Методы решения проблемы "Коктейльной вечеринки"

М.В. Губин

Научный руководитель: **Л.Б. Соколинский**

Челябинск, 2019

Проблема «коктейльной вечеринки»

Проблема «коктейльной вечеринки» –

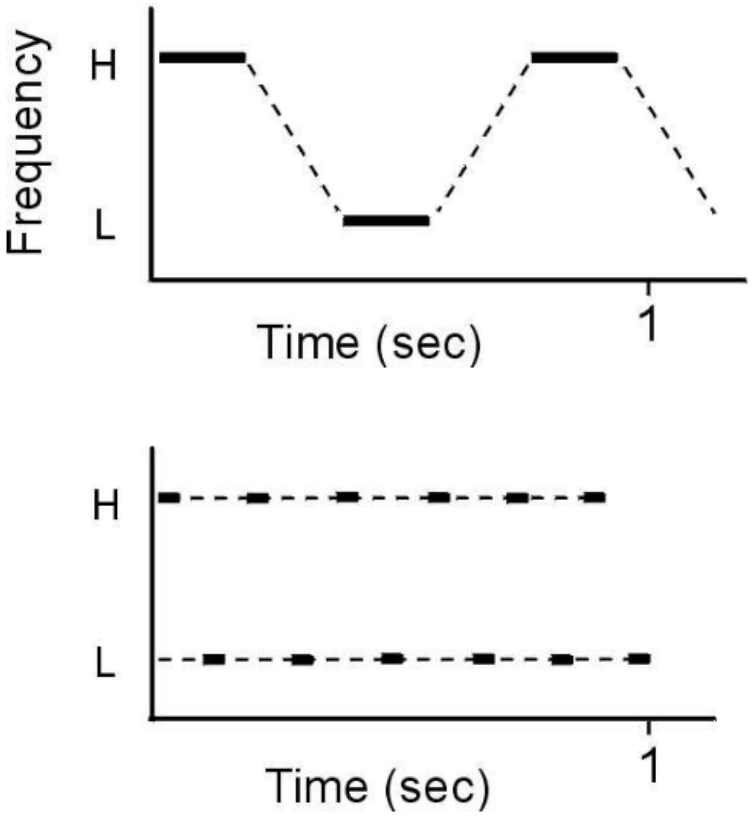
способность мозга сосредоточить свое слуховое внимание на конкретном стимуле, и отфильтровывать ряд других стимулов, например, когда человек может сосредоточиться на одном разговоре в шумной комнате.

Проблема коктейльной вечеринки была описана в 1950-х годах Колином Черри *, **дихотическая модель прослушивания**.

В 80-х Альберт Брегман начал изучать звуковую сегрегацию, назвав ее **«Анализ слуховых сцен»** (ASA, Auditory Scene Analysis).

* J.H. McDermott, “The cocktail party problem”, Current Biology, vol. 19, № 22, pp. 1024–1027, 2009

Анализ слуховых сцен, ASA



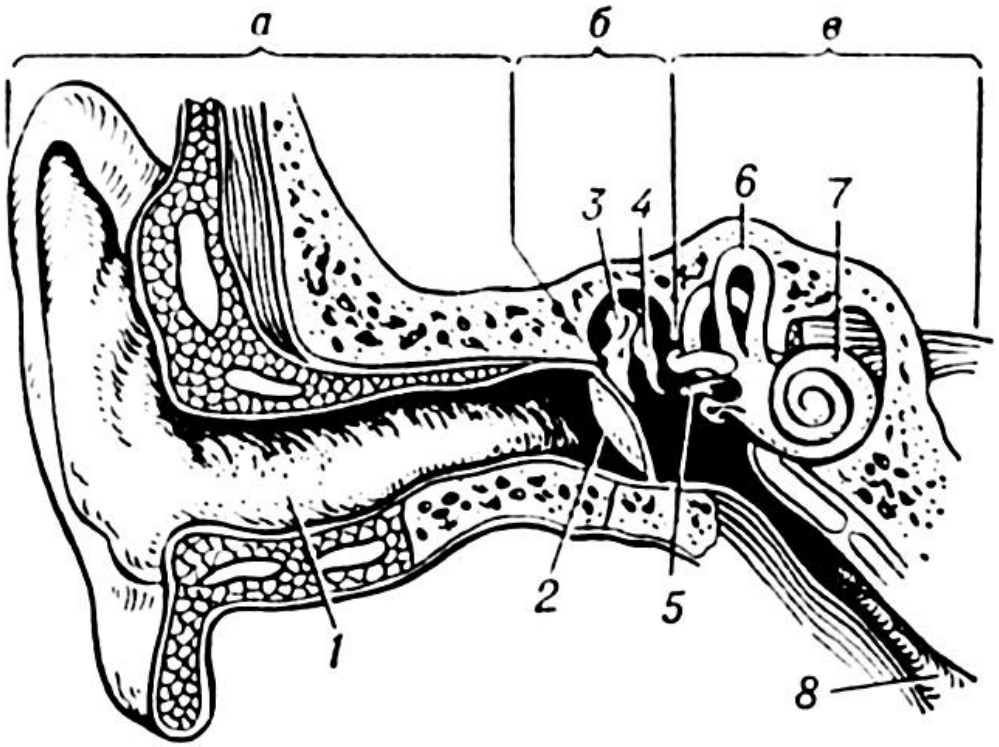
В психофизике это основная модель слухового восприятия (модель слушателя). Термин введен **Албертом Брегманом**.

Используется для понимания как слуховая система человека преобразует звук в воспринимаемые стимулы.

Три аспекта модели: сегментирование, интеграция и сегрегация.

* Albert Bregman “Auditory Scene Analysis: the perceptual Organization of Sound ”, The book, MIT Press, 1990.

Слуховая система человека



Стремечко передает свои колебания жидкости, находящейся внутри полости улитки. Дрожание жидкости воспринимается волосковыми клетками **кортиева органа**.

а – **наружное** ухо:

1 – слуховой проход,

2 – барабанная перепонка,

б – **среднее** ухо:

3 – молоточек,

4 – наковальня,

5 – стремечко,

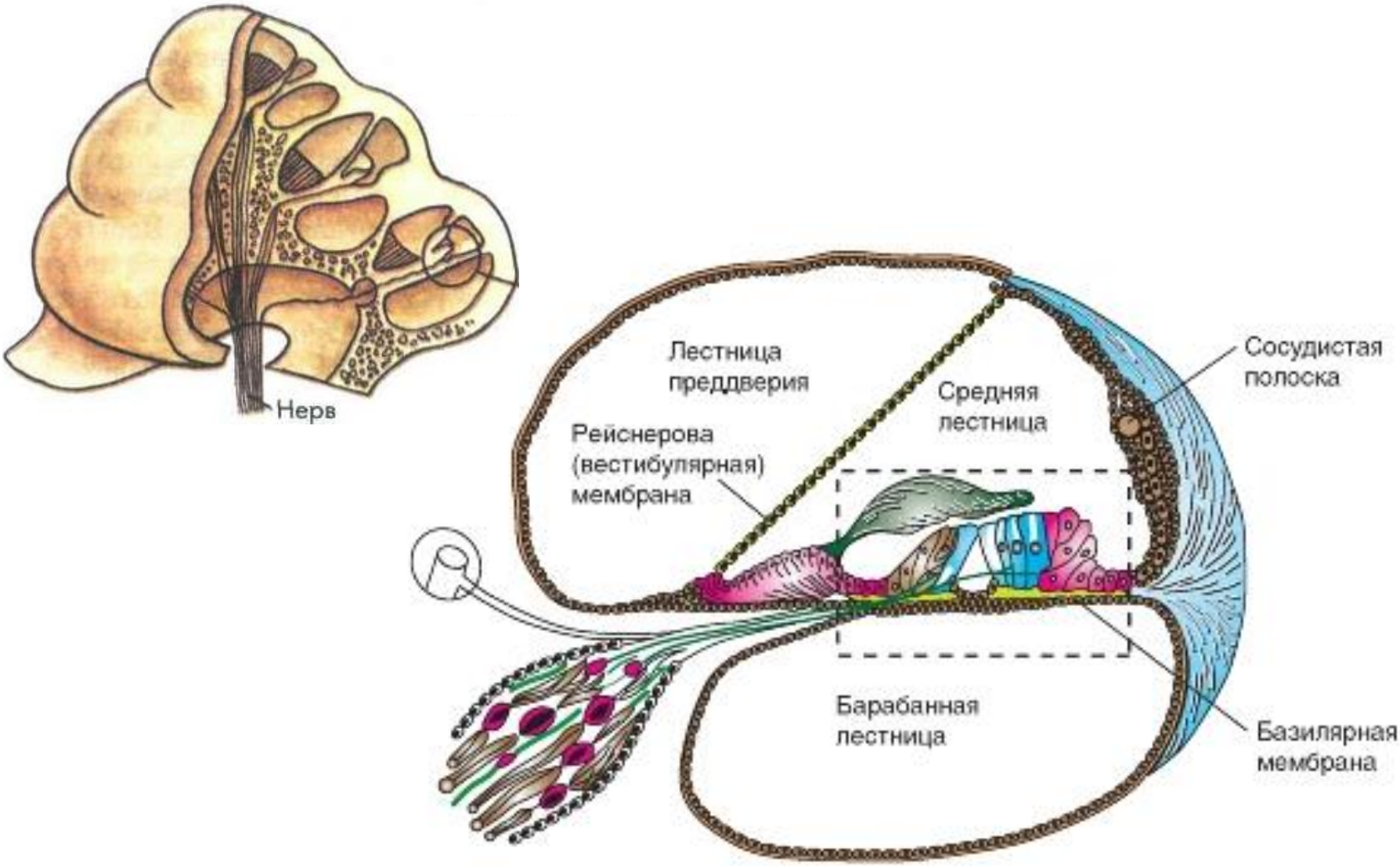
в – **внутреннее** ухо:

6 – полукружные каналы,

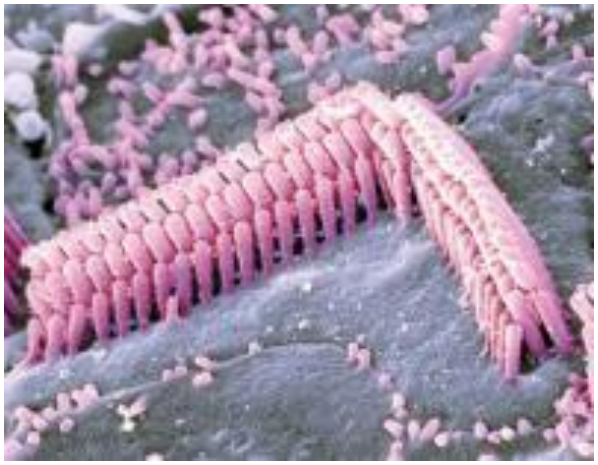
7 – улитка,

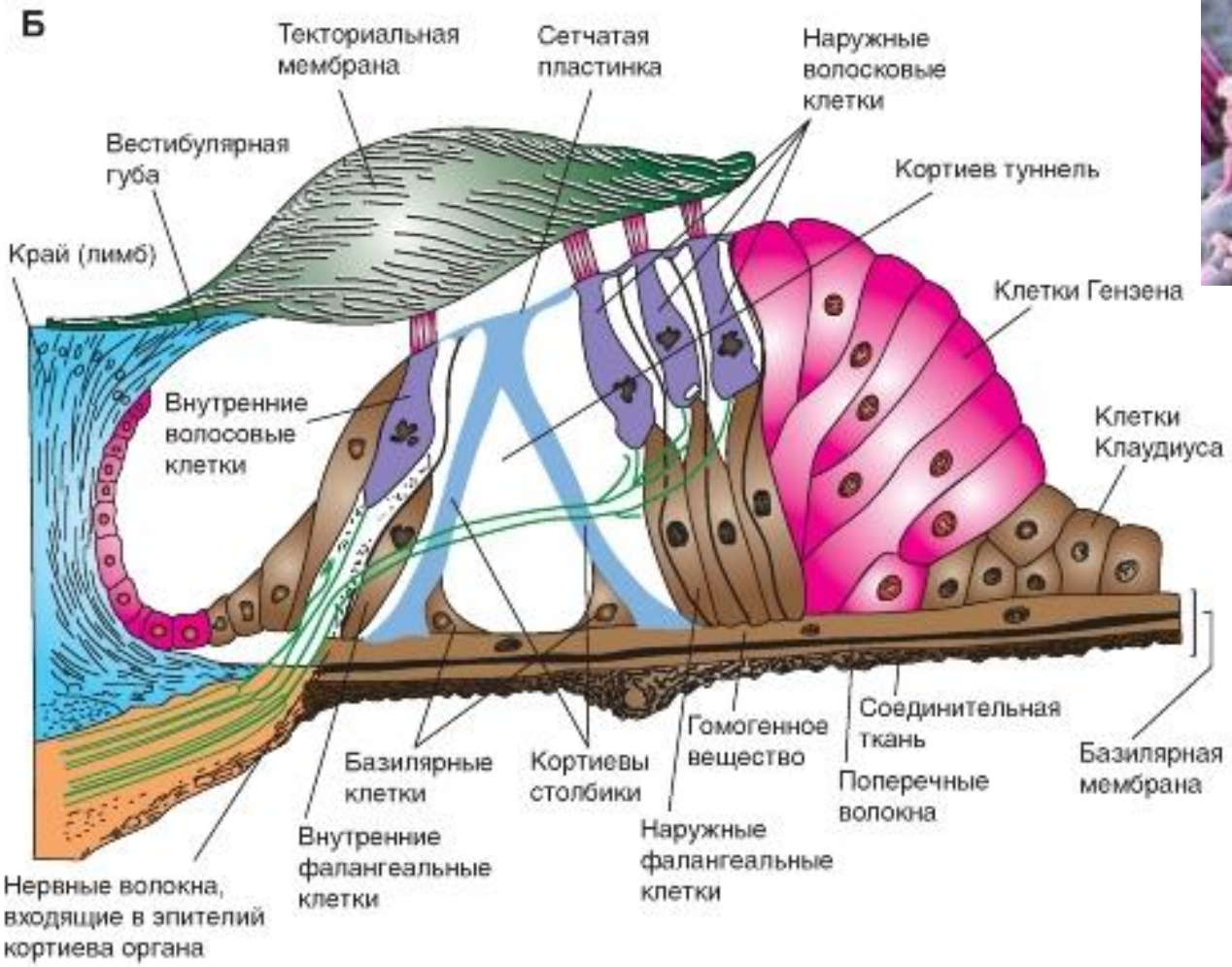
8 – евстахиева труба.

Улитка

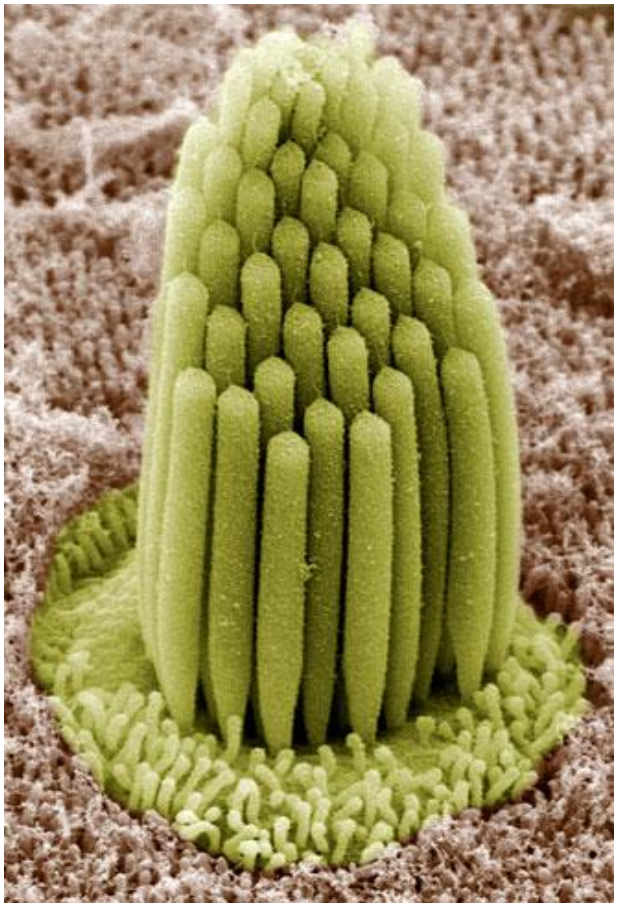

Нерв
Лестница преддверия
Средняя лестница
Сосудистая полоска
Рейснерова (вестибулярная) мембрана
Барабанная лестница
Базилярная мембрана

Кортиев орган

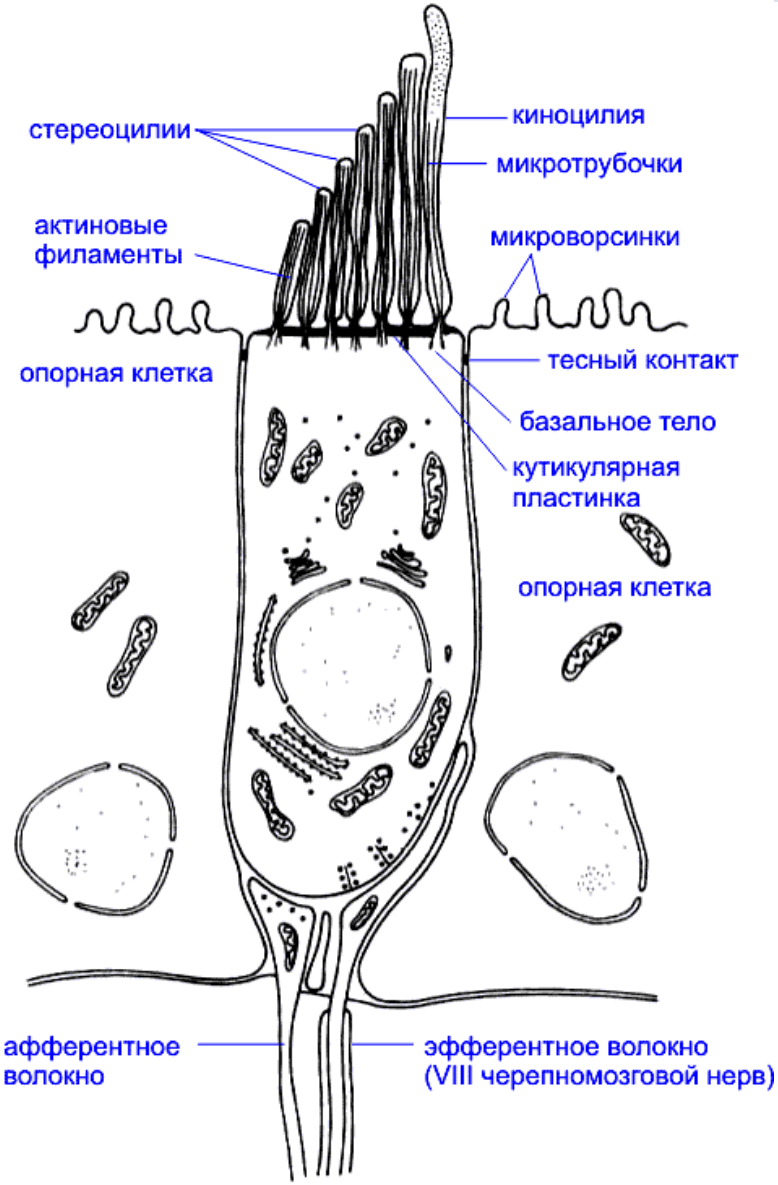




Стереоцилия внутреннего уха лягушки



Каждый пучок состоит из 30 — 150 тонких стержнеобразных отростков. Примерно 15000 волосковых клеток улитки человеческого уха.



Классификация методов разделения сигналов нескольких источников

- **Вычислительный анализ слуховых сцен** (CASA, Computational Auditory Scene Analysis):
  - Моноуральные;
  - Бинауральные;
  - Нейронные модели CASA;
  - Анализ музыкальных звуковых сигналов;
  - Нейронное перцептивное моделирование.
- **Слепое разделение сигналов** (BBS, Blind Source Separation):
  - Анализ независимых компонент (ICA, **Independent Component Analysis**);
  - Измерение диаграммы направленности **(Radiation pattern measurements).**
- **Нейронные сети.**

Вычислительный анализ слуховых сцен

CASA (Computational auditory scene analysis) –

это анализ слуховой сцены вычислительными средствами.

По сути, системы CASA являются системами **«машинного прослушивания»**, которые нацелены на разделение смесей источников звука так же, как это делают люди.

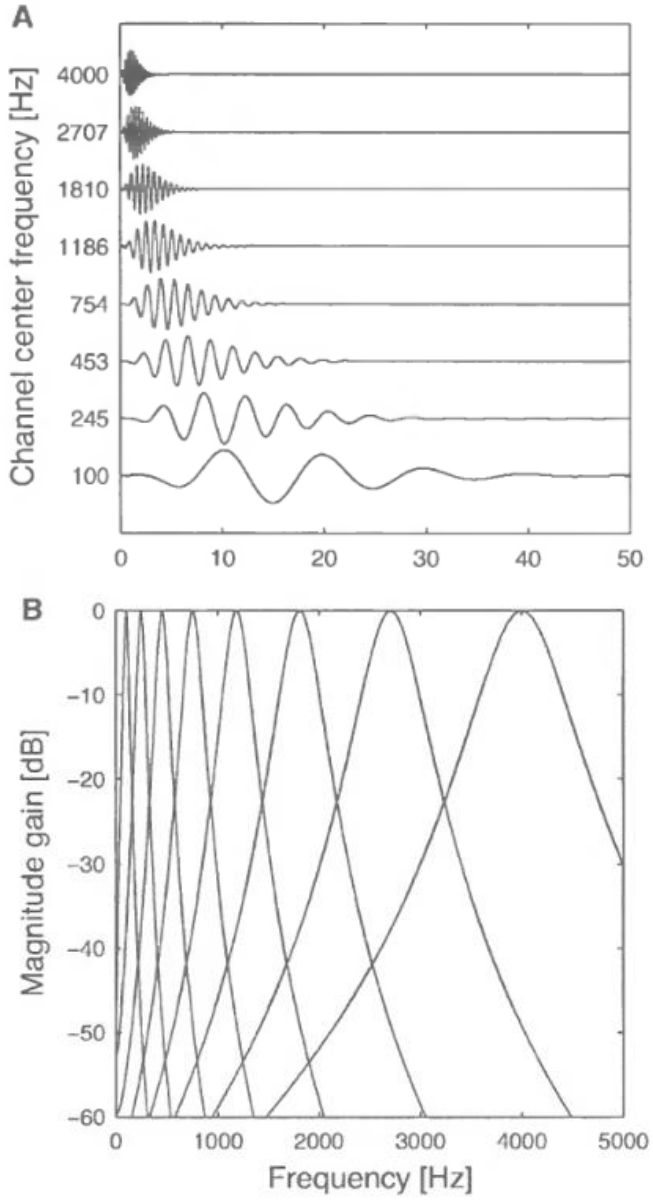
CASA отличается от области **слепого разделения сигналов** тем, что она основана на механизмах **слуховой системы человека** и, таким образом, использует не более двух микрофонных записей акустической среды.

* Computational Auditory Scene Analysis: Principles, Algorithms, and Applications (Wang, D. and Brown, G.J., Eds.; 2006)

Этапы обработки CASA

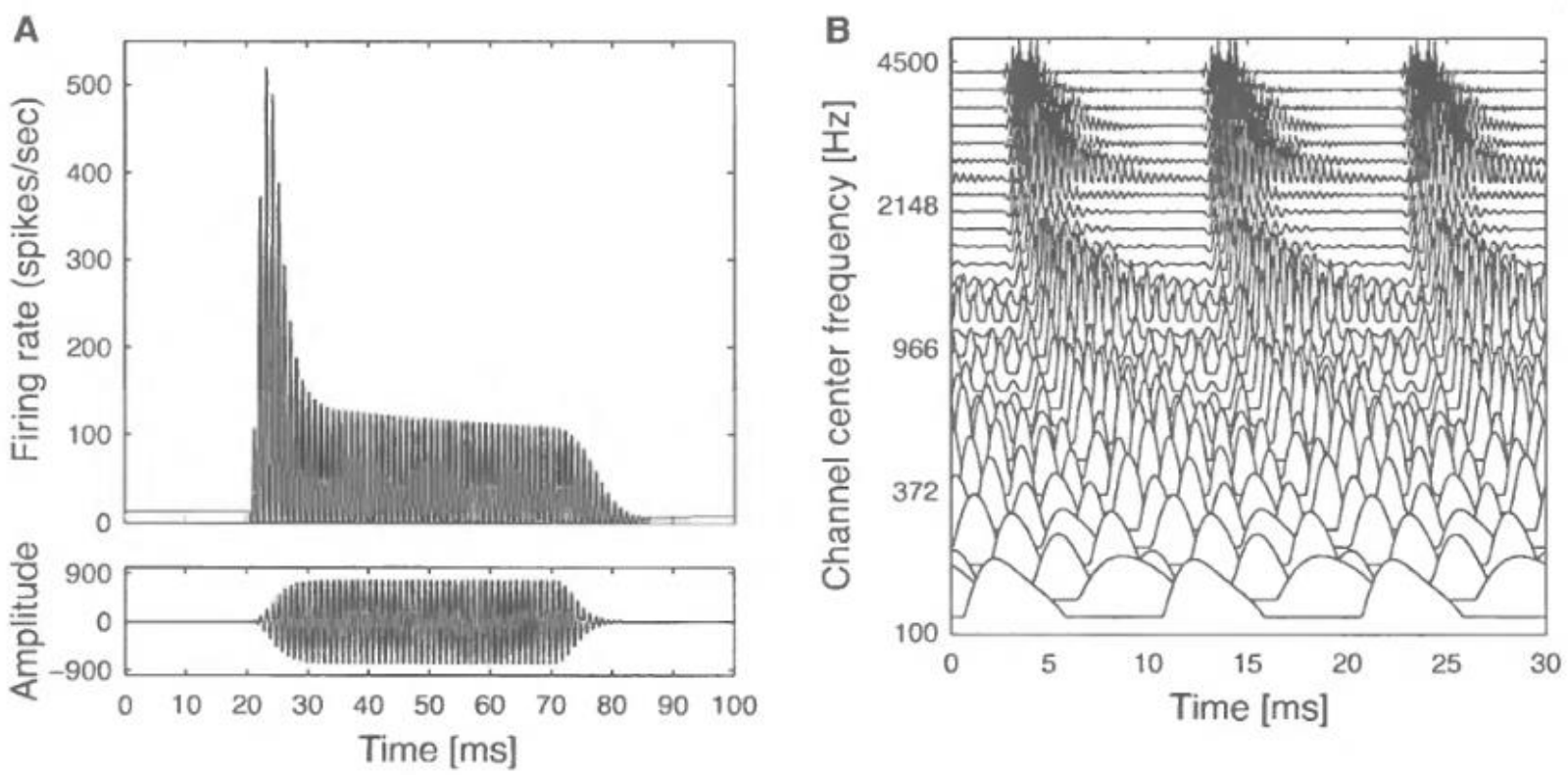
- **кохлеаграмма** – создает частотно-временное представление входного сигнала, сигнал разбивается на разные частоты, которые выбираются улиткой и волосковыми клетками (применение различных фильтров для усиления различных частот);
- **коррелограмма** – автокорреляция через частоту, позволяющую определить основной тон (частоту) источника звука;
- **кросс-коррелограмм** – уши принимают аудиосигналы в разное время, источник звука можно определить с помощью задержек, извлекаемых из двух ушей;
- **частотно-временные маски** – для отделения источника звука, системы CASA маскируют кохлеаграмму, эта маска (фильтр Винера) усиливает целевые области источника и подавляет остальные;
- **ресинтез** – реконструирует аудиосигнал из группы сегментов, достигается путем инвертирования кохлеаграммы, могут быть получены высококачественные повторно синтезированные речевые сигналы.

Кохлеаграмма



- Кохлеаграмма создает **частотно-временное представление** входного сигнала.
- Для моделирования такой избирательности применяют набор **гамматон-фильтров** (полосовые фильтры).
- На рисунке А входной сигнал разделен 8 гамматон-фильтрами с равномерным распределением частот от 100Гц до 4 кГц.
- На рисунке В представлены частотные характеристики гамматон-фильтров.

Кохлеограмма слухового нерва

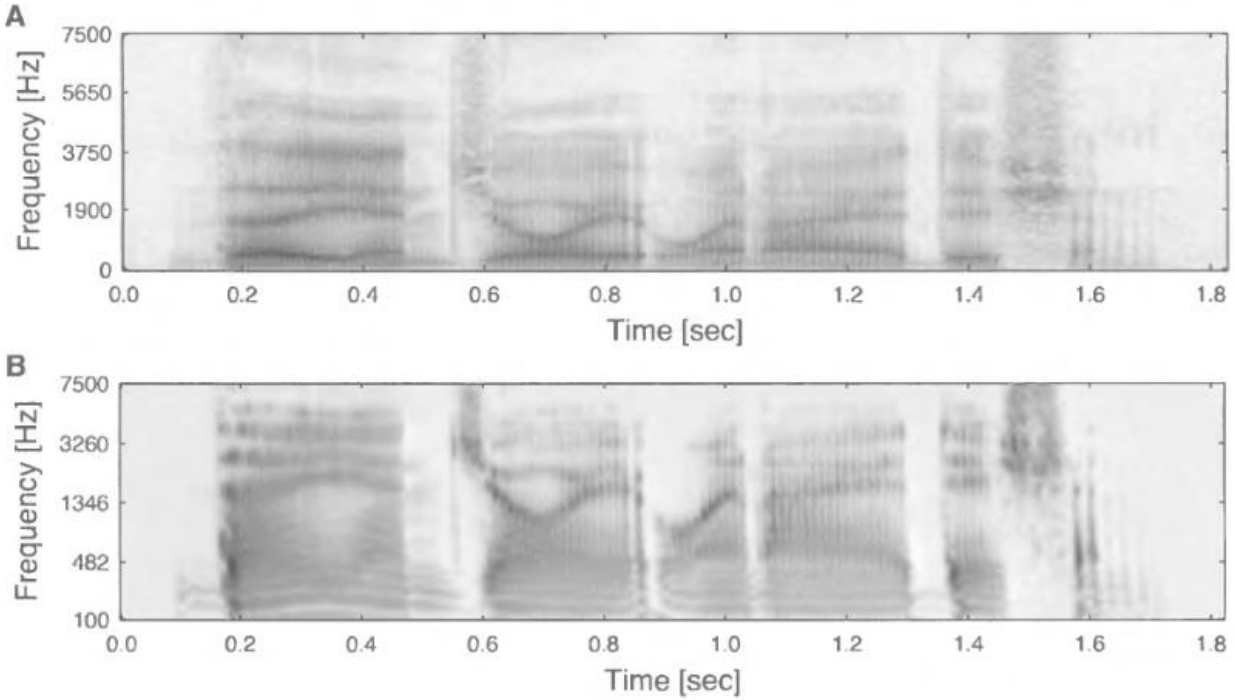


(A) Нижняя панель показывает реакцию гамматонового фильтра с центральной частотой основного тона 500 Гц , верхняя панель показывает соответствующий вывод модели волосковых клеток.

(B) Характер нейронной активности с использованием набора гамма-фильтров и модели волосковых клеток для установившегося гласного /эр/ с основной частотой 100Гц.

Кохлеограмма в спектрограмму

Представление спектрограммы может быть получено путем сглаживания временных рядов кохлеограммы, связанных с каждым частотным каналом, понижающей дискретизацией и сопоставлением результирующих значений с цветом или серой шкалой.



Спектрограмма (A) и кохлеаграмма (B) для высказывания «они наслаждаются этим, когда я слушаю». Темные пиксели указывают области с высокой энергией, а светлые пиксели указывают области с низкой энергией.

Коррелограмма

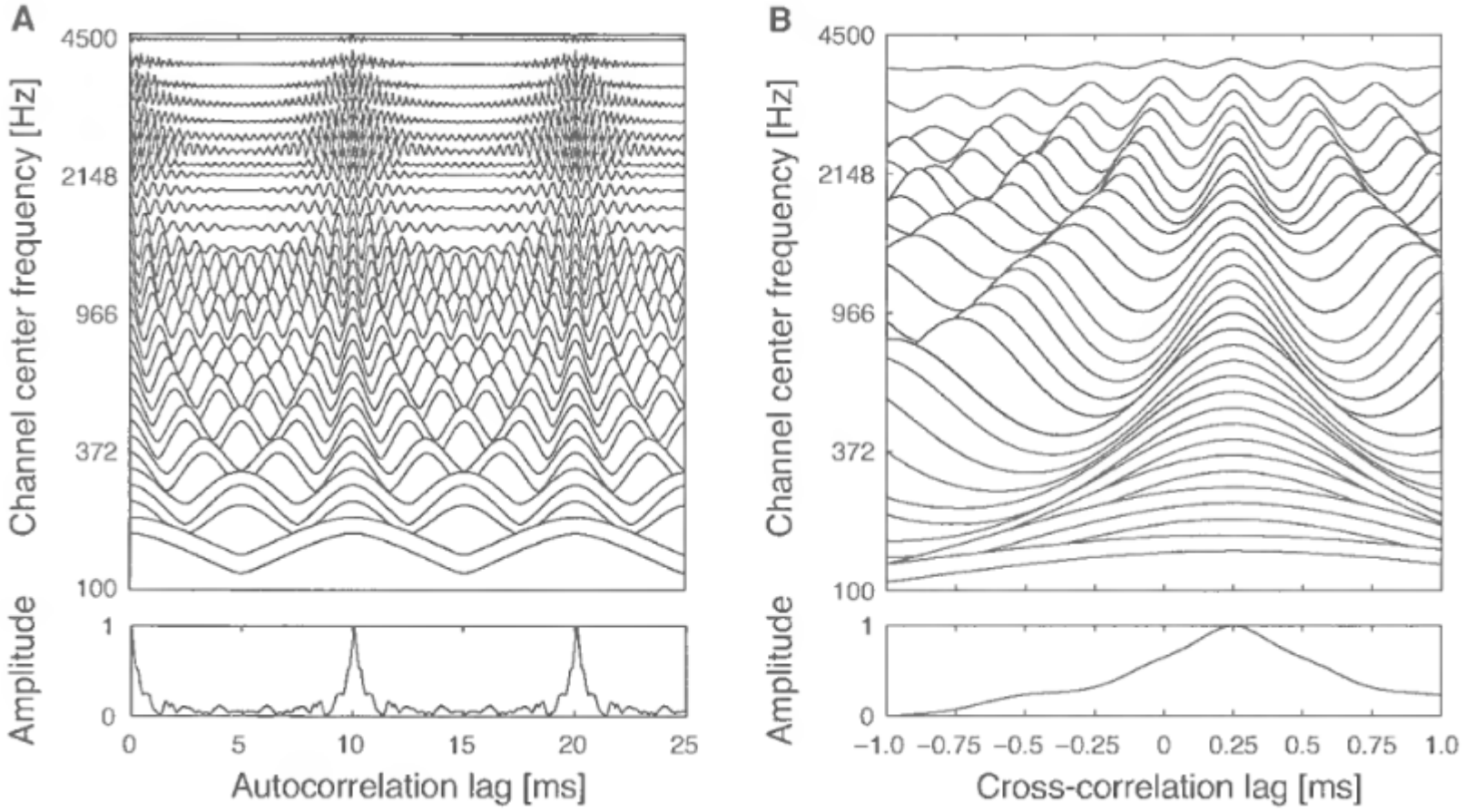
Коррелограмма

- модель восприятия основного тона, основанная на **автокорреляции** активности слухового нерва во временной области.

Составляет основу многих вычислительных моделей оценки **основной частоты** (FO) и **разделения звука** на основе FO.

Автокорреляция — статистическая взаимосвязь между последовательностями величин одного ряда, взятыми со сдвигом, например, со сдвигом по времени или частоте.

Коррелограмма и Крос-коррелограмма



(A) Коррелограмма установившегося гласного **er** с основной частотой 100 Гц (верхняя панель) и сводная коррелограмма (нижняя панель).

(B) Кросс-коррелограмма установившегося гласного **er**, представленная в **бинауральной слуховой модели** с межвременной разницей во времени 0,25 мс (верхняя панель) и сводной кросс-коррелограммой (нижняя панель).

Функции для CASA

$$g_{f_c}(t) = t^{N-1} \exp[-2\pi t b(f_c)] \cos(2\pi f_c t + \phi) u(t)$$

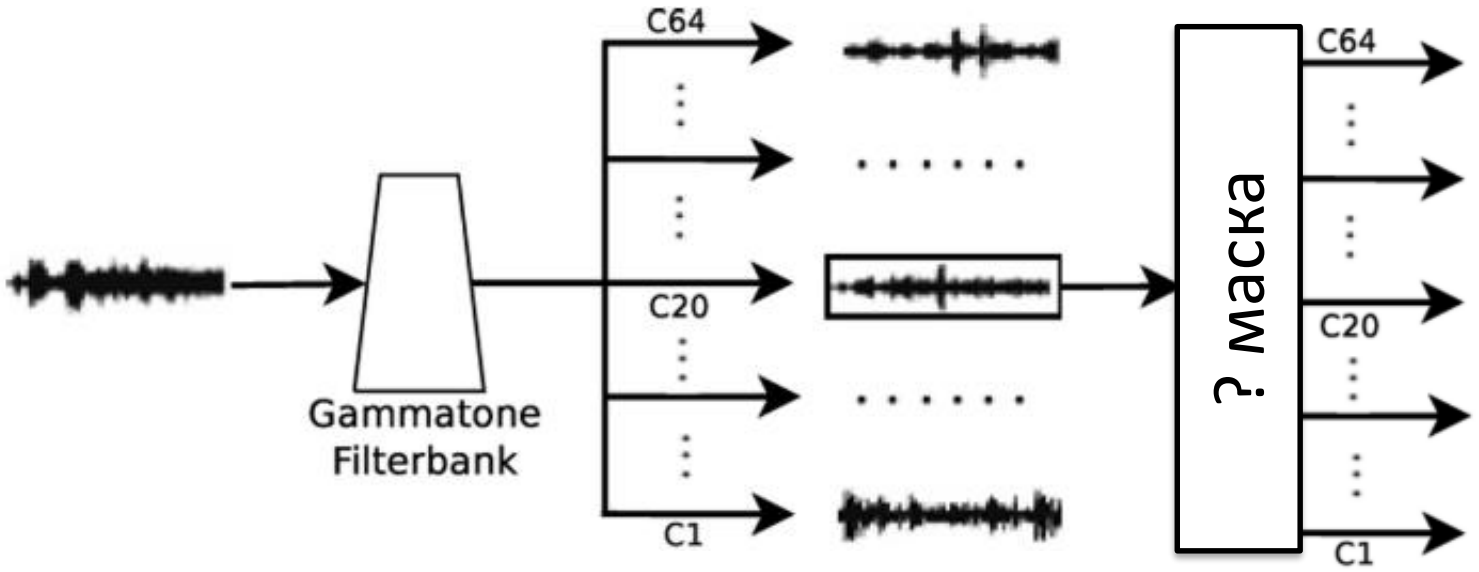
$$G(f) \approx \left[1 + \frac{j(f - f_c)}{b(f_c)} \right]^{-N} \qquad (0 < f < \infty)$$

$$acf(n, c, \tau) = \sum_{k=0}^{K-1} a(n-k, c) a(n-k-\tau, c) h(k)$$

$$sacf(n, \tau) = \sum_{c} acf(n, c, \tau)$$

$$acf(\mathbf{x}_{n,c}) = \mathrm{IDFT}(|\mathrm{DFT}(\mathbf{x}_{n,c})|^p)$$

Частотно-временная маска



Для разделения источника звука системы CASA **маскируют кохлеаграмму**. Эта маска (фильтр Винера), взвешивает целевые области источника и подавляет остальные.

Физиологическая мотивация за маской **проистекает из слухового восприятия**, когда звук становится неслышным из-за более громкого звука.

Восстановление сигнала

- Чтение маскированных каналов кохлеограммы.
- Инвертирование частотно-временного представления сигнала.
- **Преобразование Фурье**.
- Другие методы.

Преобразования Фурье

- **Прямое преобразование Фурье**
  - позволяет по заданной функции $f(t)$ находить соответствующую ей спектральную характеристику $F(j\omega)$:

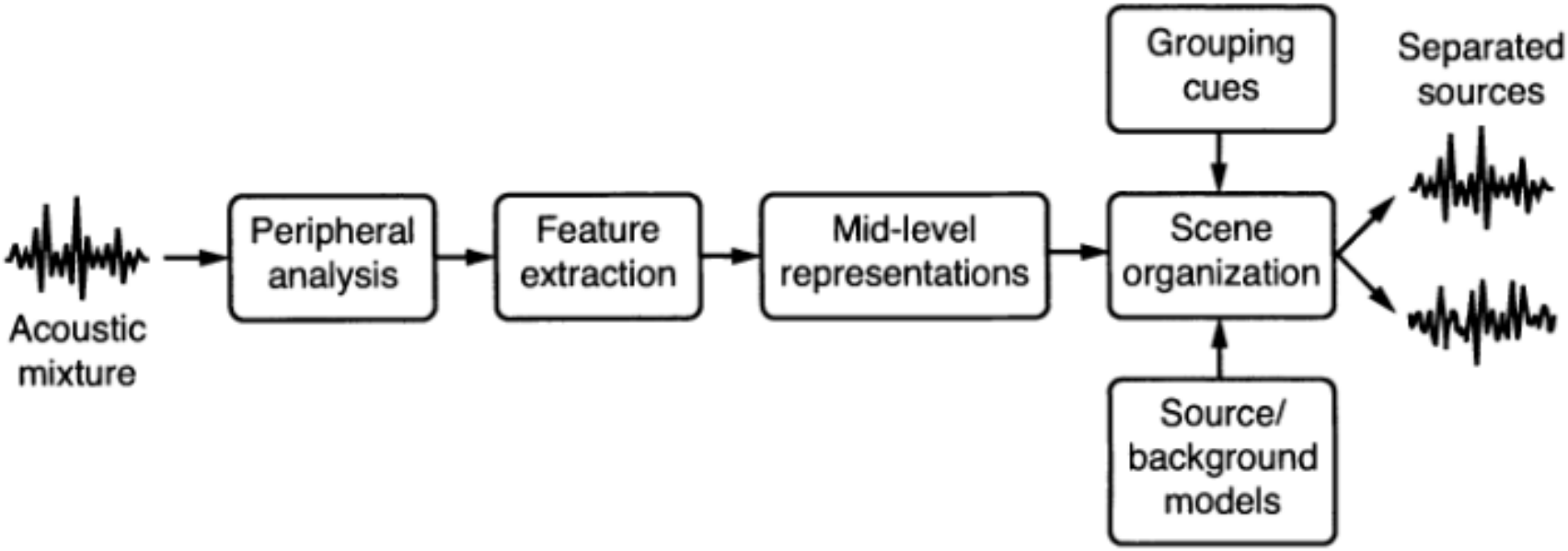
$$F(j\omega) = \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt$$

- **Обратное преобразование Фурье**
  - позволяет по известной функции $F(j\omega)$ определить соответствующую ей функцию $f(t)$:

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(j\omega) e^{j\omega t} d\omega$$

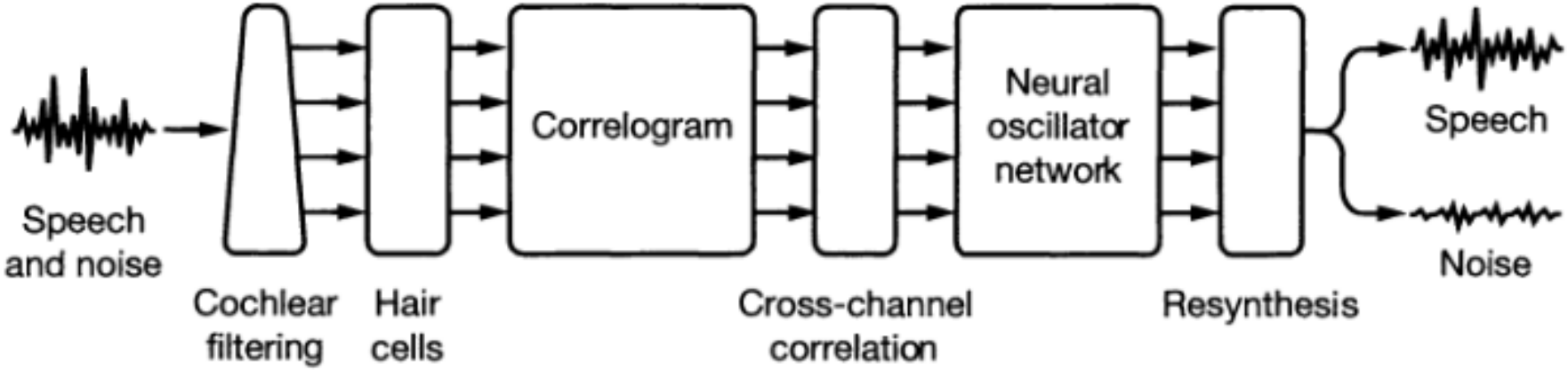
Здесь t – время, ω – частота.

Архитектура системы CASA



Акустическая смесь сначала подвергается периферийному анализу, давая **частотно-временное представление** слуховой активности (кохлеаграмма). Затем извлекаются акустические характеристики (**периодичность, начало, смещение, амплитудная модуляция и частотная модуляция)**. Представления среднего уровня подвергает анализу извлеченные характеристики. Организация сцены происходит на основе **группировки сигналов** и **обученных моделей**. Наконец, форма звукового сигнала повторно синтезируется.

Нейро-осцилляторная модель CASA



Модель слуховой группировки на основе слухового внимания, в которой нейронные осцилляторы, представляющие один воспринимаемый поток, синхронизируются и десинхронизируются для отбора звуковых потоков.

Модель предлагает механизм выделения **высоких** или **низких тонов** в различных последовательностях с чередующимися частотами.

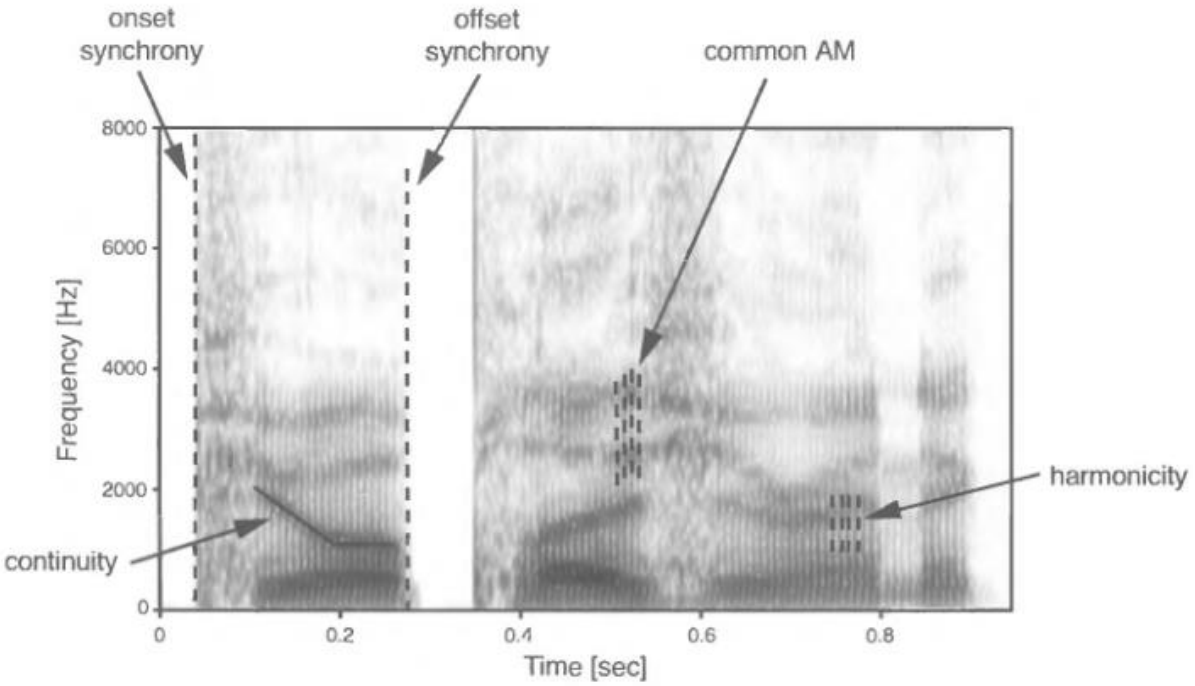
**Wrigley S.N., Brown G.J.* A neural oscillator model of auditory selective attention // Adv. Neural Inf. Process. Syst. 14. 2001. Т. 2130. № Figure 1. С. 1163–1170

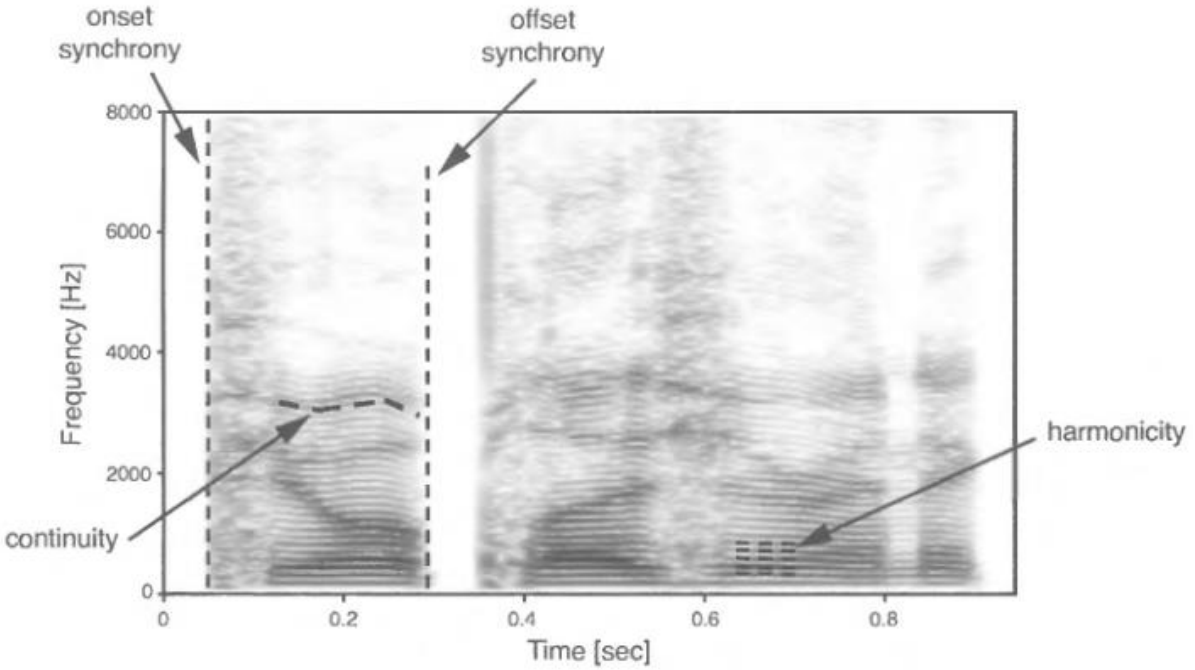
Принципы группировки сигналов

Звуковые сегменты группируются в один поток, если наблюдаются следующие признаки:

- близость по частоте и времени;
- периодичность;
- непрерывный или плавный переход;
- начало и смещение;
- амплитудно-частотная модуляция;
- ритм;
- общее пространственное расположение.

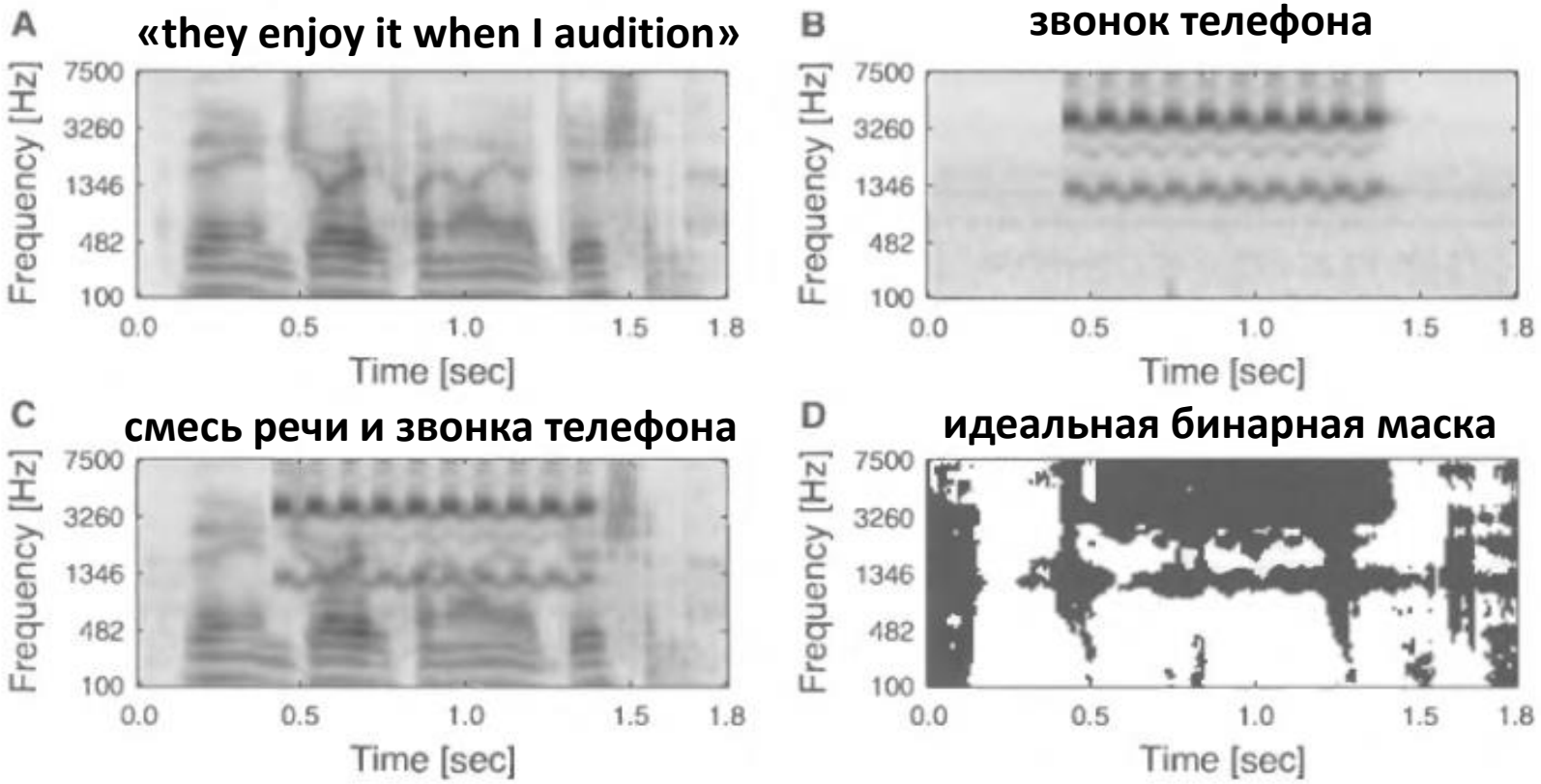
*Wang, D. L. and Brown, G. J. (Eds.) (2006). *Computational auditory scene analysis: Principles, algorithms and applications*. IEEE Press/Wiley-Interscience



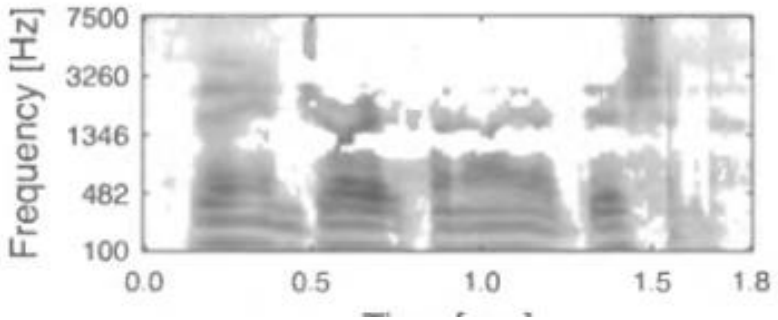


- близость по частоте и времени;
- периодичность;
- непрерывный или плавный переход;
- начало и смещение;
- амплитудно-частотная модуляция;
- ритм;
- общее пространственное расположение.

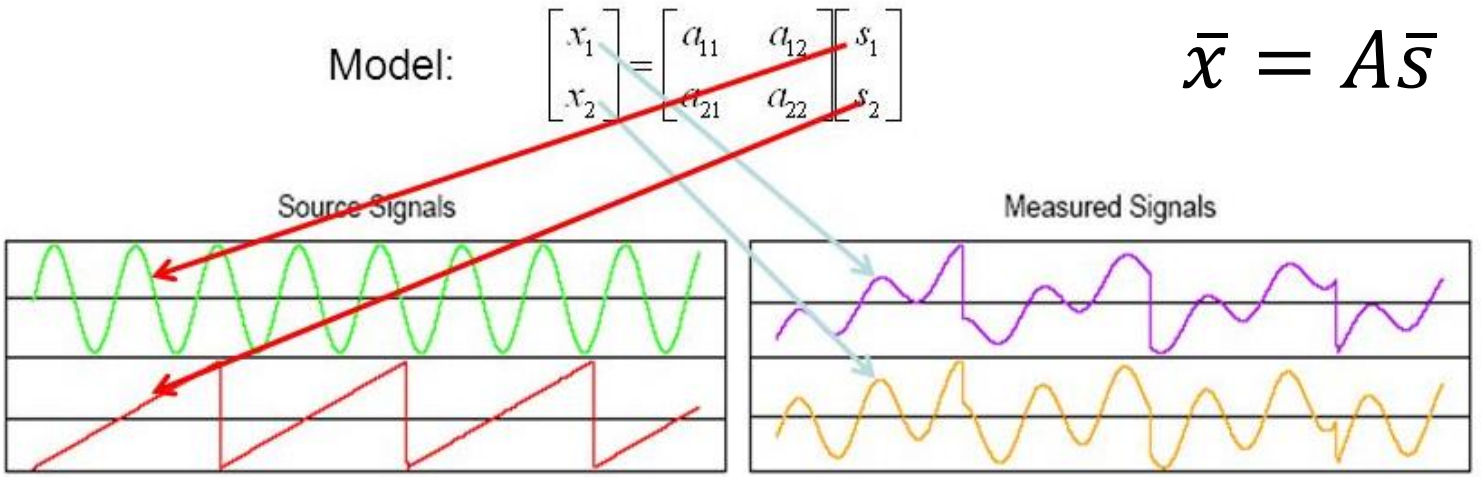
Бинарные частотно-временные маски



Результат применения IBM



Слепое разделение источников



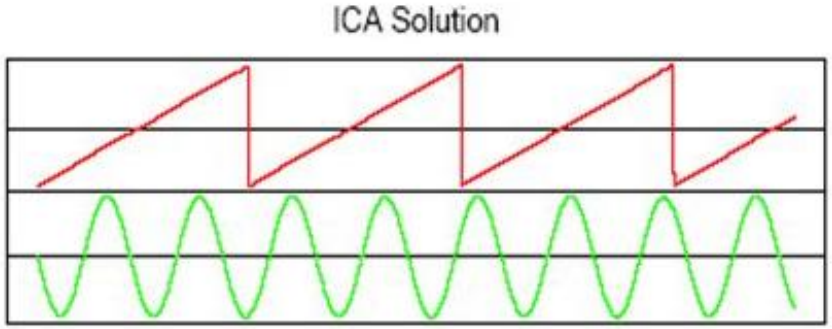
$$\bar{x} = A\bar{s}$$

$\bar{x}$ – смеси

$\bar{s}$ – источники

A - параметры

Проблема: найти $\bar{s} = W\bar{x}$, где $W=A^{-1}$



Метод слепого разделения сигналов

- BSS - это разделение набора исходных сигналов из множества смешанных сигналов, без помощи информации (или с очень небольшим количеством информации) об источнике сигналов или в процессе смешивания.

- Методы BSS - это, прежде всего, системы выделения и распознавания речи, системы телекоммуникаций, системах обработки данных в медицинских приборах.

- Методы BSS можно разделить на два больших класса:
  - методы, работающие с данными **во временной области**
    - **Ограничения**: число датчиков (микрофонов) не меньше числа источников сигнала, не должно быть постороннего шума, кроме самих сигналов, например, шум микрофонов, работа с данными очень большой размерности, большое время выполнения этих алгоритмов, да и результаты, полученные при помощи этих методов, оставляли желать лучшего
  - методы, работающие с данными **в частотной области**
    - **Ограничения**: алгоритм DUET основан на предположении, что все независимые источники имеют в любой момент времени редкий, разбросанный частотный спектр (каждая частотная компонента смешанного сигнала связана только с одним независимым источником).

- Один из методов основан на частотно-зависимой составляющей реальной функции когерентности наблюдаемых сигналов. Этот параметр позволяет обнаруживать частотно-временные зоны, в которых активен только один источник.

Методы реализации CASA

- Моноуральные
- Бинауральные
- Нейронные модели CASA
- Анализ музыкальных звуковых сигналов
- Нейронное перцептивное моделирование

Монофонические алгоритмы частотно-временного маскирования

Монауральные алгоритмы **используют собственные свойства звука** для разделения или анализа слуховой сцены.

Это такие свойства как:

гармоничность (harmonicity), начало и смещение (onset and offset), амплитудные (amplitude) и частотные модуляции (frequency modulations), временную непрерывность (temporal continuity) и обученные речевые модели (trained speech models).

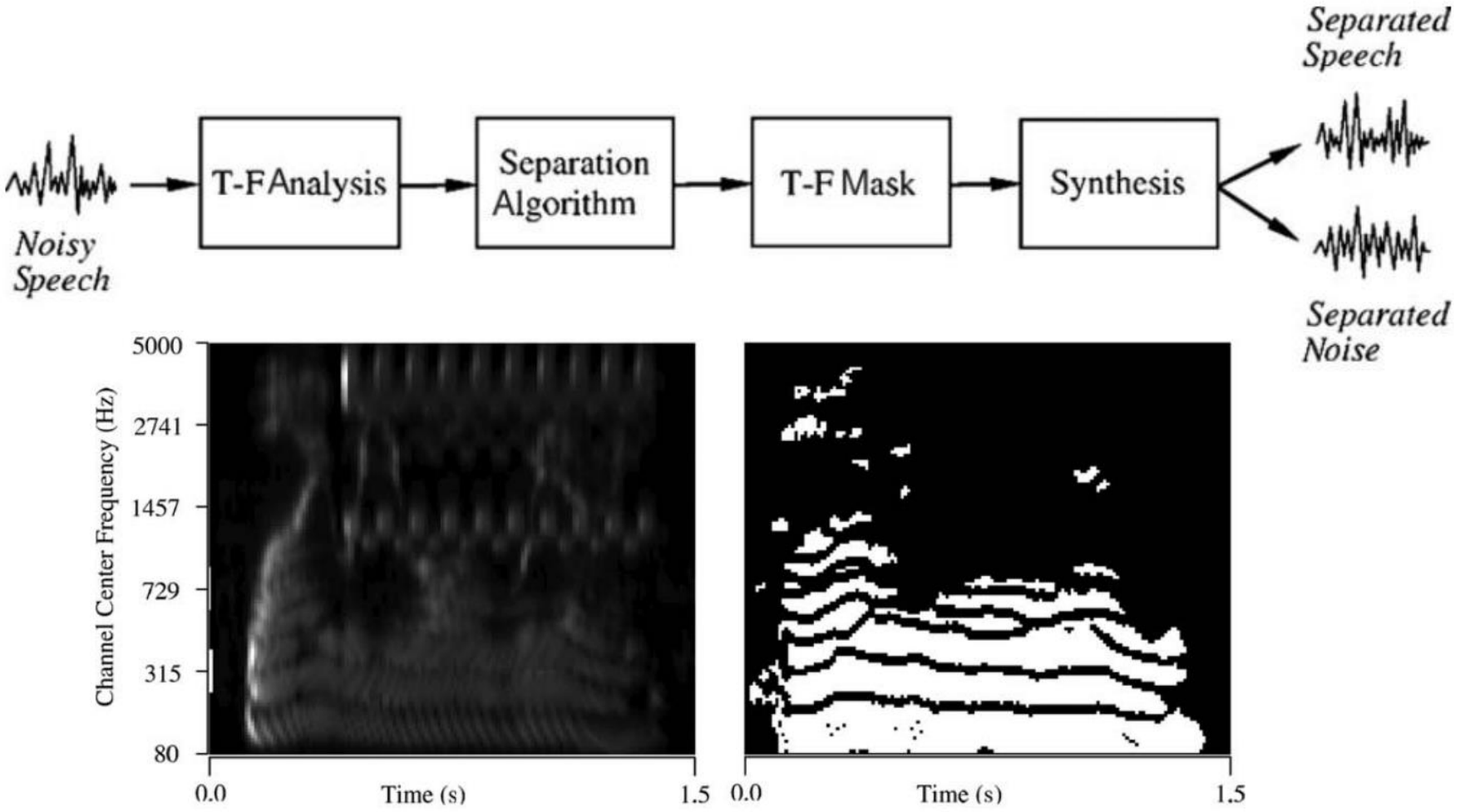
Алгоритмы сложны и непроизводительны.

Браун и Кук 1994

- Набор гамматон-фильтров.
- Выделяют слуховые карты.
- Определение автокорреляции.
- Сегментация.
- Группировка на основе сходства контуров основного тона.
- Сгруппированные потоки представляет из себя бинарную маску.

*Wang, D. L. and Brown, G. J. (Eds.) *Computational auditory scene analysis: Principles, algorithms and applications*. IEEE Press/Wiley-Interscience, 2006

Frequency Masking for Speech Separation



* *DeLiang Wang* Time–Frequency Masking for Speech Separation and Its Potential for Hearing Aid Design // Trends in Amplification, 2008

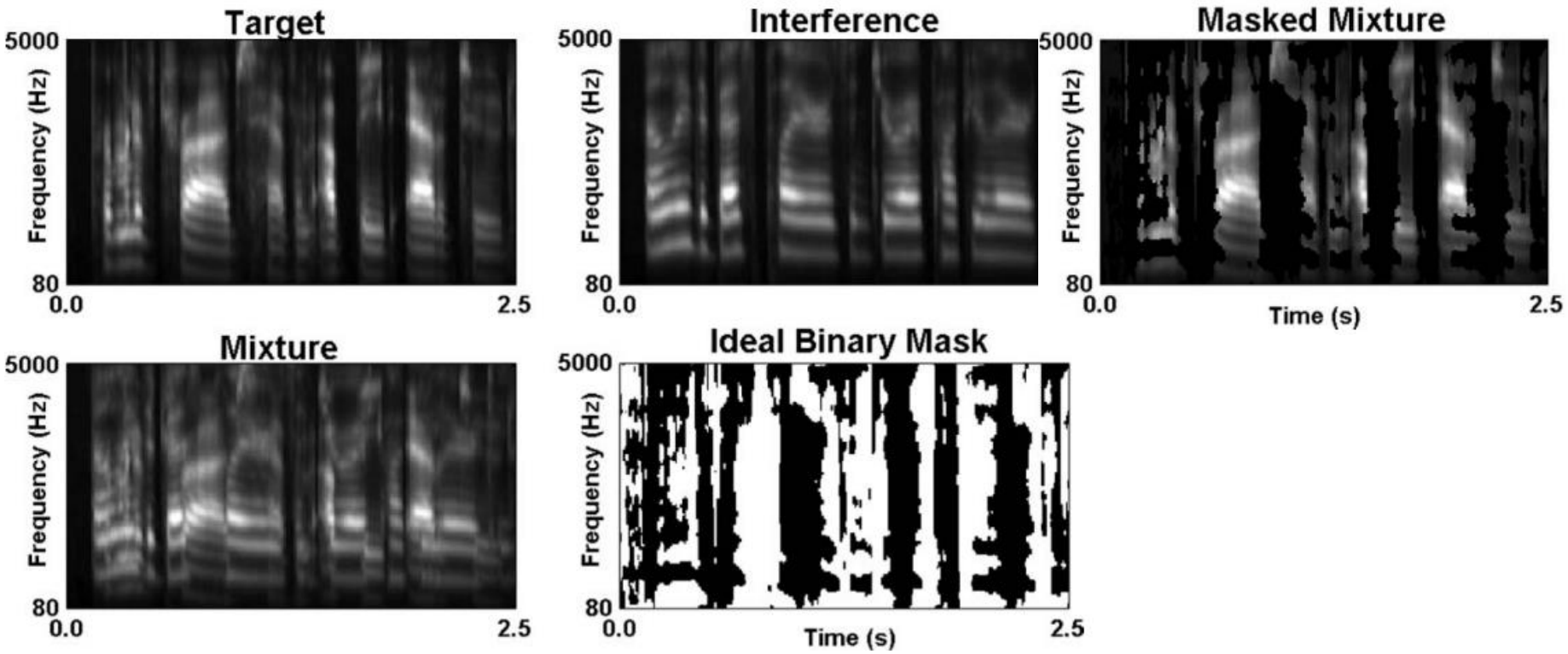
Идеальная бинарная маска F-T

Когда разные источники ортогональны, двоичной маски **достаточно** для полного извлечения одного источника из смеси.

$$IBM(t, f) = \begin{cases} 1 & \text{if } s(t, f) - n(t, f) > LC \\ 0 & \text{otherwise} \end{cases}$$

- $s(t, f)$ - целевая энергия, дБ
- $n(t, f)$ - энергия помехи, дБ
- LC - пороговое значение (отношения сигнал/шум [SNR]), дБ

Идеальная бинарная маска

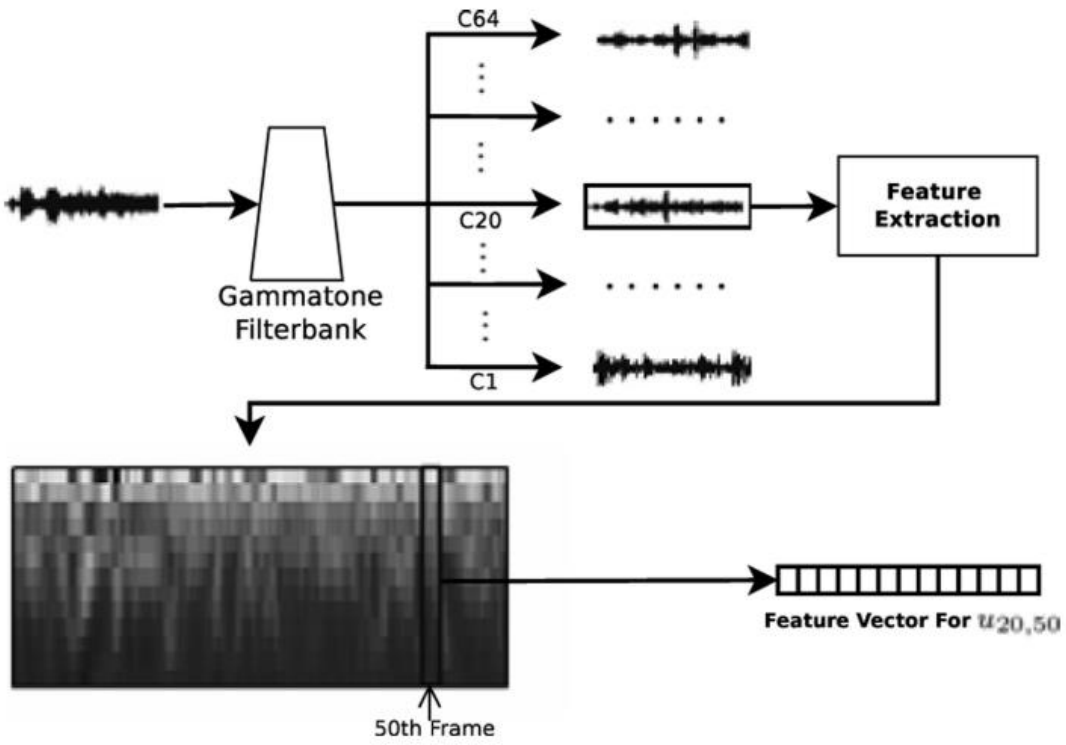


Вверху слева: кохлеаграмма целевого высказывания («Primitive tribes have an upbeat attitude») и кохлеаграмма с высказыванием («Only the best players enjoy popularity»).

Ниже: кохлеаграмма смеси, идеальная бинарная маска.

Справа: замаскированная смесь с использованием идеальной бинарной маски.

Найти идеальную бинарную маску



16КГц подается на 64-каналаьный набор гамматон-фильтров с равномерно распределенными частотами 50Гц-8КГц.

Входные потоки делятся на кадры по 20мс с перекрытием 10 мл (кохлеограмма).

Акустические **характеристики** выводятся на уровне канала на **основе тона** и **амплитуды** (много различных алгоритмов).

Для двоичной классификации используется **метод опорных векторов**. После чего выполняем межканальную корреляцию из анализа начала и смещения сегментов. На выходе получаем идеальную бинарную маску.

**Wang Y., Han K., Wang D.* Exploring Monaural Features for Classification-Based Speech Segregation // IEEE Trans. Audio. Speech. Lang. Processing. 2013. Т. 21. № 2. С. 270–279

Feature Extraction

- **Pitch-based features**

  Guoning Hu, DeLiang Wang* A **Tandem Algorithm for Pitch Estimation and Voiced Speech Segregation // IEEE Trans. Audio. Speech. Lang. Processing. 2010. T. 18. № 8. C. 2067–2079.

- **AMS features**

  * *Kim, G., Lu, Y., Hu, Y., and Loizou, P. C.* An algorithm that improves speech intelligibility in noise for normal-hearing listeners," J. Acoust. Soc. 2009. Am. 126, 1486–1494.

- **SVM CLASSIFICATION**

  **Chang, C. C., and Lin, C. J.* LIBSVM: A library for support vector machines, 2001. http://www.csie.ntu.edu.tw/?cjlin/libsvm (Last viewed November 2010)

- **AUDITORY SEGMENTATION**

  **Wang, D. L., and Brown, G. J.* "Fundamentals of computational auditory scene analysis," in Computational Auditory Scene Analysis: Principles, Algorithms and Applications. 2006. Chap. 1, pp. 1–37.

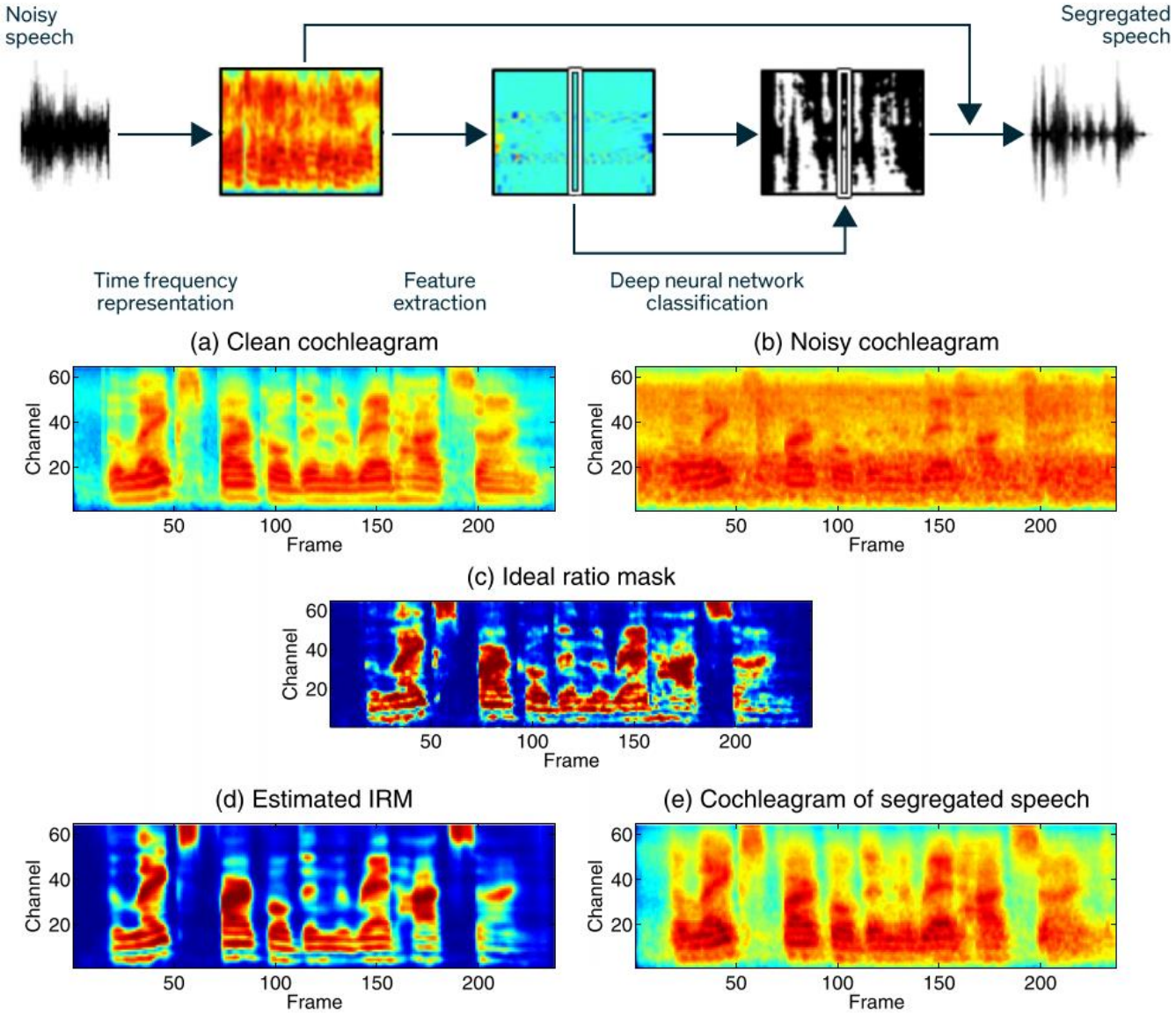
Алгоритм оценки основного тона

Используется для поиска бинарной частотно-временной маски на основе выделения основного тона:

- разложение входного сигнала с использованием 128 гамматон-фильтров;
- в каждом канале определяется коррелограмма, автокорреляционная функция сигнала;
- расчет межканальной корреляции по сходству откликов соседних фильтров;
- построение огибающей основного тона и отбрасывание неразрешенных гармоник;
- начальная оценка основного тона;
- итеративное улучшение оценки основного тона;
- сегментация;
- сегрегация.

Guoning Hu, DeLiang Wang* A **Tandem Algorithm for Pitch Estimation and Voiced Speech Segregation // IEEE Trans. Audio. Speech. Lang. Processing. 2010. Т. 18. № 8. С. 2067–2079.

DNN для работы с неизвестными шумами



- Выделение шума происходит на основе свойств шума (амплитуда, гармоничность и пр.).
- Фильтр находит такие места и вырезает их бинарной маской.
- DNN обучена на 10 000 шумах для оценки идеальной маски.
- Выделено 85 функций (обобщения звуковых событий).
- Проверена на слушателях с нарушением слуха.

Delay Wang 2017

https://spectrum.ieee.org/consumer-electronics/audiovideo/deep-learning-reinvents-the-hearing-aid

DeLay Wang

DeLay Wang - профессор кафедры компьютерных наук и инженерии и Центра когнитивных наук и наук по изучению мозга в Университете штата Огайо в Колумбусе, штат Огайо.

Руководит Лабораторией восприятия и нейродинамики OSU, которая занимается разработкой алгоритмов для решения проблем, связанных с машинным восприятием.

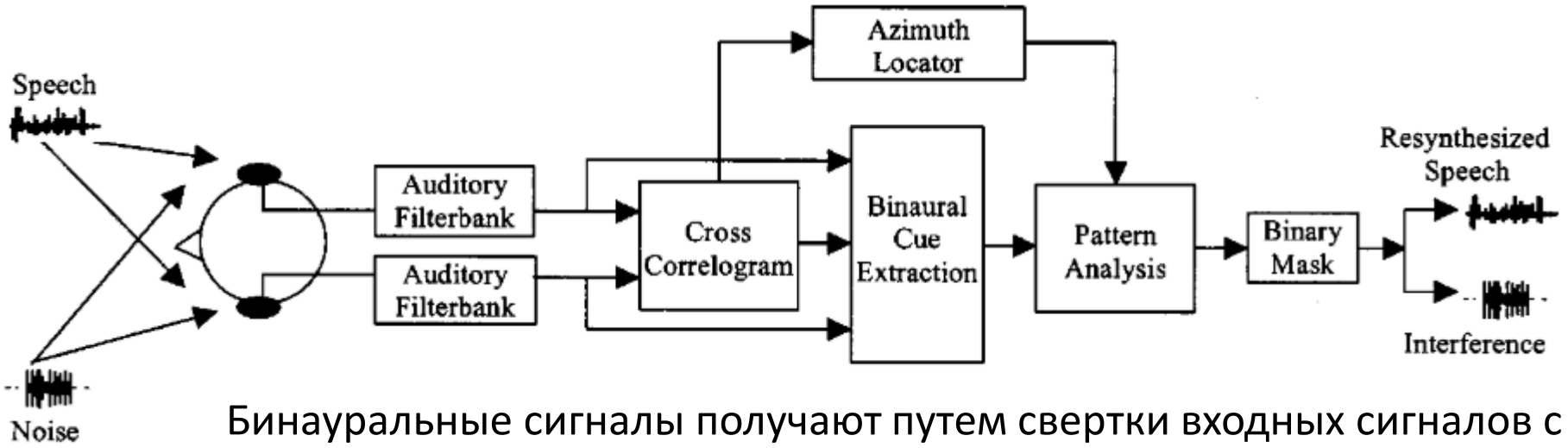
Wang получил докторскую степень в области компьютерных наук в Университете Южной Калифорнии в Лос-Анджелесе в 1991 году после получения степени бакалавра и магистра в Пекинском университете в Пекине.

Является соредактором главного журнала Neural Networks.

Binaural and Array-Based Time–Frequency Masking Algorithms

- Подход использует кластеризацию признаков в пространстве и получения IBM.
- Для левого и правого уха извлекаются **интерауральные разницы во времени** (interaural time difference , ITD) и **интенсивности** (interaural intensity difference, IID).
- Систематические изменения бинауральных сигналов приводят к **типичной кластеризации в пространстве** функций ITD-IID.

Speech segregation based on localization



Бинауральные сигналы получают путем свертки входных сигналов с измеренными импульсными характеристиками, связанными с голой манекена KEMAR, модель слуховой периферии.

Азимутальная локализация для всех источников основана на механизме взаимной корреляции **ITD** и **IID**, рассчитываются независимо для разных частотных каналов. Блок анализа структуры производит оценку идеальной двоичной маски, которая позволяет восстановить целевой сигнал и мешающий звук.

*Roman N., Wang D., Brown G.J. Speech segregation based on sound localization // J. Acoust. Soc. Am. 2003. T. 114. № 4. C. 2236–2252.

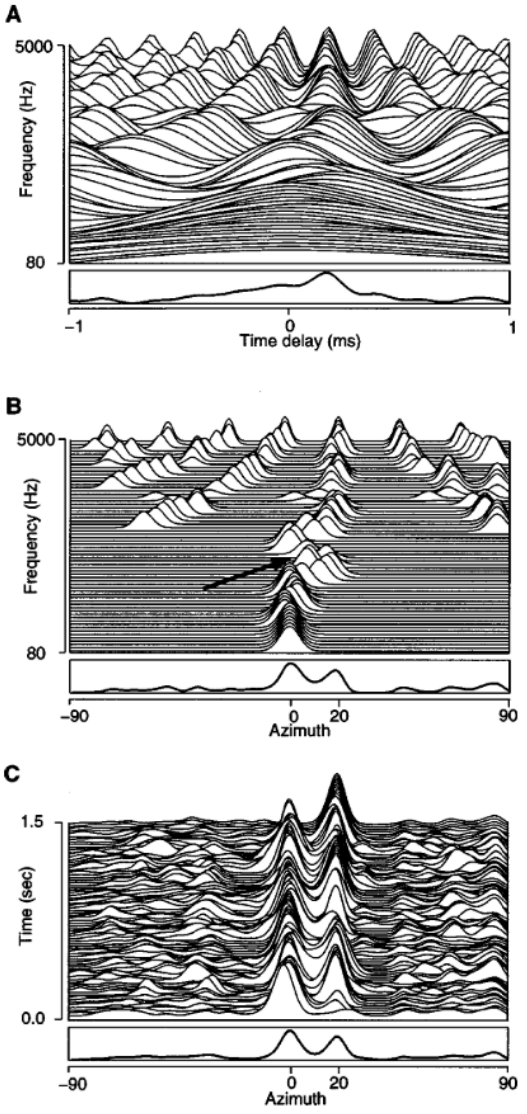
Азимут локализация

$$C(i,j,\tau) = \frac{\Sigma_{k=0}^{K-1}(l_i(j-k)-\bar{l}_i)(r_i(j-k-\tau)-\bar{r}_i)}{\sqrt{\Sigma_{k=0}^{K-1}(l_i(j-k)-\bar{l}_i)^2}\sqrt{\Sigma_{k=0}^{K-1}(r_i(j-k-\tau)-\bar{r}_i)^2}},$$

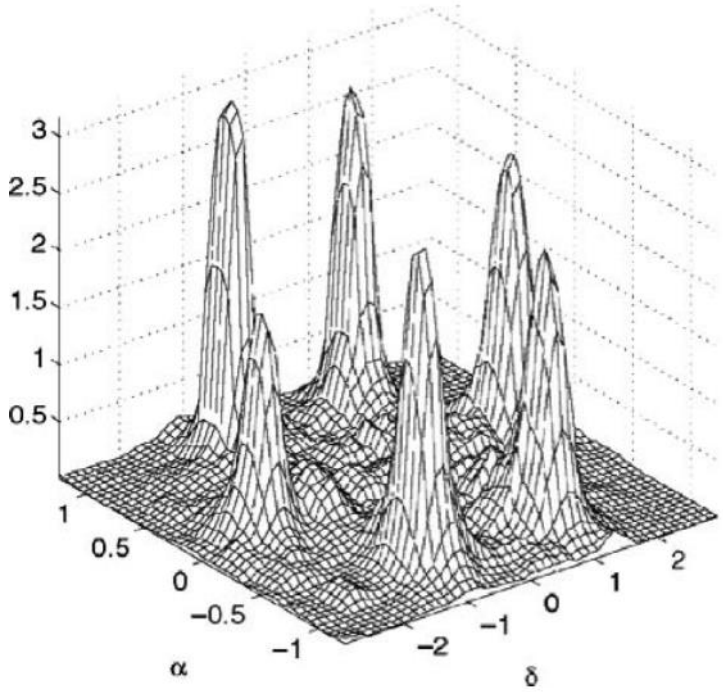
где l_i, r_i – выход левой и правой слуховой периферии *i*-ого канала, $\bar{l}_i, \bar{r}_i$ - средние их значения, τ – отставание, j - кадр.

Взаимная корреляция вычисляется для всех частотных каналов и обновляется каждые 10 мс.

На рис. мужской голос с азимутом 0, женский с азимутом 20.



Алгоритм DUET



Гистограмма 6 смесей,
где α - разность амплитуд,
а δ - разность во времени

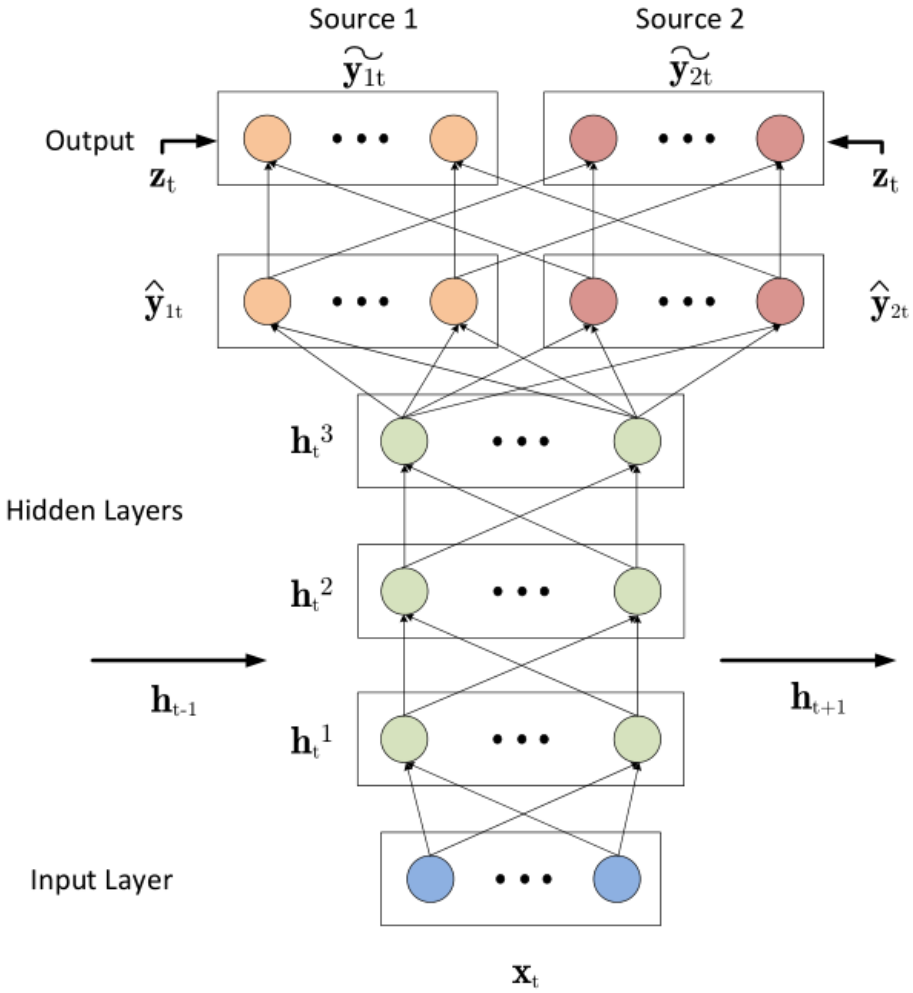
Алгоритм неконтролируемой кластеризации для извлечения отдельных источников со входом от **двух всенаправленных микрофонов**. Процесс обучения сводится к генерации двумерной гистограммы, затем гистограмма сглаживается, и располагаются пики, которые соответствуют отдельным источникам. Каждый пик используется для построения двоичной маски T-F, которая используется для восстановления отдельного источника звука из смеси.

* *Yilmaz O., Rickard S.* **Blind separation of speech** mixtures via time-frequency masking. IEEE Transactions on Signal Processing, 2004

ICA и Time–Frequency Masking

- Концепция маскирования T-F позволяет справляться с разделением **недооцененных** источников, что создает серьезные трудности для ICA.
- Стандартная формулировка ICA требует, чтобы **количество микрофонов было не меньше**, чем количество источников, что часто является непрактичным ограничением.
- С другой стороны, **впечатляющее разделение** может быть получено при выполнении предположений ICA.
- **Существует идея объединить** ICA и F-T, например: выделить один источник с помощью F-T, затем применить ICA.

DRNN архитектура для разделения



$$\tilde{\mathbf{y}}_{\mathbf{1}_t} = \frac{|\hat{\mathbf{y}}_{\mathbf{1}_t}|}{|\hat{\mathbf{y}}_{\mathbf{1}_t}| + |\hat{\mathbf{y}}_{\mathbf{2}_t}|} \odot \mathbf{z}_t$$

$$\tilde{\mathbf{y}}_{\mathbf{2}_t} = \frac{|\hat{\mathbf{y}}_{\mathbf{2}_t}|}{|\hat{\mathbf{y}}_{\mathbf{1}_t}| + |\hat{\mathbf{y}}_{\mathbf{2}_t}|} \odot \mathbf{z}_t,$$

x_t – спектры смеси,

y_{1t} и y_{2t} – спектры источников,

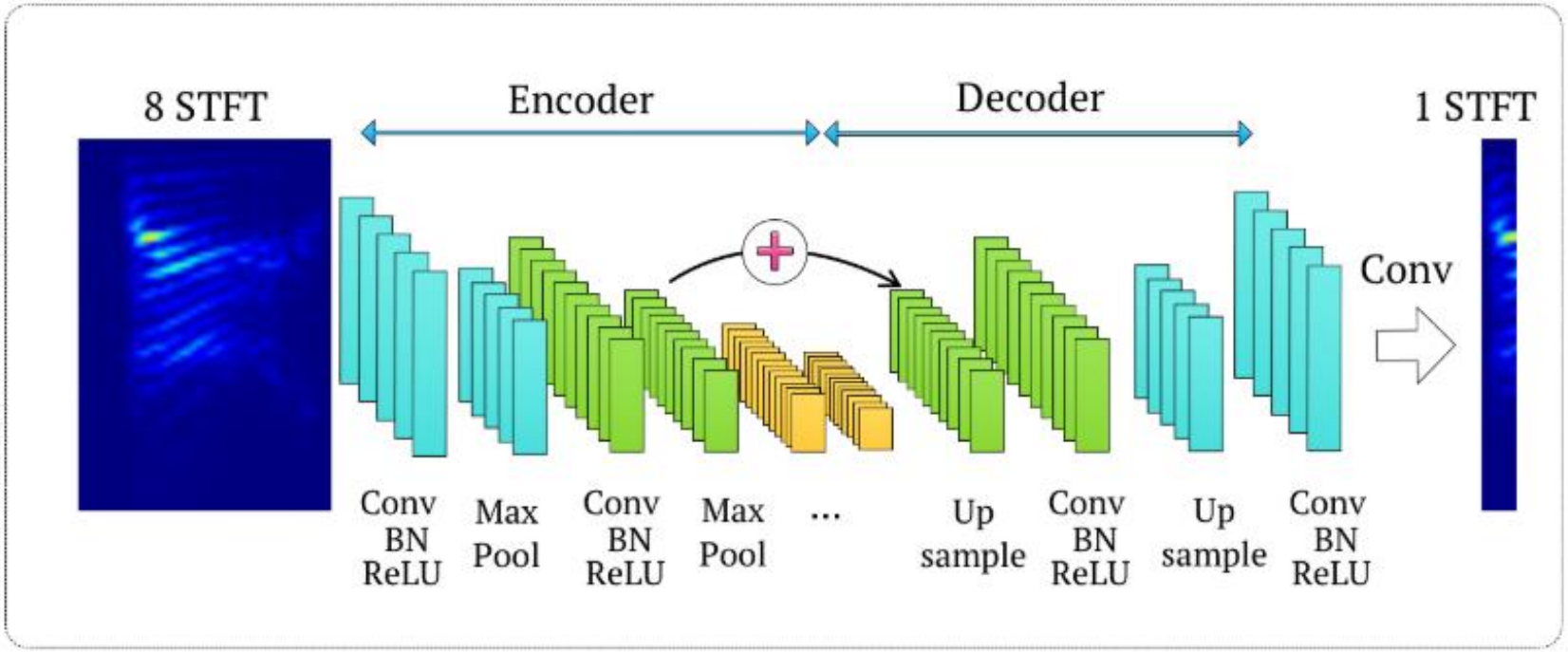
Θ - поэлементное умножение.

Сигналы восстанавливаются на основе обратного кратковременного преобразования Фурье (ISTFT).

* *Po-Sen Huang, Minje Kim M.H.-J., Smaragdis P.* Singing-Voice Separation From Monaural Recordings Using Deep Recurrent Neural Networks // 15th International Society for Music Information Retrieval Conference. : IEEE, 2014. С. 1562–1566.

CNN для подавления шума

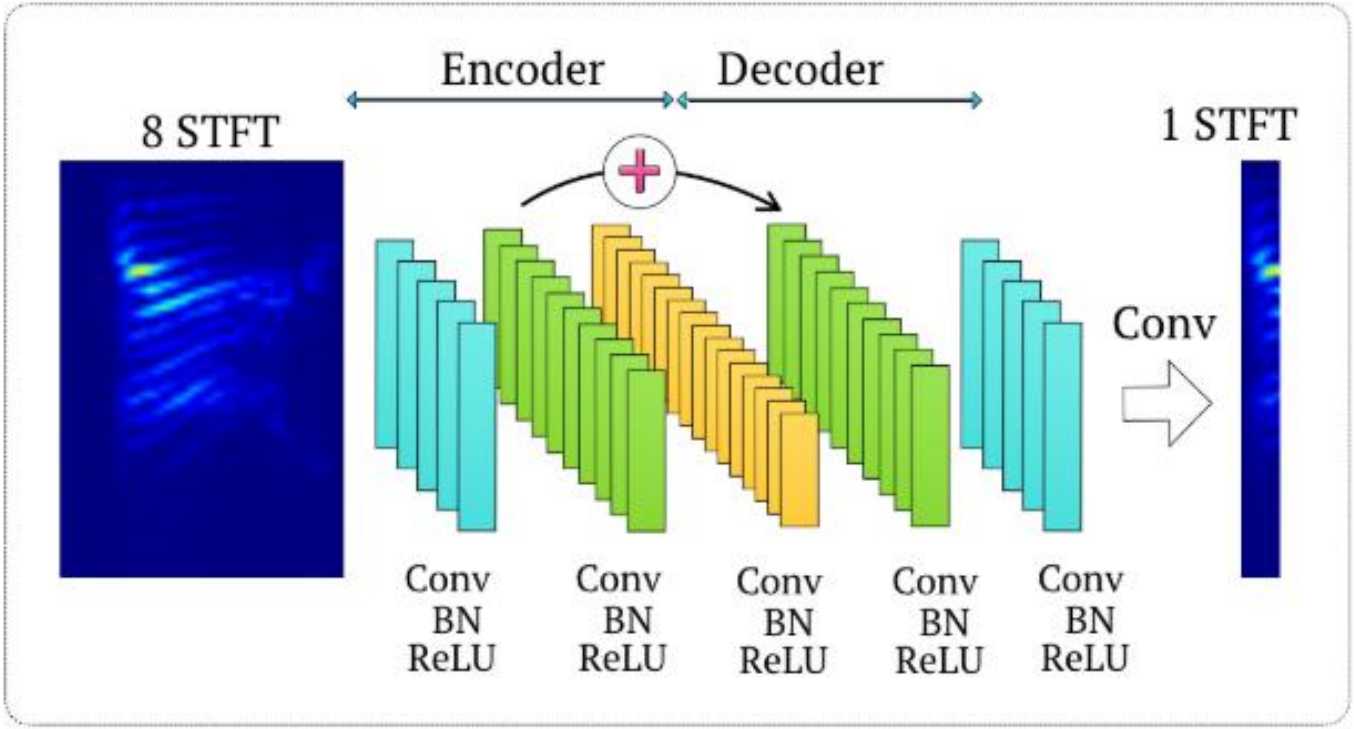
Convolutional Encoder-Decoder Network (CED)



Для подавления бормотания (babble noise)

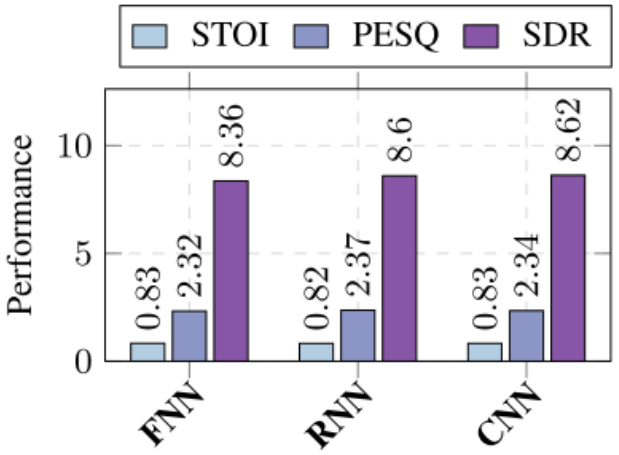
**Park S.R., Lee J.W.* A fully convolutional neural network for speech enhancement // Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH. 2017. T. 2017–Augus. № 2. C. 1993–1997

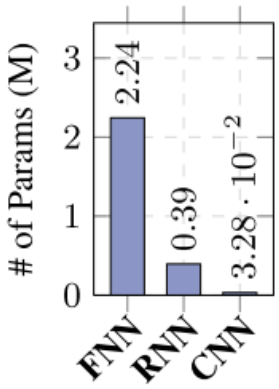
Redundant CED (R-CED)

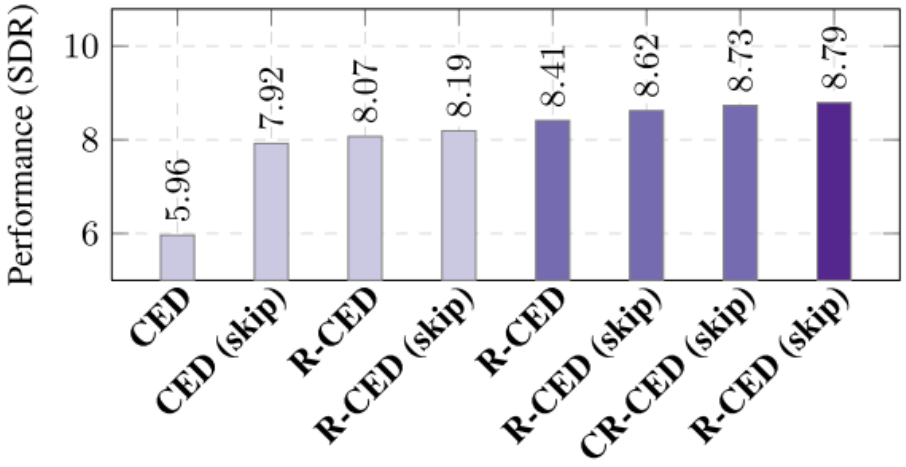


**Park S.R., Lee J.W.* A fully convolutional neural network for speech enhancement // Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH. 2017. T. 2017–Augus. № 2. C. 1993–1997

Производительность и шумоподавление





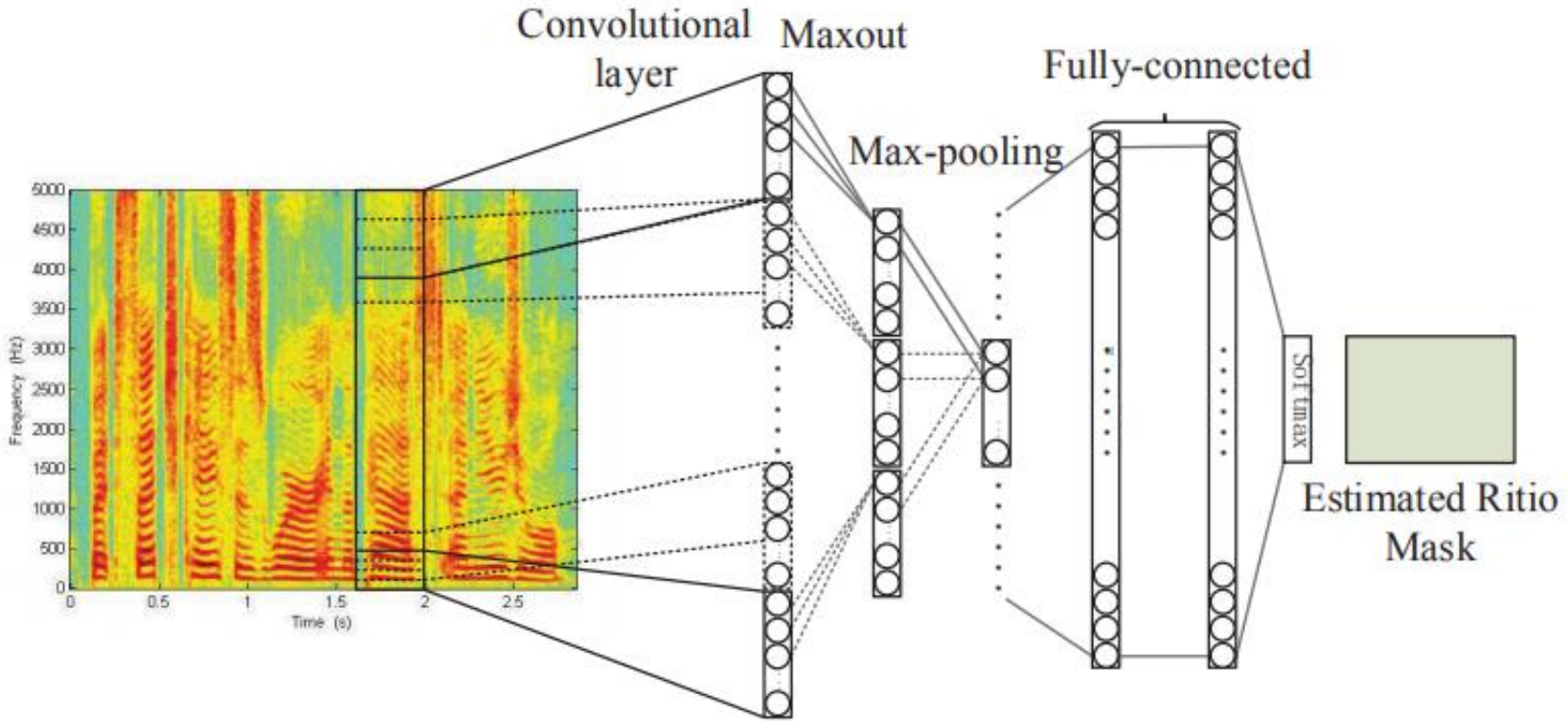


Слева: шумоподавляющие характеристики FNN, RNN и CNN и соответствующий размер сети. Размер модели CNN примерно в 68 раз меньше, чем у FNN, и примерно в 12 раз меньше, чем у RNN.

Справа: эффективность шумоподавления CED и R-CED.

Signal-to-Noise Ratio (**SNR**) – отношение сигнал шум, Perceptual Evaluation of Speech Quality (**PESQ**) - воспринимаемая оценка качества речи, Signal-to-Distortion Ratio (**SDR**) – соотношение сигнал искажение, Short-Time Objective Intelligibility (**STOI**) - кратковременная объективная разборчивость

CNN for speech separation



Принципиальная схема для оценки IRM (ideal ratio mask, мягкая маска) системы CMNN. Обучение проводится для каждого канала фильтра Gammatone.

*Like Hui, Meng Cai, Cong Guo, Liang He, Wei-Qiang Zhang, and Jia Liu Convolutional maxout neural networks for speech separation // 2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). : IEEE, 2015. C. 24–27.

Результаты разделения

TABLE I
RESULTS COMPARISON ON -5 DB TIMIT MIXTURES.

System	Babble		Factory		Pink		White	
	sSNR	PESQ	sSNR	PESQ	sSNR	PESQ	sSNR	PESQ
Noisy	-7.03	1.41	-6.90	1.30	-7.09	1.25	-7.18	1.26
DNN	-3.80	1.40	-3.43	1.39	-3.07	1.46	-4.42	1.45
CNN	-3.71	**1.43**	-3.28	**1.41**	-2.83	1.49	-3.97	1.49
CMNN	**-3.67**	1.42	**-3.24**	**1.41**	**-2.66**	**1.51**	**-3.63**	**1.51**

TABLE II
RESULTS COMPARISON ON 0 DB TIMIT MIXTURES.

System	Babble		Factory		Pink		White	
	sSNR	PESQ	sSNR	PESQ	sSNR	PESQ	sSNR	PESQ
Noisy	-4.55	1.70	-4.38	1.64	-4.59	1.59	-4.69	1.57
DNN	-1.39	1.76	-1.21	1.78	-0.79	1.84	-1.99	1.77
CNN	-1.36	**1.78**	-1.12	**1.81**	-0.63	1.87	-1.61	1.80
CMNN	**-1.27**	**1.78**	**-1.06**	1.80	**-0.46**	**1.88**	**-1.26**	**1.81**

TABLE III
RESULTS COMPARISON ON 5 DB TIMIT MIXTURES.

System	Babble		Factory		Pink		White	
	sSNR	PESQ	sSNR	PESQ	sSNR	PESQ	sSNR	PESQ
Noisy	-1.36	2.06	-1.20	2.01	-1.44	1.97	-1.56	1.92
DNN	0.71	2.09	0.80	2.11	1.13	2.19	0.23	2.08
CNN	0.70	**2.12**	0.81	**2.16**	1.21	**2.23**	0.50	2.11
CMNN	**0.82**	**2.12**	**0.88**	2.15	**1.37**	**2.23**	**0.82**	**2.12**

Результаты CMNN являются лучшими среди всех систем. Это происходит из-за сверточного слоя и слоя maxout в CMNN, которые отсутствуют в DNN.

Мы также видим, что результаты CNN лучше, чем DNN. Это указывает на то, что сети maxout могут оптимизировать функцию активации для каждого блока сетей.

Выводы

- Вдохновленный человеческой слуховой обработкой, вычислительный анализ слуховой сцены (CASA) **способен** справиться **с общими видами шумов**.
- Существует возможность заменить сложные методы различных этапов CASA просто нейронными сетями CNN, LSTM, … .
- Необходимо использовать кохлеограммы вместо спектрограмм.