# Применение синусоидального моделирования речи к задаче диаризации звука

Б. М. Нутфуллин, Е. А. Ильюшин.

***Аннотация*—Речь является видовой особенностью человека и его преимуществом над другими видами в рамках эволюции. Диаризация звука – это процесс разделения звука с учетом принадлежности к диктору. До появления глубокого обучения и наличия необходимых вычислительных ресурсов качество алгоритмов, определяющих диктора по голосу, оставляло желать лучших результатов. Диаризация имеет многочисленные приложения: умные колонки, мобильные телефоны, системы автоматического перевода речи. Но следует отметить, что существующие алгоритмы диаризации имеют недостатки, например, сложность работы при одновременном произнесении речи несколькими дикторами или результатов диаризации недостаточно для ее автоматического применения в некоторых областях. Этим обусловлена актуальность исследований в данной области.**

**Синусоидальная модель преставляет собой алгоритм отслеживания последовательностей точек в пространстве время-амплитуда-частота. В существующих исследованиях она применяется к моделированию эхолокации, человеческой речи, а также синтезу речи. На момент исследования в литературе не было найдено применений синусоидальной модели в задаче диаризации.**

**В работе рассмотрена задача диаризации и основные показатели качества, используемые при оценке решений данной задачи. Рассмотрены основные промежуточные представления звука, использующиеся в существующих решениях, и предложен алгоритм диаризации, использующий синусоидальное моделирование речи. Преимуществом предложенного алгоритма является возможность работы синусоидальных представлений как детектора наличия голоса, что в целом позволило сделать более эффективным используемый алгоритм диаризации.**

***Ключевые слова*—диаризация, звук, разделение дикторов**

## I. Введение

Во время разговора, в котором участвует множество лиц, для человека не составляет труда определить, чья речь звучит в данный момент. Обычно мы не подозреваем, что для существующих вычислительных систем эта задача более сложна и вариативна в решении, в то время как наш мозг решает ее играючи, даже не задумываясь об этом, схема использования диаризации показана на рисунке 1.

Статья получена 27.05.2021.

Нутфуллин Булат Маратович, МГУ им. М.В. Ломоносова, факультет вычислительной математики и кибернетики, Москва, Россия, (email: bulat15g@gmail.com).

Ильюшин Евгений Альбинович, МГУ им. М.В. Ломоносова, факультет вычислительной математики и кибернетики, Москва, Россия, (email: eugene.ilyushin@gmail.com).

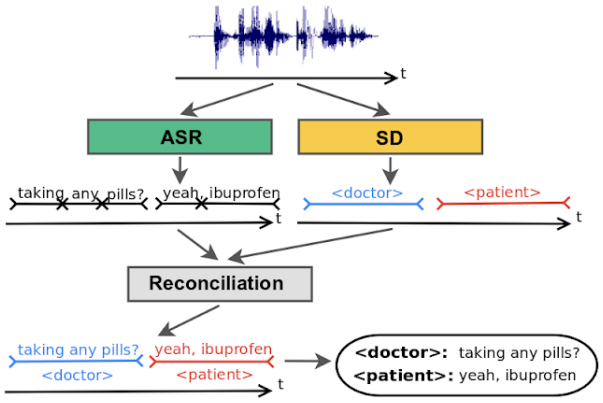


Рис. 1: Схема использования диаризации [1]

Методы определения диктора можно условно разделить по степени автоматизации, а именно:

- наивное распознавание: каждый человек без какой либо предварительной тренировки узнает голоса хорошо знакомых ему людей;
- в криминалистических расследованиях эксперты-криминалисты сравнивают различные аудиозаписи с предзаписанными образцами голоса;
- полностью автоматическое разпознавание диктора – в таком случае анализ речи и принятие решений полностью доверено компьютерным средствам. В работе будет рассматриваться только этот случай.

Автоматическая диаризация акустического сигнала – процесс разделения входящего аудиопотока на однородные сегменты в соответствии с принадлежностью звука тому или иному диктору. При разделении также могут учитываться сегменты, во время которых присутствует речь одновременно нескольких дикторов. В процессе диаризации также решаются следующие подзадачи: производится разделение речевых от неречевых сегментов, определяются моменты смены диктора. На данный момент в современных информационных системах возможно использование в следующих случаях:

- во время трансляций массовых мероприятий (футбольные матчи, концерты и т.д.);
- для помощи системам распознавания речи, что сильно улучшает результат их работы;
- во время переговоров, аудиоконференций;
- для подтверждения личности в биометрических системах;
- в системах умного дома;
- в криминалистике (в данном случае часто процесс доверяется экспертам-криминалистам).

Иллюстрация применения в информационных системах показана на рис 1, где результат диаризации совмещен с задачей распознавания речи.

## II. Задача диаризации

### A. Постановка задачи

Задача диаризации представляет собой отображение меток диктора на сегменты аудиофайла.

$$f : S \to \Delta T,$$

где:

$S$ - множество меток дикторов (или его отсутствия);

$\Delta T$ - множество сегментов $\Delta T_i(T_i, T_{i+1})$, начинающихся в момент времени $T_i$ и заканчивающихся в $T_{i+1}$.

Решением задачи диаризации является построенное отображение сегментов аудиофайла на множество дикторов. Множество дикторов может быть не определено заранее или может быть определено их количество.

### B. Метрика оценки качества диаризации

Для оценивания качества диаризации всей системы часто используют метрику Diarization Error Rate (1), она является основной для сравнения результатов различных систем.

$$DER = \frac{FA + Miss + Overlap + Confusion}{RL}, \quad (1)$$

где:

$RL$ – сумма длин всех речевых сегментов;

$FA$ – сумма длин сегментов, которые были предсказаны, как содержащие речь, но на самом деле являются неречевыми;

$Miss$ – сумма длин сегментов, содержащих речь, но предсказанных, как неречевые сегменты;

$Overlap$ – сумма длин сегментов, на которых была предсказана одновременная речь двух дикторов (перекрытие), но на самом деле сегмент относится к одному диктору. Примечание: не все системы диаризации работают, учитывая при подсчете оценки качества перекрытия несколькими дикторами одного временного промежутка;

$Confusion$ – сумма длин сегментов, на которых предсказанная метка диктора не совпала с меткой действительно говорившего человека.

## III. Обзор существующих решений

### A. VAD

Детектор наличия голоса отделяет сегменты, содержащие речь, от неречевых сегментов. От результатов работы детектора наличия голоса напрямую зависит DER (1). Для разделения сегментов производится извлечение признаков, таких, как Мел-кепстральные коэффициенты и их классификация. Для классификации используются такие модели, как DNN [26], Гауссова смесь распределений [27] и Скрытые Марковские модели [28].

### B. Предобработка аудиосигнала

В данном пункте рассмотрены часто использующиеся алгоритмы для предобработки звука, использующиеся для де-реверберации, шумоподавления, разделения речи, улучшения звука. В работах [17], [18], [19] представлен обзор алгоритмов, применяющихся в задачах разделения речи и шумоподавления. При проведении соревнований DIHARD [40] были представлены работы [20], [21], [22], описывающие решения, включающие предобработку звука. Работы [23], [24], [25] показали улучшение результатов шумоподавления с использованием глубокого обучения.

### C. Классификация алгоритмов диаризации по промежуточным признакам

В этой пункте рассмотривается классификация алгоритмов диаризации по промежуточным представлениям акустических характеристик сигнала.

*1) I-vector:* I-вектор (identity-vector) представляет кратковременные характеристики речи. Один из способов получения I-вектора – извлечение Мел-кепстральных коэффициентов (MFCC). Мел-кепстральные коэффициенты представлены в работе [8]. Шкала Мел соотносит воспринимаемую частоту или высоту чистого тона (Мел) с фактической измеренной частотой (Гц). Люди гораздо лучше различают небольшие изменения высоты звука на низких частотах, чем на высоких. Для получения MFCC необходимы следующие преобразования акустического сигнала:

- исходный сигнал преобразуется в частотную область преобразованием Фурье и вычисляются периодограммы;
- после перехода в частотную область применяется множество фильтров на Мел-шкале, переведенной на частотный спектр;
- наложенные фильтры пересекаются, а энергии фильтров достаточно коррелируют. Дискретное косинусное преобразование применяется к получившимся значениям и декоррелирует их.

Чаще всего при извлечении I-векторов входящий аудиопоток делят на фреймы, длиной от 20мс до 40мс (можно сказать, что на фреймах такой длины голос не изменяется), также пересечение соседних фреймов друг с другом составляет от 50% до 75% свой длины.

Извлечение Мел-кепстральных коэффициентов не единственный способ получения I-вектора, есть и другие алгоритмы [9]:

- Linear Frequency Cepstral Coefficients (LFCC);
- Perceptual Linear Predictive (PLP);
- Linear Predictive Coding (LPC).

Существует множество алгоритмов диаризации, использующих I-vector. В рамках обзора были отмечены следующие работы: в [10] предлагается улучшение метода кластеризации Bayesian GMM с модулем ресегментации, в работе [11] предложена MAP-адаптация GMM-UBM.

На данный момент алгоритмы, использующие I-векторы, не являются актуальными. Различные формы представлений, которые схожи с I-векторами, используются как промежуточные элементы перед извлечением X-векторов или подаются на вход end-to-end системам.

*2) X-vector:* Алгоритм, основанный на X-векторах, извлекает спектрограммы из входящего аудиопотока и возвращает с помощью моделей глубокого обучения представление о фрейме. Также возможна предобработка с помощью частотных фильтров для отсечения слишком низких или высоких частот, которые не используются в человеческой речи.

Преимущество X-vector в том, что нейронная сеть обучается специально для извлечения признаков, принадлежащих диктору, что помогает преодолеть большую вариативность, с которой может быть произнесена речь.

Алгоритмы извлечения X-векторов, предложенные в работах [29], [30] показали значительные улучшения результатов верификации.

Для кластеризации полученных X-векторов часто используются такие алгоритмы, как иерархическая кластеризация ([31], [32]), спектральная кластеризация ([33], [34], [35], [36]). Алгоритмы кластеризации, использующиеся в работах [36], [37], используют глубокое обучение для кластеризации.

В 2019 году Google представила алгоритм диаризации, использующий рекуррентную нейронную сеть с неограниченным количеством состояний (UIS-RNN) [12]. Достигнутый результат ошибки диаризации 7.6% является лучшим на момент публикации статьи. Алгоритм основан на предыдущей работе, в которой предложено использование X-векторов, а для их извлечения была использована LSTM нейронная сеть (нейронная сеть с долгой краткосрочной памятью) [13].

Системы, использующие в качестве признаков X-векторы, показывают более качественные результаты в сравнении с системами, использующими I-векторы. Это связано с большей репрезентативностью, углубленным моделированием речи, большей сложностью модели и увеличением вычислительных мощностей аппаратных средств.

*3) end-to-end системы:* Алгоритмы диаризации, относящиеся end-to-end типу, не используют промежуточных представлений звука. Такие алгоритмы принимают на входе спектрограмму, на выходе возвращая метку диктора.

Системы, которые используют кластеризацию x или I-векторов, имеют несколько проблем:

- они не могут быть оптимизированы для минимизации ошибок диаризации напрямую, потому что процедура кластеризации относится к типам неконтролируемых методов обучения;
- системы плохо работают с перекрывающимися сегментами (которые содержат пересекающиеся реплики нескольких дикторов на одном фрагменте) – они чаще всего вообще не учитываются.

Для решения данных проблем были предложены алгоритмы без промежуточных представлений [14], [15], [16], [38], [39].

## IV. Синусоидальная модель представления речи

### A. Синусоидальная модель

Идея синусоидального моедирования речи была представлена в работе [2]. В рамках статьи было предложено преобразование речи в звуке в оригинальное представление и алгоритм обратного преобразования в привычный формат аудиофаилов.

Первоначальная идея состоит в использовании алгоритма отслеживания треков в амплитудно-частотной области между фреймами:

- в каждом окне делается преобразование Фурье, на полученной амплитудно-частотной характеристике (АЧХ) берутся $n$ точек с наивысшей амплитудой;
- полученные точки сопостовляются с точками из предыдущих и следующих окон:
  - если расстояние (2) между ними меньше заданной величины, то они связываются;
  - последовательность связанных точек, имеющих длину больше пороговой величины, является треком.

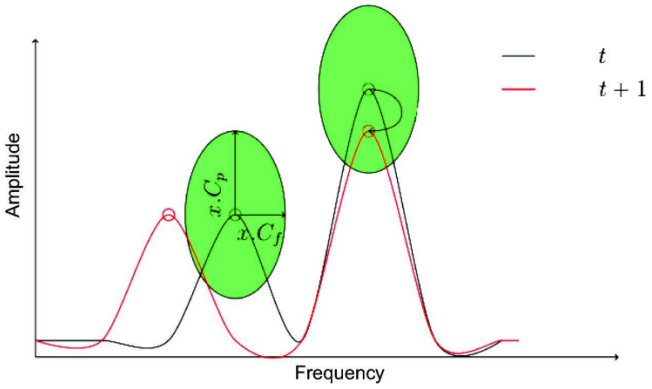


Рис. 2: Иллюстрация сопоставления пиков [3]

На рисунке 2 показана иллюстрация соединения пиков на соседних АЧХ. Два правых пика в амплитудно-частотном расстоянии меньше порогового, они будут соединены в линию. Расстояние в амплитудно-частотной области было предложено в работе [4], посвященной детектированию речи в аудиосигнале. Формула (2) для определения расстояния представлена ниже:

$$d_{i,j} = \sqrt{(\frac{f_t^i - f_{t+1}^j}{C_f})^2 + (\frac{amp_t^i - amp_{t+1}^j}{C_a})^2}, \qquad (2)$$

где:

$d_{i,j}$ – расстояние между точками;

$f_t^i; f_{t+1}^j$ – пиковые частоты, отслеженные на соседних АЧХ;

$C_f$ – эксперементально определенная константа, отображающая максимально допустимую разность частот;

$amp_t^i; amp_{t+1}^j$ – пиковые амплитуды, отслеженные на соседних АЧХ;

$C_a$ – эксперементально определенная константа, отображающая максимально допустимую разность амплитуд.

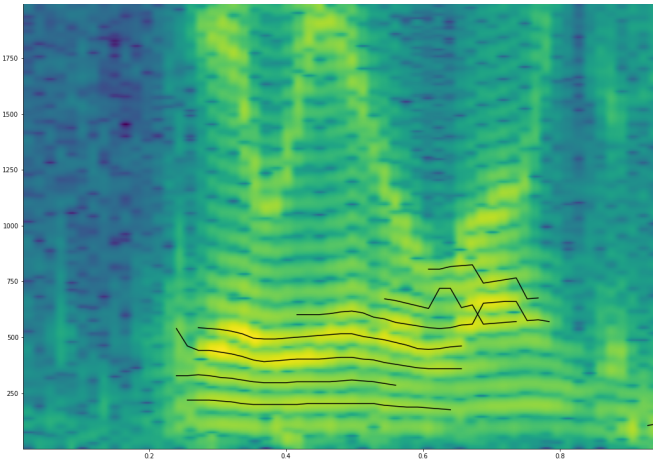


Рис. 3: Иллюстрация синусоидальной модели

На рисунке 3 показана иллюстрация модели: черные линии, нанесенные на спектрограмму, являются треком.

Треком является последовательность отслеживаемых точек в амплитудно-частотном пространстве, где каждая точка имеет три координаты – время, частоту и амплитуду.

### *B. Инструменты, предоставляющие реализацию синусоидальной модели*

В работе [3], посвященной синусоидальному моделированию эхоакустики, опубликован исходный код, реализующий синусоидальную модель и представляющий инструменты для последующего анализа и удаления шумовых треков.

Также в открытом доступе набор инструментов SMS toolbox (вебсайт [5], исходный код [6]), содержащий в том числе синтез речи из синусоидальных представлений и надстроек к ней. Данный продукт также содержит гармоническую модель представления речи [7] и ее модификаций.

## V. Проведенные эксперименты

### *A. Программная реализациия синусоидальной модели*

Программная реализация синусоидального представления была основана на исследовании синусоидального моделирования эхоакустики и эхолокации птиц [3].

Для перехода в амплитудно-частотную область применяется быстрое преобразование Фурье, размером 2048 кадров, при этом используется заполнение нулями (примечание: исходный дискретный сигнал заполняется нулями между кадрами). Такая операция позволяет приблизить результат к непрерывному преобразованию Фурье, однако при таком преобразовании утрачивается информация в фазовой области. Используется размер окна 32 мс с перекрытием 16 мс (50%). На каждом фрейме берутся 7 частот с наивысшей амплитудой, и измеряются расстояния между пиками, расстояние описано в пункте IV. При определении близости пиков на соседних АЧХ используются эксперментально подобранные константы $C_f = 300$, $C_a = 1$, описанные в формуле (2). Полученные треки, длиной менее 8 точек, убираются для фильтрации возможных шумовых треков.

Для получения треков аудиофайлы были разибиты на непересекающиеся сегменты, размером 800ms. Такой размер обусловлен относительно малой частотой дискретизации 16 кГц, и при меньших размерах разбиения информативность получившихся признаков убывает. Большой размер окна имеет недостатки, например, при частой смене дикторов много сегментов будут иметь наложения речи. Также речь может присутствовать лишь в части сегмента, если это короткое высказывание.

### *B. Полученные результаты*

В качестве набора данных был взят ISCI Corpus, длиной 70 часов и 56 различными дикторами. В рамках эксперимента были обучены модели для получения эмбеддингов: TDNN, ResNet.

Также были обучены модели для сопоставления меток диктора к эмбеддингам: алгоритмы кластеризации, UIS-RNN.

Для алгоритма, использующего спектрограмму, был получен результат $DER = 30\%$. Алгоритм, использующий синусоидальные представления, показал сравнимые результаты ($DER = 35\%$) с преимуществом работы синусоидальных представлений как детектора наличия голоса.

При оценивании качества работы системы было принято несколько допущений:

- для оценивания использовались одноканальные аудиозаписи;
- были исключены накладывающиеся реплики;
- допускались ошибки 800мс в границах сегмента.

В сравнении лучший результат был получен сочетанием ResNet сети для извлечения эмбеддингов и UIS-RNN для сопоставления меток диктора эмбеддингам.

## VI. Заключение

На сегодняшний день существует множество алгоритмов автоматической диаризации звука, использующих различные промежуточные представления речи.

В ходе исследования был воспроизведен существующий алгоритм диаризации, произведен эксперимент с использованием синусоидальной модели представления звука для получения промежуточного представления акустических характеристик сигнала и дана качественная оценка полученным результатам.

Сиусоидальная модель представления речи определяет алгоритм отслеживания треков во временом-амплитудно-частотном пространстве, при этом позволяет легко убирать шумовые треки и акустические сигналы не относящиеся к голосу. В работе [3], посвященной синусоидальному моделированию речи птиц, отмечено преимущество данного представления. На момент публикации работы, в литературе не было найдено алгоритмов диаризации, использующих синусоидальное моделирование речи.

Алгоритм, использующий синусоидальную модель в качестве промежуточного представления звука, показал сравнимые результаты в решении задачи диаризации. Предложенный алгоритм имеет преимущество в виде возможности работы синусоидальных представлений как детектора наличия голоса, что в целом позволило сделать более эффективным используемый алгоритм диаризации.

## Библиография


[1] Joint Speech Recognition and Speaker Diarization via Sequence Transduction https://ai.googleblog.com/2019/08/joint-speech-recognition-and-speaker.html

[2] R. McAulay, T. Quatieri 'Speech analysis/Synthesis based on a sinusoidal representation'

[3] Patrice Guyot, Alice Eldridge, Ying Chen Eyre-Walker, Alison Johnston, Thomas Pellegrini, et al.. Sinusoidal modelling for ecoacoustics. Annual conference Interspeech (INTERSPEECH 2016), Sep 2016, San Francisco, United States. pp. 2602-2606. ffhal-01474894f

[4] Toru Taniguchi, Mikio Tohyama, Katsuhiko Shirai 'Detection of speech and music based on spectral tracking'

[5] Spectral Modeling Synthesis Tools https://www.upf.edu/web/mtg/sms-tools

[6] Spectral Modeling Synthesis Tools code https://github.com/MTG/sms-tools

[7] Jean Larock, Yunnis Stylianou and Eric Moulines 'HNM: A Simple, Efficient Harmonic + Noise Model for Speech', Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics

[8] S. Davis and P. Mermelstein Comparison of Parametric Representations for Monosyllabic Word Recognition in

Continuously Spoken Sentences. In IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 28 No. 4, 1980.

[9] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, 'Spoken Language Processing: A Guide to Theory, Algorithm, and System Development' Prentice Hall, 2001, ISBN:0130226165

[10] Stephen H. Shum, Najim Dehak, Réda Dehak, James R. Glass 'Unsupervised Methods for Speaker Diarization:An Integrated and Iterative Approach'

[11] Giovanni Soldi, Massimiliano Todisco, Hector Delgado, Christophe Beaugeant Nicholas Evans 'Semi-supervised On-line Speaker Diarization for Meeting Data with Incremental Maximum A-posteriori Adaptation'

[12] Aonan Zhang, Quan Wang,Zhenyao Zhu,John Paisley,Chong Wang "FULLY SUPERVISED SPEAKER DIARIZATION"

[13] Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopz Moreno, "Speaker diarization with lstm," in International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 5239–5243.

[14] Yanick Lukic, Carlo Vogt, Oliver Durr, Thilo Stadelmann, 'SPEAKER IDENTIFICATION AND CLUSTERING USING CONVOLUTIONAL NEURAL NETWORKS'

[15] Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, Shinji Watanabe 'END-TO-END NEURAL SPEAKER DIARIZATION WITH SELF-ATTENTION'

[16] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe,'End-to-end neural speaker diarization withpermutation-free objectives,' inProc. Interspeech, 2019.

[17] E. Vincent, T. Virtanen, S. Gannot, Audio source separation and speech enhancement, John Wiley & Sons, 2018.

[18] D. Wang, J. Chen, Supervised speech separation based on deep learning: An overview, IEEE/ACM Transactions on Audio, Speech, and Language Processing 26 (2018) 1702–1726.

[19] R. Haeb-Umbach, S. Watanabe, T. Nakatani, M. Bacchiani, B. Hoffmeister, M. L. Seltzer, H. Zen, M. Souden, Speech processing for digital home assistants: Combining signal processing with deep-learning techniques, IEEE Signal Processing Magazine 36 (2019) 111–124.

[20] G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe, et al., Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge., in: Proceedings of the Annual Conference of the International Speech Communication Association, 2018, pp. 2808–2812.

[21] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, M. Liberman, The second DIHARD diarization challenge: Dataset, task, and baselines, Proceedings of the Annual Conference of the International Speech Communication Association (2019) 978–982.

[22] M. Diez, F. Landini, L. Burget, J. Rohdin, A. Silnova, K. Zmolıkova,O. Novotny, K. Vesely, O. Glembek, O. Plchot, et al., BUT system for DIHARD speech diarization challenge 2018., in: Proceedings of the Annual Conference of the International Speech Communication Association, 2018, pp. 2798–2802.

[23] T. Gao, J. Du, L.-R. Dai, C.-H. Lee, Densely connected progressive learning for lstm-based speech enhancement, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2018

[24] H. Erdogan, J. R. Hershey, S. Watanabe, J. Le Roux, Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2015, pp. 708–712.

[25] P. C. Loizou, Speech enhancement: theory and practice, CRC press, 2013.

[26] T. Drugman, Y. Stylianou, Y. Kida, M. Akamine, Voice activity detection: Merging source and filter-based information, IEEE Signal Processing Letters 23 (2015) 252–256

[27] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Vesely, P. Matejka, Developing a speech activity detection system for the darpa rats program, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2012, pp. 1969–1972.

[28] R. Sarikaya, J. H. Hansen, Robust detection of speech activity in thepresence of noise, in: Proceedings of the International Conference on Spoken Language Processing, volume 4, Citeseer, 1998, pp. 1455–8.

[29] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, Xvectors: Robust DNN embeddings for speaker recognition, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2018, pp. 5329–5333.

[30] D. Snyder, D. Garcia-Romero, D. Povey, S. Khudanpur, Deep neural network embeddings for text-independent speaker verification., in: Proceedings of the Annual Conference of the International Speech Communication Association, 2017, pp. 999–1003

[31] K. J. Han, S. S. Narayanan, A robust stopping criterion for agglomerative hierarchical clustering in a speaker diarization system, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2007.

[32] S. Novoselov, A. Gusev, A. Ivanov, T. Pekhovsky, A. Shulipa, A. Avdeeva, A. Gorlanov, A. Kozlov, Speaker diarization with deep speaker embeddings for dihard challenge ii., in: Proceedings of the Annual Conference of the International Speech Communication Association, 2019, pp. 1003–1007.

[33] A. Ng, M. Jordan, Y. Weiss, On spectral clustering: Analysis and an algorithm, Advances in neural information processing systems 14 (2001) 849–856.

[34] J. Luque, J. Hernando, On the use of agglomerative and spectral clustering in speaker diarization of meetings, in: Proceedings of Odyssey: The Speaker and Language Recognition Workshop, 2012, pp. 130–137.

[35] T. J. Park, K. J. Han, M. Kumar, S. Narayanan, Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap, IEEE Signal Processing Letters 27 (2019) 381–385.

[36] D. Dimitriadis, Enhancements for Audio-only Diarization Systems, arXiv preprint arXiv:1909.00082 (2019).

[37] J. Xie, R. Girshick, A. Farhadi, Unsupervised deep embedding for clustering analysis, in: Proceedings ofInternational Conference on Machine Learning, 2016, pp. 478–487.

[38] S. Horiguchi, P. Garcia, Y. Fujita, S. Watanabe, K. Nagamatsu, End-to-end speaker diarization as post-processing, arXiv preprint arXiv:2012.10055 (2020).

[39] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, K. Nagamatsu, End-to-end speaker diarization for an unknown number of speakers with encoderdecoder based attractors, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2020, pp. 269–273.

[40] Diarization hard competition https://dihardchallenge.github.io/dihard3/

# Application of sinusoidal speech modeling to the sound diarization problem

Bulat Nutfullin, Eugene Ilyushin

***Abstract*—Speech is a specific feature of human and his advantage over other species within evolution. Sound diarization is a process of sound separation, taking into account belonging to the speaker. Before the advent of deep learning and the availability of the necessary computing resources, the quality of the algorithms that determine the speaker by voice left much to be desired. Diarization has numerous applications: smart speakers, mobile phones, automatic speech translation systems. But it should be noted that the existing diarization algorithms have drawbacks, for example, the complexity of work with simultaneous speech by several speakers or the lack of diarization results for its automatic application in some areas. This explains the relevance of research in this area.**

**The sinusoidal model is an algorithm for tracking sequences of points in time-amplitude-frequency space. In existing researches, it is applied to simulations of echolocation, human speech, and speech synthesis. At the time of the study, no applications of the sinusoidal model in the problem of diarization were found in the literature.**

**The paper considers the problem of diarization and the main quality indicators used in assessing the solutions to this problem. The main intermediate representations of sound used in existing solutions are considered, and a diarization algorithm using sinusoidal speech modeling is proposed. The advantage of the proposed algorithm is the ability to operate sinusoidal representations as VAD, which in general made it possible to make the used diarization algorithm more efficient.**

***Keywords*—diarization, sound, speaker separation**

## References

[1] Joint Speech Recognition and Speaker Diarization via Sequence Transduction https://ai.googleblog.com/2019/08/joint-speech-recognition-and-speaker.html

[2] R. McAulay, T. Quatieri 'Speech analysis/Synthesis based on a sinusoidal representation'

[3] Patrice Guyot, Alice Eldridge, Ying Chen Eyre-Walker, Alison Johnston, Thomas Pellegrini, et al.. Sinusoidal modelling for ecoacoustics. Annual conference Interspeech (INTERSPEECH 2016), Sep 2016, San Francisco, United States. pp. 2602-2606. ffhal-01474894f

[4] Toru Taniguchi, Mikio Tohyama, Katsuhiko Shirai 'Detection of speech and music based on spectral tracking'

[5] Spectral Modeling Synthesis Tools https://www.upf.edu/web/mtg/sms-tools

[6] Spectral Modeling Synthesis Tools code https://github.com/MTG/sms-tools

[7] Jean Larock, Yunnis Stylianou and Eric Moulines 'HNM: A Simple, Efficient Harmonic + Noise Model for Speech', Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics

[8] S. Davis and P. Mermelstein Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. In IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 28 No. 4, 1980.

[9] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, 'Spoken Language Processing: A Guide to Theory, Algorithm, and System Development' Prentice Hall, 2001, ISBN:0130226165

[10] Stephen H. Shum, Najim Dehak, Réda Dehak, James R. Glass 'Unsupervised Methods for Speaker Diarization:An Integrated and Iterative Approach'

[11] Giovanni Soldi, Massimiliano Todisco, Hector Delgado, Christophe Beaugeant Nicholas Evans 'Semi-supervised On-line Speaker Diarization for Meeting Data with Incremental Maximum A-posteriori Adaptation'

[12] Aonan Zhang, Quan Wang,Zhenyao Zhu,John Paisley,Chong Wang "FULLY SUPERVISED SPEAKER DIARIZATION"

[13] Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopz Moreno, "Speaker diarization with lstm," in International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 5239–5243.

[14] Yanick Lukic, Carlo Vogt, Oliver Durr, Thilo Stadelmann, 'SPEAKER IDENTIFICATION AND CLUSTERING USING CONVOLUTIONAL NEURAL NETWORKS'

[15] Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, Shinji Watanabe 'END-TO-END NEURAL SPEAKER DIARIZATION WITH SELF-ATTENTION'

[16] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe,'End-to-end neural speaker diarization withpermutation-free objectives,' inProc. Interspeech, 2019.

[17] E. Vincent, T. Virtanen, S. Gannot, Audio source separation and speech enhancement, John Wiley & Sons, 2018.

[18] D. Wang, J. Chen, Supervised speech separation based on deep learning: An overview, IEEE/ACM Transactions on Audio, Speech, and Language Processing 26 (2018) 1702–1726.

[19] R. Haeb-Umbach, S. Watanabe, T. Nakatani, M. Bacchiani, B. Hoffmeister, M. L. Seltzer, H. Zen, M. Souden, Speech processing for digital home assistants: Combining signal processing with deep-learning techniques, IEEE Signal Processing Magazine 36 (2019) 111–124.

[20] G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe, et al., Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge., in: Proceedings of the Annual Conference of the International Speech Communication Association, 2018, pp. 2808–2812.

[21] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, M. Liberman, The second DIHARD diarization challenge: Dataset, task, and baselines, Proceedings of the Annual Conference of the International Speech Communication Association (2019) 978–982.

[22] M. Diez, F. Landini, L. Burget, J. Rohdin, A. Silnova, K. Zmolıkova,O. Novotny, K. Vesely, O. Glembek, O. Plchot, et al., BUT system for DIHARD speech diarization challenge 2018., in: Proceedings of the Annual Conference of the International Speech Communication Association, 2018, pp. 2798–2802.

[23] T. Gao, J. Du, L.-R. Dai, C.-H. Lee, Densely connected progressive learning for lstm-based speech enhancement, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2018

[24] H. Erdogan, J. R. Hershey, S. Watanabe, J. Le Roux, Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2015, pp. 708–712.
[25] P. C. Loizou, Speech enhancement: theory and practice, CRC press, 2013.
[26] T. Drugman, Y. Stylianou, Y. Kida, M. Akamine, Voice activity detection: Merging source and filter-based information, IEEE Signal Processing Letters 23 (2015) 252–256
[27] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Vesely, P. Matejka, Developing a speech activity detection system for the darpa rats program, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2012, pp. 1969–1972.
[28] R. Sarikaya, J. H. Hansen, Robust detection of speech activity in thepresence of noise, in: Proceedings of the International Conference on Spoken Language Processing, volume 4, Citeseer, 1998, pp. 1455–8.
[29] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, Xvectors: Robust DNN embeddings for speaker recognition, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2018, pp. 5329–5333.
[30] D. Snyder, D. Garcia-Romero, D. Povey, S. Khudanpur, Deep neural network embeddings for text-independent speaker verification., in: Proceedings of the Annual Conference of the International Speech Communication Association, 2017, pp. 999–1003
[31] K. J. Han, S. S. Narayanan, A robust stopping criterion for agglomerative hierarchical clustering in a speaker diarization system, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2007.
[32] S. Novoselov, A. Gusev, A. Ivanov, T. Pekhovsky, A. Shulipa, A. Avdeeva, A. Gorlanov, A. Kozlov, Speaker diarization with deep speaker embeddings for dihard challenge ii., in: Proceedings of the Annual Conference of the International Speech Communication Association, 2019, pp. 1003–1007.
[33] A. Ng, M. Jordan, Y. Weiss, On spectral clustering: Analysis and an algorithm, Advances in neural information processing systems 14 (2001) 849–856.
[34] J. Luque, J. Hernando, On the use of agglomerative and spectral clustering in speaker diarization of meetings, in: Proceedings of Odyssey: The Speaker and Language Recognition Workshop, 2012, pp. 130–137.
[35] T. J. Park, K. J. Han, M. Kumar, S. Narayanan, Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap, IEEE Signal Processing Letters 27 (2019) 381–385.
[36] D. Dimitriadis, Enhancements for Audio-only Diarization Systems, arXiv preprint arXiv:1909.00082 (2019).
[37] J. Xie, R. Girshick, A. Farhadi, Unsupervised deep embedding for clustering analysis, in: Proceedings ofInternational Conference on Machine Learning, 2016, pp. 478–487.
[38] S. Horiguchi, P. Garcia, Y. Fujita, S. Watanabe, K. Nagamatsu, End-to-end speaker diarization as post-processing, arXiv preprint arXiv:2012.10055 (2020).
[39] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, K. Nagamatsu, End-to-end speaker diarization for an unknown number of speakers with encoderdecoder based attractors, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2020, pp. 269–273.
[40] Diarization hard competition https://dihardchallenge.github.io/dihard3/