtained based on the results of cross-correlation between signals received by the left ear and right ear. The separation process aims to separate the target sound from the mask based on the density function of the opportunity. Separation is carried out by a binary mask which is estimated based on the pattern of changes in ITD and ILD values to the RS value of the target signal which is calculated statistically based on the opportunity density function. The quality of the separation results is measured using an objective method, namely signal-to-noise ratio (SNR). A high SNR value is 3.15 dB for the female target and 3.44 dB for the male target and 3.15 dB for the female voice target.

Keywords: binaural characteristics, binaural hearing, SNR.

Abstrak

Sistem pendengaran manusia mempu memfokuskan pendengarannya pada seorang lawan bicara dan mengabaikan sumber suara lainnya. Dalam sebuah studi awal dijelaskan bahwa lokasi dari sumber suara terhadap kedua telinga (binaural) memberikan kontribusi penting dalam sistem pendengaran manusia. Penelitian tentang pemisahan suara yang berdasarkan pada lokalisasi bunyi telah dilakukan sebelumnya. Pada penelitian tersebut, pemisahan suara tercampur hanya berdasarkan satu karakteristik binaural, yaitu *interaural time difference* (ITD). Maka pada penelitian ini, akan dilakukan pemisahan suara tercampur berdasarkan kedua karakteristik binaural, yaitu *interaural time difference* (ITD) dan *interaural level difference* (ILD). Pada penelitian ini dilakukan proses pelatihan dan proses pemisahan. Proses pelatihan dilakukan sebanyak 75 kali uji dan proses pemisahan sebanyak 1440 kali uji dengan 4 variasi sudut (30°, 20°, 10°, dan 5°) dan 3 variasi SIR (10 dB, 5 dB, dan 0 dB). Proses pelatihan bertujuan untuk memperoleh fungsi kepadatan peluang antara ITD dan ILD terhadap kekuatan relatif (RS) dari suara target. Nilai ITD dan ILD diperoleh berdasarkan hasil korelasi silang antara sinyal yang diterima oleh telinga kiri dan telinga kanan. Proses pemisahan bertujuan untuk memisahkan suara target dari masker berdasarkan fungsi kepadatan peluangnya. Pemisahan dilakukan oleh *binary mask* yang diestimasi berdasarkan pola perubahan nilai ITD dan ILD terhadap nilai RS dari sinyal target yang dihitung secara statistik berdasarkan fungsi kepadatan peluangnya. Kualitas hasil pemisahan diukur dengan menggunakan metode objektif yaitu *signal-to-noise ratio* (SNR), yaitu 3.15 dB untuk target perempuan dan 3.44 dB untuk target laki-laki dan 3,15 dB untuk target suara perempuan.

Kata kunci: karakteristik binaural, binaural hearing, SNR.

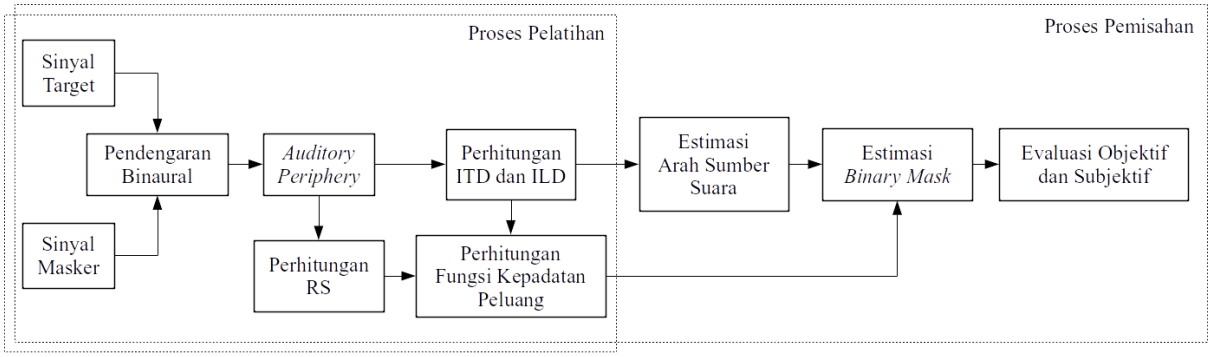
1. Pendahuluan

Sistem pendengaran manusia mampu memisahkan beberapa sumber bunyi sekaligus secara bersamaan dan memfokuskan pendengaran pada satu sumber suara meskipun bunyi latar cukup keras dan disertai percakapan beberapa orang lainnya. Fenomena ini dikenal dengan *the cocktail party effect*[1]. Istilah "*cocktail party processing*" diciptakan dalam sebuah studi awal terhadap *the cocktail party effect*, dalam studi ini menggambarkan bahwa sistem pendengaran binaural memberikan kontribusi penting dalam analisa pendengaran yang memungkinkan kita untuk memisahkan dan melokalisir sumber suara[2].

Sistem pendengaran binaural adalah sistem pendengaran yang menggunakan kedua telinga. Dalam sistem tersebut, jika posisi suatu sumber suara tidak berada dalam bidang simetri vertikal atau bidang median maka salah satu telinga akan dibayangi oleh kepala sedangkan telinga lain terbuka penuh terhadap sumber bunyi. Akibatnya terjadi perbedaan tingkat tekanan bunyi yang terdengar pada kedua telinga yang disebut *Interaural Level Difference* (ILD) serta perbedaan waktu tempuh sumber suara terhadap kedua telinga tersebut yang disebut *Interaural Time Difference* (ITD)[3].

Perubahan nilai ITD dan ILD memiliki pengaruh secara statistik terhadap perubahan kekuatan relatif suara (RS). Sehingga Binary Mask (BM) dapat diestimasi dari nilai RS yang didapat berdasarkan perubahan nilai nilai ITD dan ILD[4].

Telah dilakukan pemisahan suara berdasarkan sistem pendengaran binaural[5], namun karakteristik binaural yang digunakan adalah ITD. Maka pada penelitian ini, akan dilakukan pemisahan suara tercampur dengan dua sensor mikrofon dari dua sumber suara berdasarkan kedua karakteristik binaural ITD dan IL



Gambar 1. Metoda Penelitian

2. Metoda Penelitian

Penelitian ini dilakukan dengan 2 proses utama yaitu proses pelatihan dan proses pemisahan. Seperti yang ditunjukkan pada Gambar 1, proses pelatihan terdiri dari 5 tahap yaitu pendengaran binaural, *auditory periphery*, perhitungan nilai ITD, ILD, dan RS kemudian dilakukan perhitungan fungsi kepadatan peluang dari ketiga nilai tersebut. Sedangkan proses pemisahan adalah pendengaran binaural, *auditory periphery*, perhitungan nilai ITD dan ILD, perhitungan estimasi sudut sumber suara, estimasi BM, dan yang terakhir adalah evaluasi objektif terhadap hasil pemisahannya.

2.1. Proses Pelatihan

Hasil pendengaran binaural diperoleh dari hasil konvolusi antara suara mono terhadap data *head-related transfer function*

(HRTF) telinga kiri dan kanan. Suara mono terdiri dari suara target dan masker. Suara target adalah suara yang diinginkan sedangkan suara masker adalah suara pengganggu. Data HRTF yang digunakan adalah *HRTF CIPIC Database*.

HRTF CIPIC Database adalah data pengukuran *head-related impulse response* (HRIR) pada telinga kiri dan kanan dari KEMAR manikin dalam ruang kedap (*unechoic room*). Titik sumber suara berada pada jarak 1,4 m terhadap titik tengah kepala manusia pada bidang horisontal atau azimut φ dan bidang median atau elevasi (φ, δ). Titik sumber suara yang digunakan adalah titik 1 (0°, 0°), titik 2 (5°, 0°), titik 3 (10°, 0°), titik 4 (20°, 0°) dan titik 5 (30°, 0°)[5].

Tabel 1. Konfigurasi sinyal target dan masker pada proses pelatihan.

Sinyal Target	Azimut & Elevasi	Sinyal Masker	Azimut & Elevasi	SIR (dB)
T1	(0°, 0°)	M1	(30°, 0°)	10
T2		M2	(20°, 0°)	
T3		M3	(10°, 0°)	5
T4		M4		0
T5		M5	(5°, 0°)	

Selain variasi lokasi sumber suara, juga dilakukan variasi tingkat tekanan suara. Variasi tingkat tekanan suara ditunjukkan dalam nilai *signal-to-interference ratio* atau SIR. Konfigurasi sinyal target dan masker yang digunakan pada penelitian ini ditunjukkan pada Gambar 1.

Sinyal suara yang digunakan dalam pada percobaan ini adalah data rekaman suara percakapan laki-laki sebanyak 5 kalimat sebagai sinyal target (T1, T2, T3, T4, T5) dan suara percakapan perempuan sebanyak 5 kalimat sebagai sinyal masker (M1, M2, M3, M4, M5) berbahasa Indonesia dengan frekuensi sampling 16 kHz[5].

Konfigurasi yang dilakukan adalah sebagai berikut. Sinyal target terdiri dari 5 kalimat dan selalu pada posisi yang sama yaitu (0°, 0°). Sedangkan sinyal masker terdiri dari 5 kalimat dengan 5 posisi bergantian yaitu (30°, 0°), (20°, 0°), (10°, 0°), dan (5°, 0°). Masing-masing target dan masker memiliki 3 nilai SIR yang berbeda yaitu 10 dB, 5 dB, dan 0 dB. Sehingga total percobaan yang dilakukan adalah 5 kalimat × 1 posisi target × 5 kalimat masker × 4 posisi masker × 3 variasi SIR = 75 percobaan [4].

Auditory periphery adalah model dari telinga bagian tengah dan dalam[6]. Model ini terdiri dari filter gammatone orde 4, *middle-ear gain*, dan *hair cell activity*. Gammatone filter yang digunakan memiliki frekuensi tengah sebanyak 128 dari frekuensi 80 Hz hingga 5000 Hz[4].

Nilai ITD dan ILD diperoleh dengan metode korelasi silang $C(i, j, \tau)$ antara suara pada telinga kiri l_i dan kanan r_i hasil dari *auditory periphery* pada frekuensi tengah ke-i, *frame* ke-j dan *lag* ke-τ. Persamaan korelasi silang yang dilakukan ditunjukkan pada Persamaan 1.

$$C(i,j,\tau) = \frac{\sum_{k=0}^{K-1}(l_i(j-k)-\overline{l_K})(r_i(j-k-\tau)-\overline{r_K})}{\sqrt{\sum_{k=0}^{K-1}(l_i(j-k)-\overline{l_K})^2}\sqrt{\sum_{k=0}^{K-1}(r_i(j-k)-\overline{r_K})^2}} \quad (1)$$

Nilai RS atau kekuatan relatif adalah rasio antara sinyal target s_i terhadap sinyal tercampur, yaitu sinyal target s_i dan masker n_i, pada frekuensi tengah ke-i. Nilai RS

diperoleh dengan menggunakan persamaan berikut.

$$RS_i = \frac{\sqrt{\sum_t s(t)_i^2}}{\sqrt{\sum_t s(t)_i^2}\sqrt{\sum_t n(t)_i^2}} \quad (2)$$

Nilai ITD dan ILD digunakan sebagai masukan untuk menghitung fungsi kepadatan peluang dari RS > 0,5 dan RS ≤ 0,5. Kedua keadaan tersebut menunjukkan keadaan dari sinyal target terhadap masker. Kondisi pertama (RS > 0,5) adalah sinyal target lebih dominan daripada sinyal masker dan kondisi kedua (RS ≤ 0,5) adalah sebaliknya. Fungsi kepadatan peluang dihitung menggunakan Persamaan 3.

$$p(x|H_i) = \hat{f}_i(x) = \sum_{i=1}^{n} \frac{1}{nh_1 \dots h_d} \prod_{j=1}^{d} K\left(\frac{x_i - x_{ij}}{h_j}\right) \quad (3)$$

2.2. Proses Pemisahan

Pada proses pemisahan, langkah yang dilakukan sama seperti pada proses pelatihan namun sinyal keluaran dari *auditory periphery* merupakan sinyal suara tercampur antara sinyal target dan sinyal masker. Sinyal suara target yang digunakan pada proses pemisahan ini sebanyak 120 kalimat dengan sinyal suara masker yang selalu sama/tetap.

Tabel 2. Konfigurasi sinyal target dan masker pada proses pemisahan

Sinyal Target	Azimut dan Elevasi	Sinyal Masker	Azimut dan Elevasi	SIR (dB)
Suara laki-laki sebanyak 120 kalimat yang berbeda	(0°,0°)	Suara perempuan sebanyak 1 kalimat	(30°,0°)	10
			(20°,0°)	5
			(10°,0°)	0
			(5°,0°)	

Pada Tabel 2, proses pemisahan yang pertama dilakukan adalah pada titik 1 untuk sinyal target dan titik 5 untuk sinyal masker dengan nilai SIR 10 dB, 5 dB dan 0 dB. yang. Setelah suara target dan masker berhasil dipisah, maka akan dilakukan pemisahan untuk sinyal masker yang posisinya semakin mendekati sinyal target yaitu titik 4, titik 3 dan titik 2. Sehingga total percobaan adalah 12 kali dengan masing-masing percobaan sebanyak 120 kalimat.

Estimasi arah sumber suara diperoleh dari lokasi puncak hasil korelasi silang pada Persamaan 1. Lokasi puncak menunjukkan arah dari sumber suara dan jumlah puncak menunjukkan jumlah dari sumber suara.

BM adalah filter dalam ranah waktu-frekuensi. BM akan bernilai 1 jika $p(x|H_1) > p(x|H_2)$ dan akan bernilai 0 jika $p(x|H_1) \leq p(x|H_2)$.

Evaluasi objektif yang dipakai adalah *signal to noise ratio* (SNR) sesuai dengan Persamaan 4.

$$SNR = 20 \times \log_{10} \frac{\sum_t S_T(t)^2}{\sum_t (S_T(t)^2 - \sum_t S_E(t)^2)} \quad (4)$$

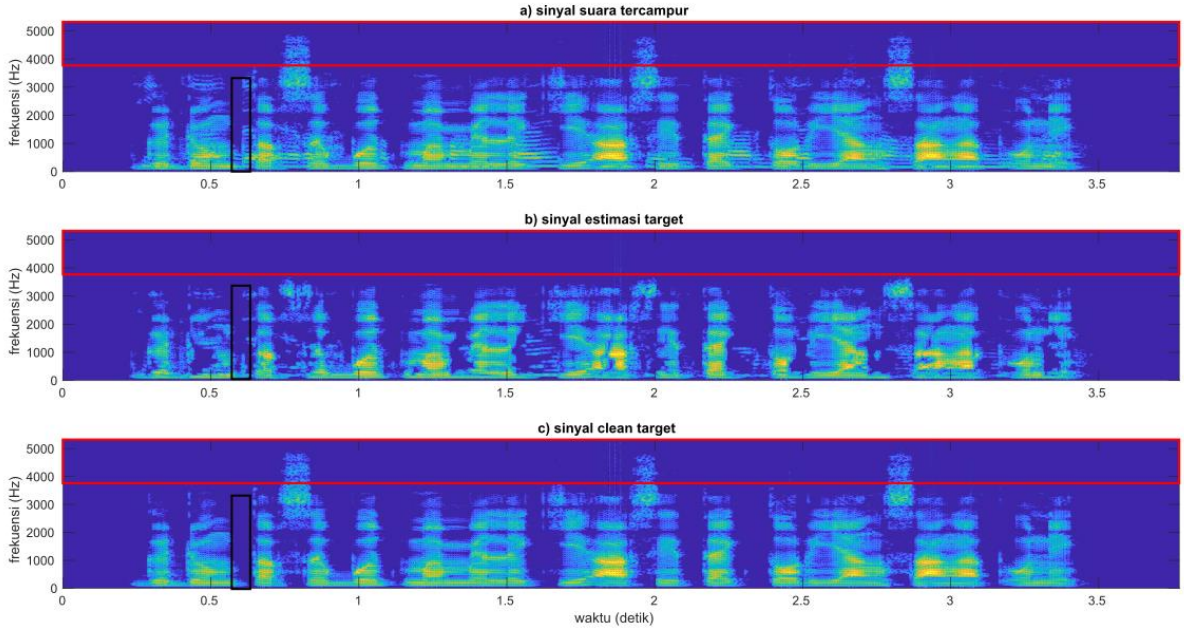
S_T adalah sinyal suara target sebelum pencampuran dan S_E adalah sinyal suara target hasil estimasi. Perhitungan SNR dilakukan pada 120 kalimat kemudian dirata-rata pada setiap konfigurasi yang dilakukan sebelumnya.

3. Hasil Penelitian

Gambar 2 adalah grafik spektrogram dari sinyal (a) suara tercampur, (b) estimasi target, dan (c) *clean* target. Sinyal suara tercampur adalah sinyal campuran antara sinyal target dan masker. Sinyal suara estimasi target adalah sinyal target hasil

pemisahan yang didapatkan dari penerapan filter BM terhadap sinyal tercampur. Sinyal suara *clean* target adalah sinyal suara target sebelum dicampur dengan sinyal masker.

Pada gambar tersebut, area bergaris hitam dan merah adalah salah satu wilayah yang difilter oleh BM. Namun tidak semua sinyal difilter dengan benar. Pada area berwarna merah, sinyal target yang seharusnya tidak difilter namun difilter oleh BM. Sebaliknya pada area berwarna hitam. Akibatnya sinyal suara estimasi target memiliki *musical noise*.

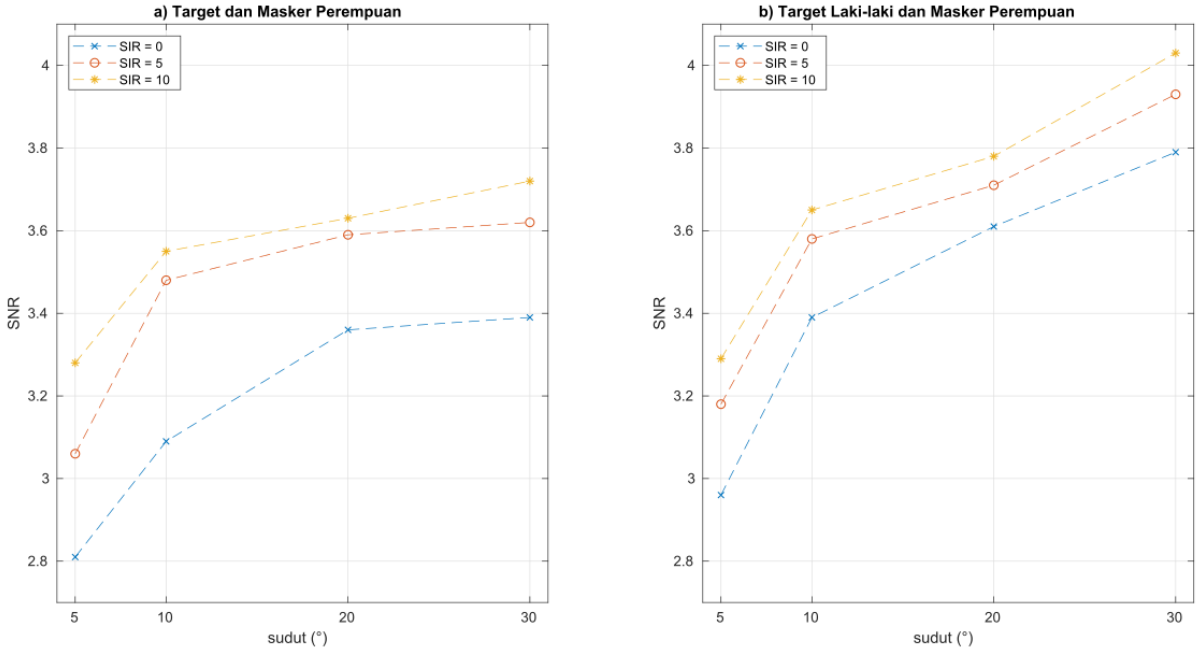


Gambar 2. Spektrogram sinyal a) tercampur, b) estimasi target, dan c) clean target

Musical noise terjadi karena hasil pemisahan yang diperoleh terdapat informasi suara yang ikut difilter oleh BM. Salah satu informasi yang difilter adalah wilayah frekuensi di atas 3700 Hz. Hilangnya informasi pada frekuensi diatas 3700 Hz karena nilai ITD yang diperoleh tidak linier seperti pada frekuensi dibawah 3700 Hz. Ketidaklinieran ini terjadi karena terdapat banyak puncak yang diperoleh pada saat perhitungan ITD yang disebabkan oleh resonansi yang terjadi pada kanal telinga. Resonansi ini disebabkan karena kanal telinga, yang memiliki panjang sebesar 2,3 cm, dimodelkan seperti tabung tertutup. Sehingga tabung tertutup tersebut akan mengalami resonansi pada panjang gelombang 9,2 cm atau frekuensi 3700 Hz.

Gambar 3 a) adalah SNR dengan target dan masker suara perempuan. Gambar 3 b) adalah SNR dengan target suara laki-laki dan masker suara perempuan. Pada Gambar 3 a) dan b), nilai SNR meningkat lebih tajam saat sudut 5° hingga 10° daripada saat sudut 10° hingga 30°. Hal ini terjadi karena sudut 5° adalah sudut dimana nilai ambiguitas tertinggi[7]. Secara keseluruhan, kedua grafik tersebut menunjukkan nilai SNR yang meningkat selaras dengan membesarnya sudut pemisah antara target dan masker. Hal ini terjadi karena efek bayangan kepada atau disebut *headshadow effect*[8]. Selain itu, nilai SNR juga meningkat selaras dengan

meningkatnya nilai SIR. Hal ini terjadi karena semakin besar nilai SIR maka energi suara target terhadap masker juga semakin besar.



Gambar 3. Hasil Evaluasi Objektif

Perbedaan jenis suara pada sinyal target menyebabkan perbedaan nilai SNR. Target suara laki-laki memiliki nilai SNR yang lebih tinggi daripada target suara perempuan, yaitu 3,44 dB untuk target suara laki-laki dan 3,15 dB untuk target suara perempuan. Hal ini disebabkan frekuensi dasar dari sinyal target. Ketika sinyal target memiliki jenis suara yang sama maka frekuensi dasar keduanya adalah sama. Sehingga ketika proses pemisahan suara target juga difilter oleh BM. Sebaliknya jika sinyal target memiliki frekuensi dasar yang berbeda dengan sinyal masker, suara target tidak difilter oleh BM.

4. Kesimpulan

Pemisahan suara tercampur berdasarkan karakteristik binaural-nya telah dilakukan. Kualitas hasil pemisahan ditunjukkan oleh nilai SNR yang tinggi yaitu 3.15 dB untuk target perempuan dan 3.44 dB untuk target laki-laki. Nilai SNR yang tinggi terjadi pada sumber suara target dan masker yang memiliki frekuensi dasar yang berbeda.

5. Saran

Saran dari penulis untuk penelitian selanjutnya adalah perhitungan *Interaural Time Difference* selain *Cross-Correlation* seperti *Threshold Detection* atau *Linier Phase Fit* untuk memperoleh hasil pemisahan yang lebih baik.

6. Daftar Pustaka

[1] E. C. Cherry, “Some Experiments on the Recognition of Speech, with One and with Two Ears,” *J. Acoust. Soc. Am.*, vol. 25, no. 5, pp. 975–979, Sep. 1953.

[2] J. F. Culling, M. L. Hawley, and R. Y. Litovsky, “The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources,” *J. Acoust. Soc. Am.*,

vol. 116, no. 2, pp. 1057–1065, Aug. 2004.

[3] A. Kohlrausch, J. Braasch, D. Kolossa, and J. Blauert, "An introduction to binaural processing," in *The Technology of Binaural Listening*, J. Blauert, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 1–32.

[4] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *Proc. Int. Jt. Conf. Neural Networks*, vol. 4, no. 4, pp. 2861–2866, Oct. 2001.

[5] G. R. Karthik and P. K. Ghosh, "Binaural speech source localization using template matching of interaural time difference patterns," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2018, vol. 2018-April, pp. 5164–5168.

[6] R. Meddis, "Simulation of mechanical to neural transduction in the auditory receptor," *J. Acoust. Soc. Am.*, vol. 79, no. 3, pp. 702–711, Mar. 1986.

[7] Y. Zhou, L. Balderas, and E. J. Venskytis, "Binaural ambiguity amplifies visual bias in sound source localization," *J. Acoust. Soc. Am.*, vol. 144, no. 6, pp. 3118–3123, Dec. 2018.

[8] J. G. W. Bernstein, G. I. Schuchman, and A. L. Rivera, "Head shadow and binaural squelch for unilaterally deaf cochlear implantees," *Otol. Neurotol.*, vol. 38, no. 7, pp. e195–e202, Aug. 2017.