Компьютерные инструменты в образовании, 2019
№ 1: 5–28
УДК: 004.934.5
http://ipo.spb.ru/journal
doi:10.32603/2071-2340-2019-1-5-28

информатика

# СИНТЕЗ РЕЧИ: ПРОШЛОЕ И НАСТОЯЩЕЕ

Калиев А.[1], аспирант, kaliyev.arman@yandex.kz
Рыбин С. В.[1], доцент, svrybin@itmo.ru

[1]Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, Кронверкский пр., д. 49, 197101, Санкт-Петербург, Россия

**Аннотация**

В статье представлено описание развития методов синтеза интонационной речи от истоков до настоящего времени. Рассмотрены основные подходы, сыгравшие важную роль в становлении научного направления синтеза речи, а также современные перспективные методы. Приведена объемная библиография по данному вопросу.

**Ключевые слова:** *синтез интонационной речи, речевые сигналы, эмоциональная речь, Unit Selection, глубокие нейронные сети, просодика, акустические параметры.*

**Цитирование:** Калиев А., Рыбин С. В. Синтез речи: прошлое и настоящее // Компьютерные инструменты в образовании. 2019. № 1. С. 5–28. doi: 10.32603/2071-2340-2019-1-5-28

## 1. РАННЯЯ ИСТОРИЯ

**Артикулярный синтез речи.** Первые попытки имитации человеческой речи с помощью говорящей машины начались во 2-й половине XVIII века [1]. В 1773 году ученому Кристиану Кратзенштейну, профессору физиологии в Копенгагене, действительному члену Российской Академии Наук, удалось получить гласные звуки с помощью резонансных трубок, подключаемых к музыкальному инструменту Орган [2]. Позже Вольфганг фон Кемпелен построил «Акустико-механическую речевую машину» (1791) [3] в Вене, а в середине 1800-х годов Чарльз Уитстон [4] на основе подхода фон Кемпелена построил свою версию говорящей машины. Используя резонаторы, сделанные из кожи, его машина в ручном режиме могла изменять конфигурацию для производства различных речевых звуков, как показано на рис. 1. Хороший обзор ранней истории синтеза речи можно найти в [6].

**Появление вокодеров.** Разработанная Гельмгольцем [7] в конце XIX века теория резонаторов дала новый импульс в развитии синтеза речи. Вокальный тракт человека стал рассматриваться как последовательность резонаторов. При этом гласные звуки различаются резонансными частотами, впоследствии названными формантами.

В первой половине XX века исследования, проведенные в лаборатории Белла под руководством Флетчера [8], установили взаимосвязь между спектром речи (распределение потока силы речевого звука по частоте) и его звуковыми характеристиками, а также его разборчивостью, воспринимаемой человеческим ухом. В 30-х годах XX века Гомер Дадли под значительным влиянием исследований Флетчера разработал синтезатор речи

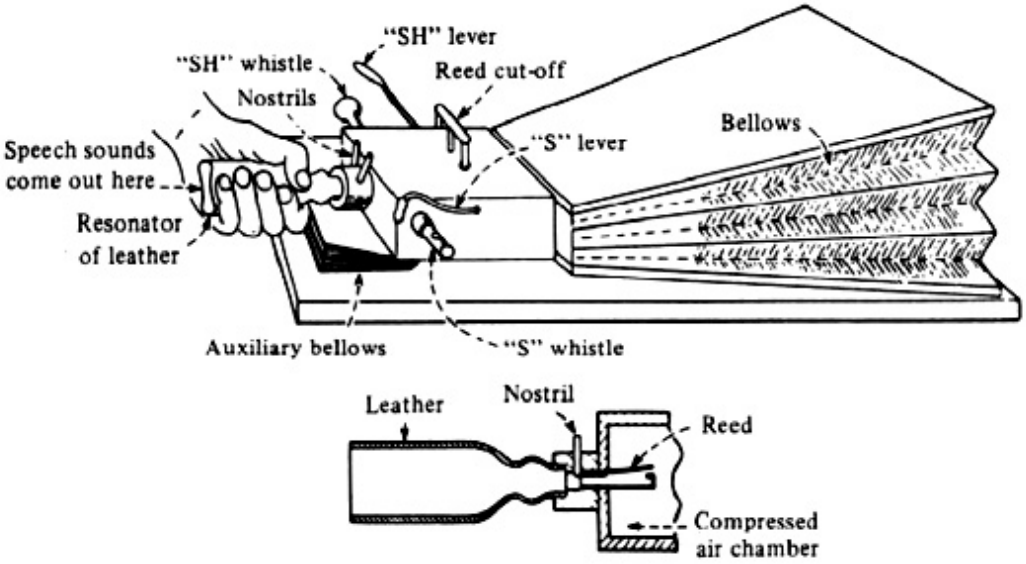


**Рис. 1.** Версия Уитстона говорящей машины Кемпелена (Фланагэн [5])

под названием VODER [9, 10], представляющий собой электрический аналог (с механическим управлением) механической говорящей машины Уитстона. На рис. 2 показана блок-схема аппарата Дадли VODER, состоящего из рычага для выбора смягченного осциллятора или шума и педали для управления частотой осциллятора (высотой звука синтезированного голоса). VODER был продемонстрирован на Всемирной выставке в Нью-Йорке в 1939 году (показано на рис. 3) и был признан важной вехой в эволюции говорящих машин.

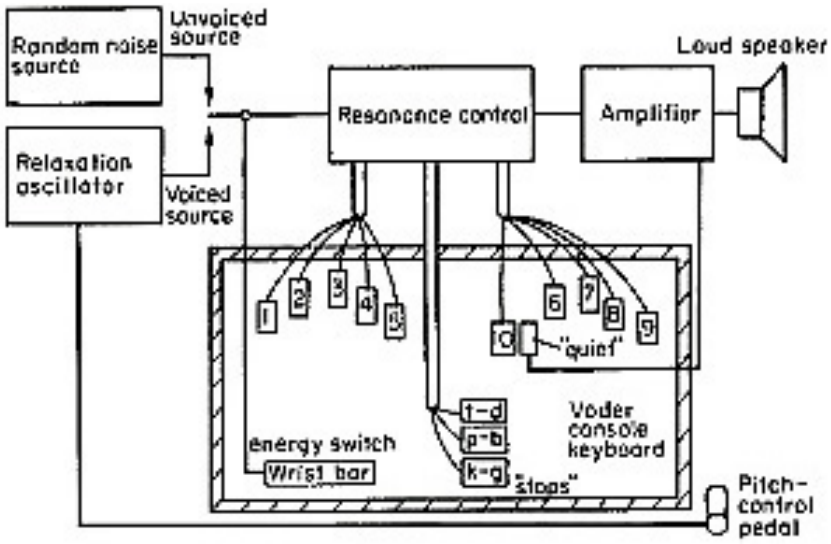


**Рис. 2.** Блок-схема аппарата Гомера Дадли VODER [9]

Таким образом, в результате исследований Флетчера и Дадли была установлена важность спектрального представления сигнала для надежной идентификации фонетических характеристик речи [11].

**Развитие вокодеров в СССР**[1]**.** Поскольку вокодер активно использовался в областях, связанных с государственной безопасностью, до сих пор есть пробелы в истории его развития. Даже в Германии, где был выдан первый патент на устройство, эти пробелы закрыты лишь частично [12, 13].

Такая ситуация справедлива и для истории развития вокодеров в Советском Союзе, представление о которой в англоязычной литературе преимущественно основано на романе А. И. Солженицына «В круге первом» (1968, издание 1978 [14]). В нем описаны некоторые подробности о работе автора в «шарашке» (специальной лаборатории МВД — МГБ) по анализу и кодированию речи, расположенной на территории бывшего Александро-Мариинского приюта в Марфино. Это описание послужило основным источником информации зарубежных работ по истории советского вокодера, например, в монографиях по истории развитии речевой техники М. Р. Шредера [15] и по истории вокодера Д. Томпкинса [16].

После окончания холодной войны появился доступ ко многим документам в бывшем Советском Союзе, в частности, были опубликованы воспоминания некоторых сотрудников лаборатории, например биографические заметки о В. А. Котельникове [17] и книга одного из ведущих инженеров К. Ф. Калачева [18] об истории лаборатории Марфино.


**Рис. 3.** VODER на Всемирной выставке в Нью-Йорке в 1939 г.

[1] Данный раздел написан по материалам работы [11].

Исследования в области вокодеров в Марфино дали несколько замечательных результатов, например:

— была разработана модификация вокодера, в которой часть речевого сигнала была оставлена во временной области, в то время как энергия сигнала в полосах частот передавалась параметрически, позже этот подход стал известен как полувокодер [19];
— был предложен принцип формантного вокодера в форме, также описанной Мансоном и Монтгомери в 1950 году [20].

С середины 1950-х годов в отечественной литературе появились открытые публикации по компрессии речи и применении вокодера, например замечательный учебник [21].

**Формантный синтезатор речи.** Первые аппараты для спектральной визуализации сигналов появились в телекоммуникационных отраслях [5]. Так, в 1946 году учеными лаборатории Белла во главе с Кёнигом [22] был представлен первый спектрограф.

Возможность визуализации речевых сигналов, как, например, визуализация акустического линейного сигнала, стала настоящим прорывом и изменила сам подход в исследовании речевых технологий [5]. Сейчас спектрограф редко упоминается в литературе, но, как и спектр стал естественным средством для фонетического анализа речевого сигнала. Немного позже спектрографа команды Кёнига в лаборатории Хаскинс был разработан синтезатор Pattern Playback, который конвертировал визуальный шаблон сигнала на спектрографе обратно в звуковой ряд [23, 24].

В 1960-х годов был разработан ряд синтезаторов речи, все они основывались на простом методе формантного синтеза, хорошо изученного за предыдущие десятилетие. Как следствие, к началу 1960-х было накоплено большое количество знаний о фонетике и акустике речи, что в дальнейшем вылилось в развитие первых систем синтеза речи по правилам [25]. Но просодические характеристики в этих системах еще не учитывалась, чаще всего они просто настраивалась путем ручного задания длительности фонем и частоты основного тона до приемлемого уровня [26]. В 1987 году Денис Клэтт опубликовал статью, где он более детально дал пояснение к каждой разработке того времени вместе с их хронологией [26].

Благодаря физиологам и фонетистам 1940-х и 1950-х годов и разработкам Pattern Playback, период с 1950-х вплоть до 1970-х годов стал десятилетиями, когда были более точно изучены семантическое, синтаксическое и лексическое влияния на фонетические свойства речи [25, 26]. Так, в развитии порождающих фонологий большую роль сыграли Хомский и Хэйлл, которые в серии своих публикаций представили базовые правила вычисления фонологических представлений из потенциального бесконечного множества предложений [27]. А в 1968 году Игнатуш Маттингли на защите своей диссертации представил первый просодический синтезатор по правилам [26].

Первое в мире использование компьютера для синтеза речи произошло в лаборатории Белла в 1962 г., где Джон Л. Келли использовал для этого свой компьютер IBM 704 [28]. Он также сумел синтезировать ритмичный голос, поющий песню Дэйзи Белл (Daisy Bell), что позже вдохновило друга Джона Джона Пирса использовать синтезированную музыку для сцены из фильма «2001 год: Космическая одиссея» [29, 30].

Как программное приложение или интегральная часть операционной системы синтез речи появился в начале 1980-х в компьютерах, таких как Apple Macintosh и Commodore Amiga [29]. До середины 1980-х исследования в этой области могли позволить себе только крупные лаборатории и компаний, но появление относительно

дешевых и мощных компьютеров способствовало распространению исследований по многим университетам и лабораториям. В дальнейшем, с увеличением компьютерной памяти и вычислительной мощности, исследователи стали искать подходы для улучшения качества синтезируемой речи, что породило множество методов конкатенаций речи [31].

**Формантный синтез речи в СССР.** К сожалению, разработки речевых технологий в Советском Союзе еще плохо изучены и требуют дальнейших исследований в изучении научных материалов, подготовленных в этой области между 60-ми и 90-ми годами прошлого века. Также надо отметить, что большой вклад в развитие технологии синтеза речи в СССР и на постсоветском пространстве внесли такие замечательные ученые, как Б. М. Лобанов, Е. А. Мурзин, М. Ф. Деркач, О. Ф. Кривнова, Л .В. Бондарко и др.

По материалу [32] первый форматный синтезатор речи для русского языка «ФОНЕМОН-1» появился в начале 70-х годов в Минске. В дальнейшем нам известно о серии промышленных синтезаторов речи «ФОНЕМОН», разработанных в СССР. Так, «ФОНЕМОН-4» имел англоязычную версию, а «ФОНЕМОН-5» был интегрирован в компьютеры класса ЕС-1840 и IBM-XT.

В конце 80-х — начале 90-х годов финансирование работ по синтезу речи в СССР практически прекратилось, и исследования продолжались только в академическом плане, что негативно сказалось на качестве разработок в этом направлении.

**Конкатативный синтез речи.** Конкатативный синтез обычно ограничивается одним диктором и использует минимальный речевой корпус. Исследователи сами, по своему усмотрению и опыту, выбирали, какие фонетические единицы использовать для склеивания — чаще это были дифоны. Корпус должен был состоять из всевозможных выбранных фонетических единиц языка. В процессе синтеза, целевая просодика (англ. target prosody) предложения склеивается из этих фонетических единиц с помощью таких методов обработки сигналов, как PSOLA [33]. Синтезированная речь, таким образом, страдала звуковыми артефактами из-за многочисленных склеек и роботизированным звуком характерным для форматных синтезаторов.

Несмотря на огромные усилия исследователей моделировать физические процессы генерации речи с помощью артикулярной модели вокального тракта, а затем на основе модели синтезировать речь, используя временные свойства речи (а позже с помощью конкатенации речевых элементов), качество синтеза речи оставалась неестественным и неприемлемым для человеческого слуха [34]. Одной из причин неудачи синтеза речи с помощью конкатенации стало то, что элементы, использовавшиеся для склеивания, были записаны в лабораторных условиях, где речь была специально записана в просодическом нейтральном тоне. Хотя речевые элементы и содержали соответствующие спектральные характеристики для заданной звуковой последовательности, они не могли достаточно правильно моделировать различные динамические артикуляторные характеристики этой последовательности в разных контекстах [35].

**Корпусный подход.** Следующий прорыв в технологии синтеза речи произошел в институте Современных Телекоммуникационных Исследований (Advanced Telecommunications Research) в Японии в конце 1980-х — начале 1990-х годов, где Иосинори Сагисака использовал обширную базу данных, хранившую множество различных речевых контекстов для каждого дифона [31, 34, 36]. Для поиска лучшей комбинации дифонов использовалась функция акустической дистанции, которая минимизировала акустические искажения между двумя фонетическими единицами [36]. Основной мотивацией использования обширных баз данных стало предположение, что при использовании большого количества фонем с разными просодическими представлени-

ями и спектральными характеристиками должна синтезироваться более естественная речь, чем это могло быть сделано из небольшого множества речевых элементов [35, 37]. В теории было показано, что при достаточном количестве дифонов и «правильной» их комбинации можно собрать высококачественную речь, максимально близкую к естественной. Однако такой подход с использованием обширных баз данных с тысячами дифонов создал новую проблему, связанную с масштабными вычислениями для поиска «правильных» дифонов.

В это же время под влиянием успехов технологии распознавания речи появились первые попытки использования для синтеза речи методов машинного обучения на основе больших корпусов. Так, в синтезе речи стали применяться скрытые марковские модели для оценки гладкости конкатенации между двумя элементами и для сглаживания спектральных разрывов [38].

**CART для предсказывания просодики.** Деревья принятий решений для предсказывания просодики первыми применила Хиршберг вместе со своими коллегами [39, 40], затем ряд ученых для той же цели использовали деревья принятий решений в комбинации с Марковскими процессами [41], скрытые Марковские модели [42] и обучение по правилам [43]. Сильверман в 1993 году показал, что просодика намного лучше предсказывается, когда сама модель обучалась на предметно-ориентированном корпусе [44].

В дальнейшем деревья принятий решения оказались чуть ли не самым успешным и распространённым решением для предсказывания просодики. Практически во всех системах CART успешно справлялся со своей задачей. Основной причиной такого успеха является несложный алгоритм и небольшие вычислительные ресурсы, что для систем, требующих непрерывного моделирования просодических характеристик, было крайне важно.

Несмотря на то, что ввод стандартных корпусов был большим шагом вперед, оставалась проблема справедливого сравнения результатов. Многие исследовательские команды публиковали зачастую недостаточно достоверные данные, связанные с их методом оценки и тестирования. Решить эту проблему предполагалось ежегодными соревнованиями, проводимым независимым институтом. Одним из самых первых среди них было соревнование DARPA Resource Management project в 1987 году [29].

В это время параллельно подходу CART для решения задач на отдельных этапах синтеза речи пытались применять искусственные нейронные сети [45–47]. Однако, хотя теоретически было известно, что нейронную сеть с несколькими скрытыми слоями можно использовать для эффективного моделирования, на практике обучить такую сеть в то время было нереально из-за запредельной стоимости вычислений [48].

## 2. UNIT SELECTION

В 1995 году Роб Донован при защите своей диссертации PhD [49] и параллельно Хант и Блак с системой CHATR [37, 50] в институте Современных Телекоммуникационных Исследований продемонстрировали использование алгоритма Unit Selection, который в последующем стал настоящим трендом в исследованиях синтеза речи.

**Описание Unit Selection.** Перед началом работы алгоритма Unit Selection на предыдущих этапах работы синтезатора речи производится сегментация речи на фонетические элементы (англ. units), и для каждого элемента определяется вектор его просодических, лингвистических и акустических параметров. Когда требуемые параметры элементов получены, наступает очередь применения метода Unit Selection для выбора оптимальной последовательности их реализаций из звуковой базы данных [37, 49, 50].

Для того чтобы определить, насколько тот или иной элемент базы подходит для синтеза данной единицы, вводятся функции стоимости замены (*англ.* target cost) и стоимости связи (*англ.* concatenation cost).

Функция стоимости замены $T(u_i, t_i)$ определяет расстояние между выбранным элементом и целевым сегментом.

Функция стоимости связи $J(u_i, u_{i-1})$ определяет расстояние между двумя последовательно выбранными элементами.

Лучшая последовательность из $n$ элементов определяется как минимальная общая стоимость согласно формуле:

$$\sum_{i-0}^{n} T(u_i, t_i)\theta_t + J(u_i, u_{i-1})\theta_j,$$

где $\theta_t$ и $\theta_j$ — веса, настраиваемые ручным способом. К началу 2010-х годов Unit Selection становится самым популярным методом синтеза, синтезированная речь которого прямолинейно зависела от качества записей. На соревнованиях Blizzard Challenge 2007 14 из 15 представленных статей работали с Unit Selection синтезом [29].

Несмотря на появившиеся к 2018 году новые подходы к синтезу речи, Unit Selection не утратил своей актуальности. На последних соревнованиях Blizzard Challenge 2018 победу одержала гибридная система синтеза с Unit Selection [51].

## 3. СТАТИСТИЧЕСКИЙ ПАРАМЕТРИЧЕСКИЙ СИНТЕЗ РЕЧИ

В начале 2000-х годов, наряду с методом Unit Selection, ростом популярности отметился метод статистического параметрического синтеза речи [52–55]. Впервые такой подход для синтеза речи был предложен в [56]. Статистический параметрический синтез речи может быть описан как система, генерирующая среднее из множества похожих речевых сегментов. Это резко контрастирует с методом Unit Selection, который склеивает естественные речевые единицы для генерации речи.

Несмотря на то, что сторонники статистического параметрического синтеза речи соглашались с мнением, что лучшие образцы Unit Selection конкатативного метода синтеза речи работают лучше, чем любой другой метод статистического параметрического синтеза речи, последний стал само собой отдельным, широко распространённым научным направлением.

**Описание статистического параметрического синтеза речи.** В типичной системе статистического параметрического синтеза речи сначала выделяются параметрические представления речи, включая спектральные и параметры возбуждения из речевого корпуса. Затем они моделируются с помощью множества генеративных моделей (например с помощью скрытых Марковских моделей (СММ)). Обычно критерий максимального правдоподобия используется для оценки параметров моделей как

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}} \{p(O|W, \lambda)\},$$

где $\lambda$ — множества параметров модели, $O$ — обучающая выборка и $W$ — множества слов соотносящихся с $O$. Затем генерируются речевые параметры $o$ для синтеза последовательности слов $w$ из множества вычисленных моделей $\hat{\lambda}$, так чтобы максимизировать

$$\hat{o} = \underset{o}{\operatorname{argmax}} \{p(o|w, \lambda)\}.$$

И, наконец, речевой сигнал реконструируется из параметрических представлений речи. Несмотря на то, что здесь может быть использована любая генеративная модель, широкое распространение получили СММ. Статистический параметрический синтез речи на основе СММ также хорошо известен как синтез речи на основе СММ [52].

Многие исследователи отмечают преимущество статистического параметрического синтеза речи над другими подходами в его гибкости [57–60], небольшом количестве речевых артефактов [61–63] и надежности [64]. Благодаря этому стало возможно изменять голосовые характеристики, стили речи и эмоций [65, 66].

Однако основным болевым местом такого подхода стало качество синтезируемой речи. Хейга Зен и другие в своей статье [67] отметили три основных фактора, которые снижают качество синтезируемой речи: вокодинг, точность акустической модели и сглаженность сигнала.

Для улучшения моделирования акустики ученые, наконец, обратились к нейронным сетям [48, 68]. Прогресс в аппаратных средствах (например появление GPU) и в программном обеспечении (например [69] и т. д.) дал возможность обучать многослойные нейронные сети на больших объемах данных.

**Глубокие нейронные сети.** С 2010 года глубокие нейронные сети были успешно применены для обучения акустической модели в распознавании речи [70–74], кодировании спектрограмм [75], обнаружении голосовой активности [76].

Кроме того, они также были успешно применены для задач статистического параметрического синтеза речи [77–81], преобразования голоса [82–84] и улучшения качества речи [85–87]. В 2015 году Жен Хуа и другие [68] опубликовали статью с полным обзором применения глубоких нейронных сетей для обучения акустической модели в системах статистического параметрического синтеза речи. В статье они разделили их на три подхода. В первом подходе рассмотрены методы применения глубоких нейронных сетей для обучения акустической модели для каждого акустического кластера отдельно. Во втором и третьем подходе использовалась глубокая нейронная сеть для предсказывания акустических параметров, где входными данными служили лингвистические представления. Отличие было в том, что если во втором подходе ученые моделировали входные и выходные данные с помощью совместного распределения вероятностей, то в третьем случае — с помощью условного распределения.

**Применения LSTM для синтеза речи.** В 2015 году Хейга Зен и Хасим Сак представили синтез речи на основе нейронных сетей долгой краткосрочной памяти (*англ.* Long-Short Term Memory, сокращённо LSTM) [88]. В статье LSTM используется для предсказания длительности пауз, фонем и отдельно для предсказывания акустических параметров. В дальнейшем применение LSTM в статистическом параметрическом синтезе речи для обучения акустической модели станет нормой и классическим подходом. Общая процедура алгоритма может быть представлена следующим образом:

1. Проводится анализ заданного текста.
2. Извлекаются лингвистические представления $x^i$ для всех фонем $i \leqslant N$ , где $N$ — общее количество фонем для заданного текста.
3. Для каждой фонемы $x^i$ предсказывается длительность фонемы $\hat{d}^i$ нейронной сетью LSTM $\Lambda_d$.
4. Для каждого кадра $\tau \leqslant \hat{d}^i$ фонемы $x^i$
   (a) составляется вектор лингвистического представления кадра $x^i_\tau$,
   (b) предсказывается акустический вектор кадра $\hat{y}^i_\tau$ нейронной сетью LSTM $\Lambda_a$ с учетом $x^i_\tau$, где элементами $\hat{y}^i_\tau$ могут быть, например, мел-частотные

кепстральные коэффициенты (MFCC) и частота основного тона (F0) аудио сигнала.

5. С помощью вокодинга $\hat{y}_\tau^i$ преобразуется в аудио сигнал.

Согласно полученным результатам, оценка MOOC по качеству синтезированной речи на английском языке составила 3.723 ± 0.105. После успешного применения LSTM для синтеза речи уже в 2016 году Хейга Зен и другие представили оптимизированную версию синтезатора речи на основе LSTM, способную работать на мобильных устройствах [89].

**Генеративные состязательные сети.** Генеративные состязательные сети (*англ.* Generative adversarial network, сокращённо GAN) показали хорошие результаты акустического моделирования, в частности, благодаря тому, что GAN успешнее решают проблему сглаженности речевого сигнала [90]. Также GAN использовались для улучшения качества вокодера путем моделирования формы голосового сигнала как волны возбуждения [91] в автоматическом распознавании речи для прямого повышения помехоустойчивости акустической модели [92] и в [93] для прогнозирования эмоций из речевого сигнала.

GAN состоят из двух конкурирующих нейронных сетей, которые можно условно разделить на генератор $G$ и дискриминатор $D$. Генератор генерирует из лингвистического представления $x$ акустический вектор $\hat{y} : G(x) : x \to \hat{y}$. В то же время в дискриминатор $D$ подаются акустические параметры $\hat{y}$ — сгенерированные с помощью генератора, и $y$ — полученные из базы данных. Во время обучения дискриминатор учится определять, какие акустические параметры получены из реального речевого сигнала, а какие «не настоящие». А генератор обучается обманывать дискриминатор, генерируя акустические параметры, максимально близкие к реальной разметке.

На практике во время такой схемы обучения, к сожалению, генератор стремится генерировать такое акустическое распределение, которое «обмануло» бы дискриминатор, а не являлось бы близким к естественной речи.

В 2017 году Юки Сайто и другие [90] представили GAN для обучения акустической модели, где генератором служил предварительно обученный LSTM. Во время обучения GAN проводилось только несколько итераций, и, таким образом, удалось улучшить акустическое распределение.

Архитектурная конструкция GAN получила большую популярность и широкое применение в современных исследованиях. Она остаётся одним из самых перспективных способов обучения акустической модели для статистического параметрического синтеза речи.

## 4. END-TO-END МОДЕЛИ

Благодаря появлению больших вычислительных возможностей, в середине 2010-х годов широкое распространение получил end-to-end метод генерирования речи. На рис. 4 представлено генеалогическое дерево, где метод end-to-end является новым направлением развития технологии синтеза речи.

Данный метод предполагает использование одной нейронной сети для генерирования сигнала речи из лингвистических параметров, соединяя, таким образом, акустическую модель с вокодером в одну нейронную сеть. Несмотря на то, что данный метод способен выдавать высококачественный речевой сигнал, он плохо применим для работы в реальном режиме, так как требует очень больших вычислительных ресурсов.

**WaveNet новая модель.** В сентябре 2016 года группа ученых из исследовательской компании DeepMind в городе Лондон представила глубокую нейронную сеть WaveNet

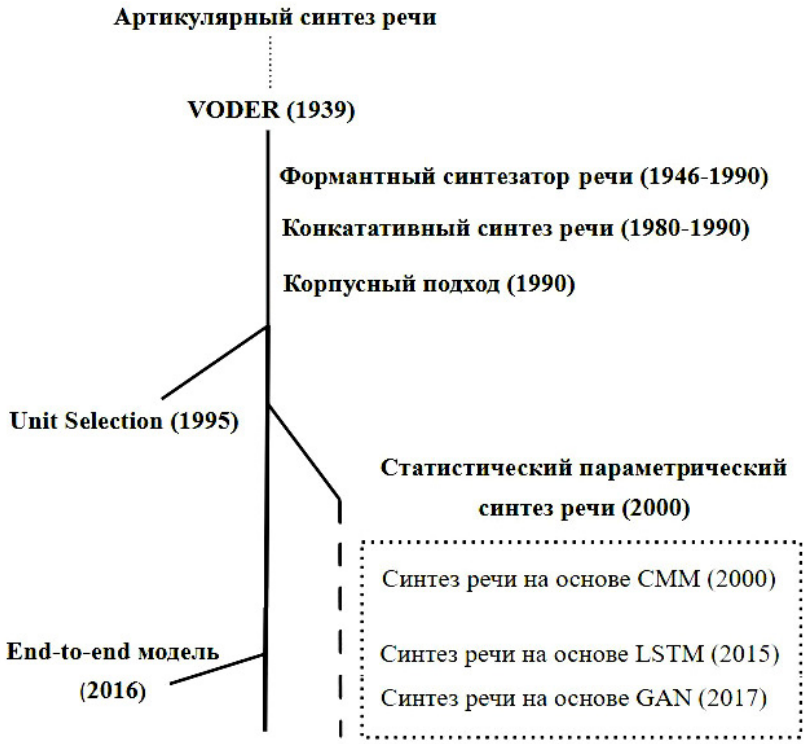


**Рис. 4.** Генеалогическое дерево развития технологии синтеза речи

[94] для генерирования необработанного аудио сигнала. WaveNet представляет собой авторегрессивную модель, которая комбинирует каузальные фильтры с расширенной свёрткой (dilated convolutions), что позволяет ей увеличивать рецептивные поля (receptive fields) экспоненциально к глубине. Согласно исследованиям авторов, увеличение рецептивных полей усиливает моделирование долгосрочных временных зависимостей в аудио сигналах.

Для демонстрации возможностей WaveNet она была применена для решения ряда задач, включая синтез речи, где входными данными служили лингвистические представления текста и логарифмическая частота основного тона, а выходными данными был сам аудио сигнал. В экспериментах WaveNet превзошла статистический параметрический и конкатативный (Unit Selection) синтез речи для английского и китайского языка. А качество синтезированной речи было максимально похоже на естественную человеческую речь, что подтверждено оценками MOOC 4.21 ± 0.081 для английской синтезированной речи и 4.08 ± 0.085 — для китайской синтезированной речи. В самой статье было только частично дано описание архитектуры нейронной сети WaveNet (рис. 5). В частности, была представлена схема обработки долговременных зависимостей речи, но не было описания схемы обработки лингвистических зависимостей речи.

Через год, в марте в 2017, группа ученых из Стэндфордского Университета Серкан О. Арик и другие представили свою версию нейронной сети WaveNet [95] для компаний Baidu (рис. 6). В опубликованной статье ученые предложили свою схему обработки лингвистических зависимостей речи. Однако, несмотря на эти две сенсационные работы, исследования в этой области ограничены из-за сложности разработки такого рода нейронных сетей.

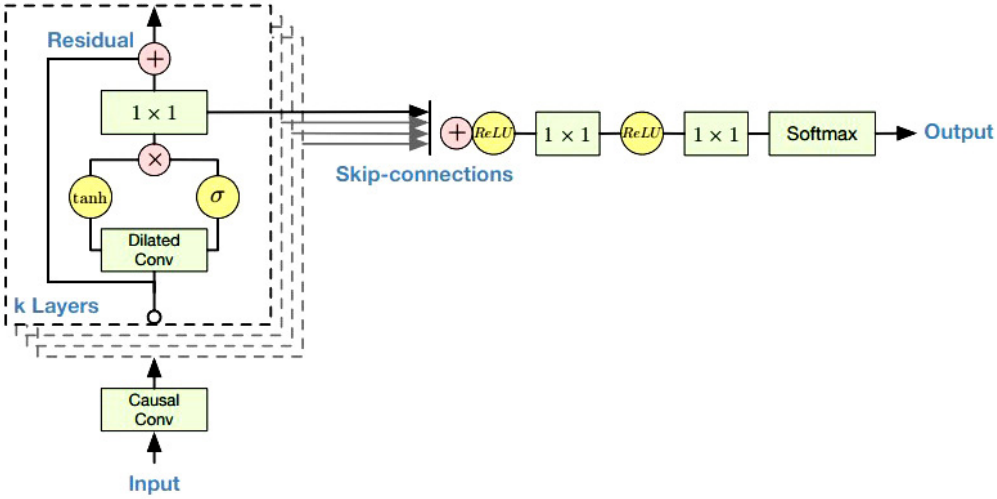


**Рис. 5.** Общая схема остаточного блока (residual block) архитектуры нейронной сети WaveNet, представленной компанией DeepMind [94]

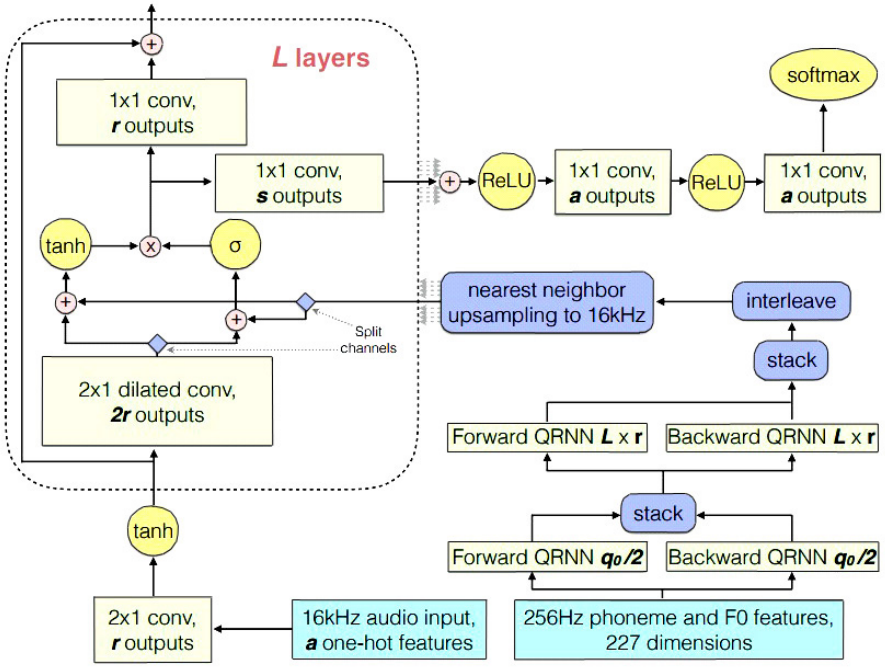


**Рис. 6.** Общая блок схема архитектуры нейронной сети WaveNet, представленной компанией Baidu [95]

**Tacotron.** Система Tacotron [96] также позволяет конвертировать лингвистические данные в аудио сигнал. Однако отличительной особенностью Tacotron является то, что она включает в себя текстовый анализатор, акустическую модель и детектор длины фонем. Система Tacotron2, которая является комбинацией систем Tacotron и WaveNet, успешно генерировала речевой сигнал, близкий к естественной речи и с очень высокой MOOC оценкой [97].

Однако достижения Tacotron и Tacatron2 были подтверждены только для английского языка, для других языков было проведено очень мало аналогичных исследований [98]. Это вызвано тем, что для реализации Tacotron для других языков в первую очередь должны быть разработаны указанные составные части системы.

## 5. СИНТЕЗ ЭМОЦИОНАЛЬНОЙ РЕЧИ

Несмотря на то, что наша повседневная речь имеет большое количество различных экспрессий, большинство исследований фокусировались на упрощённых случаях, ограничиваясь небольшим количеством категорий экспрессивной речи, включая эмоции. Как правило, в высказываниях и предложениях присутствовал только один стиль.

**Системы с явным контролем. Формантный синтез.** В ранних исследованиях эмоциональной речи использовался формантный синтезатор, поскольку он предоставлял гибкий и относительно удобный контроль над акустическими параметрами речи [99, 100]. Задачей исследователя было найти просодические правила для каждой категории эмоциональной речи и применить эти правила для синтеза эмоциональной речи из нейтрально синтезированной речи.

В 1989 году Кан разработал первый синтезатор эмоциональной речи с помощью формантного синтеза [99–101], где основные параметры формантного синтезатора настраивались ручным образом для каждой эмоциональной категории. А в 2000 году Буркхардт [102] с помощью формантного синтеза определил основные акустические параметры речи для различных эмоциональных категорий. Целью исследования Буркхардта было выявить акустические характеристики, влияющие на эмоциональное восприятие, путем изменений акустических параметров нейтрально выраженных высказываний. Согласно его экспериментам, при формантном синтезе эмоциональной речи определяющее значение имеют следующие параметры: основной тон, его среднее значение и диапазон изменения, скорость речи, фонация, точность определения гласных.

Поэтому вполне ожидаемо, что первые эмоционально экспрессивные системы синтеза речи были созданы на основе формантного синтезатора DECTalk [103].

**Конкатенации дифонов.** Естественно было ожидать, что эмоциональная речь, синтезированная путем конкатенации дифонов и изменений частоты и длительности полученного акустического сигнала с помощью алгоритма PSOLA, будет предпочтительней, чем результат формантного синтеза. Однако эксперименты по конкатенации дифонов, записанных в нейтральном тоне, и изменение их основных акустических параметров согласно выявленным соотношениям для эмоциональной речи с помощью алгоритма PSOLA не дали однозначных результатов [99].

У исследователей пока нет единого мнения о механизме того, как просодические и акустические параметры влияют на ту или иную выразительную речь, поскольку результаты экспериментов показали, что некоторые эмоциональные стили могут быть достигнуты различными комбинациями просодики и голоса [103]. Можно предположить, что существует множество стратегий для комбинаций просодики и голоса для достижения требуемой эмоциональности речи [103, 104].

Для формантного и дифонного синтеза эмоциональной речи требовалось явным образом связать эмоциональные стили с акустикой. Параметры, с помощью которых эмоций выражаются, должны были четко указаны, а их влияние на речь должно быть четко описано. Обычно правила основывались на описаниях в специализированных отчетах и литературе, на анализе собственных данных или на каких-либо других исследованиях.

Таким образом, экспрессивный синтез речи, разработанный с помощью формантного синтеза или дифонной конкатенации, является примером систем с явным контролем. Под эту категорию попадают также методы, разработанные для преобразований нейтральной речи в экспрессивную [103].

**Корпусный подход.** Ещё одним методом синтеза эмоциональной речи является запись всех дифонов с различными голосами одного диктора и последующей их конкатенацией. Такой подход часто называют корпусным. В большинстве случаев для каждой эмоциональной категории речи создаются отдельные корпуса.

Иида и Кэмпбелл [105] создали систему, которая могла синтезировать эмоциональную речь трех категорий: положительную, негативную и нейтральную. Для этого авторы записали три базы данных, состоящие соответственно из записей в положительном, нейтральном и негативном тоне. Используя алгоритм Unit Selection, они синтезировали речь в тоне той базы данных, из которой она было собрана. Джонсон и другие [106] использовали похожий подход для синтеза военной речи. Их база данных состояла из военных команд и военных речей в разной тональности. По той же технологии Питрелли и другие [107] записали базы данных для плохих и хороших новостей.

Для корпусного подхода можно выделить пять связанных между собой основных методов воспроизведения и контроля различных эмоциональных выражений и стилей разговора: моделирование, адаптация, интерполяция, контроль и оценка стиля. Моделирование стиля — это методика моделирования и генерирования определенного стиля с соответствующим корпусом для обучения. Адаптация стиля уменьшает стоимость подготовки данных, используя модель адаптации из нейтрального стиля. Промежуточные экспрессивные стили могут быть сгенерированы с помощью интерполяции двух и более стилей. Контроль стиля позволяет интуитивно управлять мерой выразительности стиля синтезированной речи. И наоборот, интенсивность стиля речи оценивается с помощью анализа обратного процесса контроля стиля.

Преимущества корпусного подхода в естественности синтезируемой экспрессивной речи. Однако сложность такого подхода — в дороговизне создания множества речевых корпусов для каждого стиля, кроме того, при таком подходе иногда трудно интерполировать стили, например, для извиняющегося вопроса.

**Комбинированная.** В 2004 году Хамза и другие [108] предложили комбинированную систему синтеза экспрессивной речи, которая включала в себя корпусный подход и подход, основанный на правилах. В своей статье такую технологию, основанную на правилах, они назвали просодико-фонологическим подходом.

В просодико-фонологическом подходе сначала на одном большом корпусе статистически моделируются акустические параметры. Затем создается словарь экспрессивных стилей, где каждый стиль соотносится с правилами последовательностей просодических разметок, полученных из небольших корпусов.

По их утверждению, некоторые экспрессивные стили больше поддаются корпусному подходу, чем просодическому фонологическому подходу. Например, передача информация в стиле «хороших новостей» лучше реализовывалась корпусным подходом из-за его сложного и системного воздействия на речевой сигнал, тогда как «акцент (эмфазис)» более подходил к просодико-фонологическому подходу из-за своего более простого, локализованного проявления. Соответственно, корпусный подход использовался для синтеза хороших новостей, плохих новостей, вопросов, в то время как просодико-фонологический подход использовался для контрастного акцента (контрастной эмфазы).

Несмотря на то, что просодико-фонологический подход в основном решает проблемы корпусного подхода, он также имеет уязвимые места. А именно: сложности в оптимизации правил, написанных от руки, и необходимость большой работы по просодической разметке корпуса. Зато в рамках такого подхода добавление нового стиля требует внесения небольших дополнений к правилам стилей просодической разметки, а не добавления целиком нового корпуса.

**Больше данных. Аудиокниги.** Поскольку создание корпусов требует больших затрат финансовых и временных, ученые начали интенсивно исследовать возможность использования аудиокниг [109–112] как возможный источник эмоциональных высказываний для подготовки обучающих данных.

Методы сбора акустических параметров для различных категорий эмоций и стилей речи в аудиокнигах часто основываются на обработке текстовой информации самих аудиокниг. Основная идея заключается в использовании текстовых описаний различных ролей и стилей их подачи. Соответственно, появляется возможность по тексту разметить аудио сигнал по характеру и стилю речи.

Например, высказывание в скобках как

*Он радостно произнёс: «Как раз вовремя!!! Мы с Саней не зря готовились!»*

будет соответственно размечено в радостном тоне согласно ключевым словам «радостно» и «произнёс». Как следствие аудиозапись этого отрезка будет проецироваться в радостную эмоциональную категорию соответствующего автора. Главные и основные требования такого метода — это наличие аудиокниг, записанных эмоционально экспрессивными дикторами, и многочисленное присутствие различных стилей речи.

Сикей и другие [113] предложили метод проекций акустических параметров к различным эмоциональным категориям речи, применяя для этих целей аудиокниги. Для этого были использованы самоорганизующиеся карты Кохонена и алгоритм кластеризации $k$-средних для кластеризации акустических параметров. Следуя этой работе, Флориан Эйбен и другие [113] построили систему синтеза экспрессивной речи на основе CMM, показав, что системы такого рода могут хорошо отображать человекоподобную экспрессию.

## Список литературы


1. *Juang B. H., Rabiner L. R.* Automatic Speech Recognition — A Brief History of the Technology Development // UC Santa Barbara. 2004. URL: http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/354_LALI-ASRHistory-final-10-8.pdf. (Дата обращения: 15.01.2019).
2. *Kratzenstein C. G.* Sur la naissance de la formation des voyelles // Journal de physique. 1782. Vol. 21. P. 358–380.
3. *Dudley H., Tarnoczy T. H.* The speaking machine of Wolfgang von Kempelen // The Journal of the Acoustical Society of America. 1950. Vol. 22. №. 2. P. 151–166.
4. *Wheatstone C.* The Scientific Papers of Sir Charles Wheatstone. The Physical Society of London, 1879.
5. *Jones M. J., Knight R. A.* The Bloomsbury companion to phonetics. A&C Black, 2013.
6. *Соломенник А. И.* Технология синтеза речи в историко-методологическом аспекте // Речевые технологии. 2013. № 1. С. 42–57.
7. *Von Helmholtz H., Ellis A. J.* On the Sensations of Tone as a Physiological Basis for the Theory of Music. London: Longmans, Green and Company, 1875. P. 576.
8. *Fletcher H.* The nature of speech and its interpretation // The Bell System Technical Journal. 1922. Vol. 1. № 1. P. 129–144.
9. *Dudley H.* The Vocoder // Bell Labs Record. 1939. Vol. 17. P. 122–126.
10. *Dudley H., Riesz R. R., Watkins S. S. A.* A synthetic speaker // Journal of the Franklin Institute. 1939. Vol. 227. № 6. P. 739–764.

11. *Hoffmann R., Birkholz P., Gabriel F., Jäckel R.* From Kratzenstein to the Soviet Vocoder: Some Results of a Historic Research Project in Speech Technology // International Conference on Speech and Computer. Springer, Cham, 2018. P. 215–225. doi: 10.1007/978-3-319-99579-3_23
12. *Hoffmann R.* Zur Entwicklung des Vocoders in Deutschland // Jahrestagung für Akustik, DAGA. 2011. P. 149–150.
13. *Hoffmann R., Gramm G.* The Sennheiser vocoder goes digital: On a German R&D project in the 1970s // 2nd Inernational. Workshop on the History of Speech Communication Research (HSCR 2017). TUDpress 2017, 2017. P. 35–44.
14. *Солженицын А. И.* В круге первом. М.: ИНКОМ НВ. 1991.
15. *Schroeder M. R.* Computer speech: recognition, compression, synthesis. Springer Science & Business Media, 2013. Vol. 35.
16. *Tompkins D.* How to wreck a nice beach: The vocoder from World War II to hip-hop, The machine speaks. Melville House. 2011. doi: 10.12801/1947-5403.2012.04.02.04
17. *Котельников В. А.* Судьба, охватившая век. Том 2: Н. В. Котельникова об отце. М.: Физматлит, 2011.
18. *Калачев, К. Ф.* В круге третьем. Воспоминания и размышления о работе Марфинской лаборатории в 1948–1951 годах. М., 1999.
19. *Schroeder M. R., David E. E.* A vocoder for transmitting 10 kc/s speech over a 3.5 kc/s channel // Acta Acustica united with Acustica. 1960. Vol. 10. № 1. P. 35–43.
20. *Munson W. A., Montgomery H. C.* A speech analyzer and synthesizer // The Journal of the Acoustical Society of America. 1950. Vol. 22. № 5. P. 678–678.
21. *Сапожков М. А.* Речевой сигнал в кибернетике и связи. М.: Связьиздат, 1963.
22. *Koenig W., Dunn H. K., Lacy L. Y.* The sound spectrograph // The Journal of the Acoustical Society of America. 1946. Vol. 18. № 1. P. 19–49.
23. *Cooper F. S., Liberman A. M., Borst J. M.* The interconversion of audible and visible patterns as a basis for research in the perception of speech // Proceedings of the National Academy of Sciences of the United States of America. 1951. Vol. 37. № 5. P. 318.
24. *Young R. W.* Review of U.S. Patent 2,432,321, Translation of Visual Symbols, R. K. Potter, assignor (9 December 1947) // The Journal of the Acoustical Society of America. Vol. 20. P. 888–889. doi: 10.1121/1.1906454
25. *Sproat R. W., Olive J. P.* Text-to-Speech Synthesis // AT&T Technical Journal. 1995. Vol. 74. № 2. P. 35–44.
26. *Klatt D. H.* Review of text-to-speech conversion for English // The Journal of the Acoustical Society of America. 1987. Vol. 82. № 3. P. 737–793.
27. *Goldsmith J., Laks B.* Generative phonology: its origins, its principles, and its successors. 2016.
28. *Mullennix J.* Computer Synthesized Speech Technologies: Tools for Aiding Impairment: Tools for Aiding Impairment. IGI Global, 2010.
29. *Suendermann D., Höge H., Black A.* Challenges in speech synthesis // Speech Technology. Springer, Boston, 2010. P. 19–32.
30. *Stork D. G.* HAL's Legacy: 2001's Computer as Dream and Reality. Mit Press, 1997.
31. *Black A. W., Lenzo K. A.* Building synthetic voices // Language Technologies Institute, Carnegie Mellon University and Cepstral LLC. 2003. Vol. 4. № 2. P. 62.
32. *Лобанов Б. М., Цирульник Л. И.* Компьютерный синтез и клонирование речи // Минск: Белорусская Наука, 2008. 342 с.
33. *Charpentier F., Stella M.* Diphone synthesis using an overlap-add technique for speech waveforms concatenation // ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, 1986. Vol. 11. P. 2015–2018. doi: 10.1109/ICASSP.1986.1168657
34. *Taylor P.* Text-to-speech synthesis. Cambridge university press, 2009.
35. *Campbell N., Black A. W.* Prosody and the selection of source units for concatenative synthesis // Progress in speech synthesis. Springer, NY, 1997. P. 279–292.
36. *Sagisaka Y., Kaiki N., Iwahashi N., Mimura K.* ATR $\mu$-Talk Speech Synthesis System // Second International Conference on Spoken Language Processing. 1992.
37. *Hunt A. J., Black A. W.* Unit selection in a concatenative speech synthesis system using a large speech database // 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings. IEEE, 1996. Vol. 1. P. 373–376.
38. *Ostendorf M., Bulyko I.* The impact of speech recognition on speech synthesis // Proceedings of 2002 IEEE Workshop on Speech Synthesis, 2002. IEEE, 2002. P. 99–106.

39. *Hirschberg J.* Pitch accent in context predicting intonational prominence from text // Artificial Intelligence. 1993. Vol. 63. № 1–2. P. 305–340.
40. *Wang M. Q., Hirschberg J.* Automatic classification of intonational phrase boundaries // Computer Speech & Language. 1992. Vol. 6. № 2. P. 175–196.
41. *Ross K., Ostendorf M.* Prediction of abstract prosodic labels for speech synthesis // Computer Speech & Language. 1996. Vol. 10. № 3. P. 155–185.
42. *Taylor P., Black A. W.* Assigning phrase breaks from part-of-speech sequences // Computer Speech & Language. 1998. Vol. 12. № 2. P. 99–117.
43. *Fordyce C. S., Ostendorf M.* Prosody prediction for speech synthesis using transformational rule-based learning // Fifth International Conference on Spoken Language Processing. 1998.
44. *Silverman K. E. A.* On customizing prosody in speech synthesis: Names and addresses as a case in point // Proceedings of the workshop on Human Language Technology. Association for Computational Linguistics, 1993. P. 317–322.
45. *Narendranath M., Murthy H. A., Rajendran S., Yegnanarayana B.* Transformation of formants for voice conversion using artificial neural networks // Speech communication. 1995. Vol. 16. № 2. P. 207–216.
46. *Watanabe T., Murakami T., Namba M., Hoya T., Ishida Y.* Transformation of spectral envelope for voice conversion based on radial basis function networks // Seventh International Conference on Spoken Language Processing. 2002.
47. *Karaali O.* Speech synthesis with neural networks // Proceedings of the 1996 World Congress on Neural Networks. 1996. P. 45–50.
48. *Zen H., Senior A., Schuster M.* Statistical parametric speech synthesis using deep neural networks // 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013. P. 7962–7966. doi: 10.1109/ICASSP.2013.6639215
49. *Donovan R. E., Woodland P. C.* Improvements in an HMM-based speech synthesizer // Fourth European Conference on Speech Communication and Technology. 1995.
50. *Campbell N., Black A. W.* Prosody and the selection of source units for concatenative synthesis // Progress in speech synthesis. Springer, NY, 1997. P. 279–292. doi: 10.1007/978-1-4612-1894-4_22
51. *Jiang Y., Zhou X., Ding C., Hu Y. J., Ling Z. H., Dai L. R.* The USTC system for Blizzard Challenge 2018 // Blizzard Challenge Workshop. 2018.
52. *Yoshimura T., Tokuda K., Masuko T., Kobayashi T., Kitamura T.* Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis // Sixth European Conference on Speech Communication and Technology. 1999.
53. *Ling Z. H., Wang R. H.* HMM-based unit selection using frame sized speech segments // Ninth International Conference on Spoken Language Processing. 2006.
54. *Black A. W.* CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling // Ninth International Conference on Spoken Language Processing. 2006.
55. *Zen H., Toda T., Nakamura M., Tokuda T.* Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005 // IEICE transactions on information and systems. 2007. Vol. 90. № 1. P. 325–333.
56. *Tokuda K., Yoshimura T., Masuko T., Kobayashi T., Kitamura T.* Speech parameter generation algorithms for HMM-based speech synthesis // 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100). IEEE, 2000. Vol. 3. P. 1315–1318. doi: 10.1109/ICASSP.2000.861820
57. *Tamura M., Masuko T., Tokuda K., Kobayashi T.* Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR // 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. IEEE, 2001. Vol. 2. P. 805–808. doi: 10.1109/ICASSP.2001.941037
58. *Yoshimura T., Tokuda K., Masuko T., Kobayashi T., Kitamura T.* Speaker interpolation in HMM-based speech synthesis system // 5th European Conference on Speech Communication and Technology. 1997.
59. *Shichiri K., Sawabe A., Tokuda K., Masuko T., Kobayashi T., Kitamura T.* Eigenvoices for HMM-based speech synthesis // Seventh International Conference on Spoken Language Processing. 2002. P. 1269–1272.
60. *Nose T., Yamagishi J., Masuko T., Kobayashi T.* A style control technique for HMM-based expressive speech synthesis // IEICE TRANSACTIONS on Information and Systems. 2007. Vol. 90. № 9. P. 1406–1413.

61. *Morioka Y., Kataoka S., Zen H., Nankaku Y., Tokuda K., Kitamura T.* Miniaturization of HMM-based speech synthesis // Autumn Meeting of ASJ. 2004. P. 325–326.
62. *Kim S. J., Kim J. J., Hahn M.* HMM-based Korean speech synthesis system for hand-held devices // IEEE Transactions on Consumer Electronics. 2006. Vol. 52. № 4. P. 1384–1390. doi: 10.1109/TCE.2006.273160
63. *Gutkin A., Gonzalvo X., Breuer S., Taylor P.* Quantized HMMs for low footprint text-to-speech synthesis // Eleventh Annual Conference of the International Speech Communication Association. 2010. P. 837–840.
64. *Yamagishi J., Nose T., Zen H., Ling Z.-H., Toda T., Tokuda K., King S., Renals S.* Robust speaker-adaptive HMM-based text-to-speech synthesis // IEEE Transactions on Audio, Speech, and Language Processing. 2009. Vol. 17. № 6. P. 1208–1230. doi: 10.1109/TASL.2009.2016394
65. *Tachibana M., Izawa S., Nose T., Kobayashi T.* Speaker and style adaptation using average voice model for style control in HMM-based speech synthesis // 2008 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2008. P. 4633–4636. doi: 10.1109/ICASSP.2008.4518689
66. *Nose T., Tachibana M., Kobayashi T.* HMM-based style control for expressive speech synthesis with arbitrary speaker's voice using model adaptation // IEICE Transactions on Information and Systems. 2009. Vol. 92. № 3. P. 489-497. doi: 10.1587/transinf.E92.D.489
67. *Zen H., Tokuda K., Black A. W.* Statistical parametric speech synthesis // Speech communication. 2009. Vol. 51. № 11. P. 1039–1064. doi: 10.1016/j.specom.2009.04.004
68. *Ling Z.-H., Kang S.-Y., Zen H., Senior A., Schuster M., Qian X.-J., Meng H., Deng L.* Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends // IEEE Signal Processing Magazine. 2015. Vol. 32. № 3. P. 35–52. doi: 10.1109/MSP.2014.2359987
69. *Dean J. et al.* Large scale distributed deep networks // Advances in neural information processing systems. 2012. P. 1223–1231.
70. *Yu D., Deng L., Dahl G.* Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition // Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning. 2010.
71. *Dahl G., Yu D., Deng L., Acero A.* Large vocabulary continuous speech recognition with context-dependent DBN-HMMS // ICASSP. 2011. P. 4688–4691.
72. *Dahl G., Yu D., Deng L., Acero A.* Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition //IEEE Transactions on audio, speech, and language processing. 2012. Vol. 20. № 1. P. 30–42. doi: 10.1109/TASL.2011.2134090
73. *Mohamed A., Dahl G. E., Hinton G.* Acoustic modeling using deep belief networks // IEEE Transactions on Audio, Speech, and Language Processing. 2012. Vol. 20. № 1. P. 14–22. doi: 10.1109/TASL.2011.2109382
74. *Sainath T. N., Kingsbury B., Soltau H., Ramabhadran B.* Optimization techniques to improve training speed of deep neural networks for large speech tasks // IEEE Transactions on Audio, Speech, and Language Processing. 2013. Vol. 21. № 11. P. 2267–2276. doi: 10.1109/TASL.2013.2284378
75. *Deng L., Seltzer M. L., Yu. D., Acero A., Mohamed A. R., Hinton G.* Binary coding of speech spectrograms using a deep auto-encoder // Eleventh Annual Conference of the International Speech Communication Association. 2010. P. 1692–1695.
76. *Zhang X. L., Wu J.* Deep belief networks based voice activity detection // IEEE Transactions on Audio, Speech, and Language Processing. 2013. Vol. 21. № 4. P. 697–710.
77. *Ling Z. H., Deng L., Yu. D.* Modeling spectral envelopes using restricted Boltzmann machines for statistical parametric speech synthesis // 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013. P. 7825–7829. doi: 10.1109/ICASSP.2013.6639187
78. *Ling Z. H., Deng L., Yu. D.* Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis // IEEE transactions on audio, speech, and language processing. 2013. Vol. 21. № 10. P. 2129–2139. doi: 10.1109/TASL.2013.2269291
79. *Kang S., Qian X., Meng H.* Multi-distribution deep belief network for speech synthesis // 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013. P. 8012–8016. doi: 10.1109/ICASSP.2013.6639225
80. *Fernandez R., Rendel A., Ramabhadran B., Hoory R.* F0 contour prediction with a deep belief network-Gaussian process hybrid model // 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013. P. 6885–6889. doi: 10.1109/ICASSP.2013.6638996

81. *Lu H., King S., Watts O.* Combining a vector space representation of linguistic context with a deep neural network for text-to-speech synthesis / /Eighth ISCA Workshop on Speech Synthesis. 2013.
82. *Chen L.-H., Ling Z.H., Song Y., Dai L. R.* Joint spectral distribution modeling using restricted boltzmann machines for voice conversion // Interspeech. 2013. P. 3052–3056.
83. *Nakashika T., Takashima R., Takiguchi T., Ariki Y.* Voice conversion in high-order eigen space using deep belief nets // Interspeech. 2013. P. 369–372.
84. *Wu Z., Chng E. S., Li H.* Conditional restricted boltzmann machine for voice conversion // 2013 IEEE China Summit and International Conference on Signal and Information Processing. IEEE, 2013. P. 104–108. doi: 10.1109/ChinaSIP.2013.6625307
85. *Lu X., Tsao Y., Matsuda S., Hori C.* Speech enhancement based on deep denoising autoencoder // Interspeech. 2013. P. 436–440.
86. *Xia B., Bao C.* Speech enhancement with weighted denoising auto-encoder // INTERSPEECH. 2013. P. 3444–3448.
87. *Xu Y., Du J., Dai L. R., Lee C. H.* An experimental study on speech enhancement based on deep neural networks // IEEE Signal processing letters. 2014. Vol. 21. № 1. P. 65–68. doi: 10.1109/LSP.2013.2291240
88. *Zen H., Sak H.* Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis // 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015. P. 4470–4474. doi: 10.1109/ICASSP.2015.7178816
89. *Zen H., Agiomyrgiannakis Y., Egberts N., Henderson F., Szczepaniak P.* Fast, compact, and high quality LSTM-RNN based statistical parametric speech synthesizers for mobile devices // arXiv preprint arXiv:1606.06061. 2016.
90. *Saito Y., Takamichi S., Saruwatari H.* Statistical parametric speech synthesis incorporating generative adversarial networks // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2018. Vol. 26. № 1. P. 84–96. doi: 10.1109/TASLP.2017.2761547
91. *Bollepalli B., Juvela L., Alku P.* Generative adversarial network-based glottal waveform model for statistical parametric speech synthesis // Interspeech. 2017. P. 3394–3398.
92. *Liu B., Nie S., Zhang Y., Ke D., Liang S., Liu W.* Boosting noise robustness of acoustic model via deep adversarial training // 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018. P. 5034–5038. doi: 10.1109/ICASSP.2018.8462093
93. *Han J., Zhang Z., Ren Z., Ringeval F., Schuller B.* Towards conditional adversarial training for predicting emotions from speech // 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018. P. 6822–6826. doi: 10.1109/ICASSP.2018.8462579
94. *Van Den Oord A. et al.* WaveNet: A generative model for raw audio // SSW. 2016. Vol. 125.
95. *Arik S. Ö. et al.* Deep voice: Real-time neural text-to-speech // Proceedings of the 34th International Conference on Machine Learning. JMLR, 2017. P. 195-204.
96. Wang Y. et al. Tacotron: Towards end-to-end speech synthesis // arXiv preprint arXiv:1703.10135. 2017.
97. *Shen J. et al.* Natural tts synthesis by conditioning WaveNet on mel spectrogram predictions // 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018. P. 4779–4783. doi: 10.1109/ICASSP.2018.8461368
98. *Yasuda Y., Wang X., Takaki S., Yamagishi J.* Investigation of enhanced Tacotron text-to-speech synthesis systems with self-attention for pitch accent language // ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019. P. 6905–6909.
99. *Schröder M.* Emotional speech synthesis: A review // Seventh European Conference on Speech Communication and Technology. 2001. P. 561–564.
100. *Govind D., Prasanna S. R. M.* Expressive speech synthesis: a review // International Journal of Speech Technology. 2013. Vol. 16. № 2. P. 237–260.
101. *Cahn J. E.* The generation of affect in synthesized speech // Journal of the American Voice I/O Society. 1989. Vol. 8. № 1. P. 1–19.
102. *Burkhardt F., Sendlmeier W. F.* Verification of acoustical correlates of emotional speech using formant-synthesis // ISCA Tutorial and Research Workshop (ITRW) on speech and emotion. 2000.
103. *Schröder M.* Expressive speech synthesis: Past, present, and possible futures // Affective information processing. Springer, London, 2009. P. 111–126. doi: 10.1007/978-1-84800-306-4_7
104. Schröder M. Can emotions be synthesized without controlling voice quality // Phonus. 1999. Vol. 4. P. 35–50.

105. *Iida A., Campbell N.* Speech database design for a concatenative text-to-speech synthesis system for individuals with communication disorders //International Journal of Speech Technology. 2003. Vol. 6. № 4. P. 379–392. doi: 10.1023/A:1025761017833
106. *Johnson W. L., Narayanan S. S., Whitney R., Das R., Bulut M., LaBore C.* Limited domain synthesis of expressive military speech for animated characters // Proceedings of 2002 IEEE Workshop on Speech Synthesis, 2002. IEEE, 2002. P. 163–166. doi: 10.1109/WSS.2002.1224399
107. *Pitrelli J. F., Bakis R., Eide E. M., Fernandez R., Hamza W., Picheny M. A.* The IBM expressive text-to-speech synthesis system for American English // IEEE Transactions on Audio, Speech, and Language Processing. 2006. Vol. 14. № 4. P. 1099–1108. doi: 10.1109/TASL.2006.876123
108. *Hamza W., Eide E., Bakis R., Picheny M., Pitrelli J.* The IBM expressive speech synthesis system // Eighth International Conference on Spoken Language Processing. 2004.
109. *Zhao Y., Peng D., Wang L., Chu M., Chen Y., Yu P., Guo J.* Constructing stylistic synthesis databases from audio books // Ninth International Conference on Spoken Language Processing. 2006.
110. *Prahallad K., Toth A. R., Black A. W.* Automatic building of synthetic voices from large multi-paragraph speech databases // Eighth Annual Conference of the International Speech Communication Association. 2007.
111. *Braunschweiler N., Gales M. J. F., Buchholz S.* Lightly supervised recognition for automatic alignment of large coherent speech recordings // Eleventh Annual Conference of the International Speech Communication Association. 2010.
112. *Eyben F., Buchholz S., Braunschweiler N., Latorre J., Wan V., Gales M. J., Knill, K.* Unsupervised clustering of emotion and voice styles for expressive TTS // 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2012. P. 4009–4012. doi: 10.1109/ICASSP.2012.6288797
113. *Székely E., Cabral J. P., Cahill P., Carson-Berndsen J.* Clustering expressive speech styles in audiobooks using glottal source parameters // Twelfth Annual Conference of the International Speech Communication Association. 2011. P. 2409–2412.

*Поступила в редакцию 15.01.2019, окончательный вариант — 21.02.2019.*

Computer tools in education, 2019
№ 1: 5–28
http://ipo.spb.ru/journal
doi:10.32603/2071-2340-2019-1-5-28

# Speech Synthesis: Past and Present

Kaliev A.[1], postgraduate student, kaliyev.arman@yandex.kz
Rybin S. V.[1], associate professor, svrybin@itmo.ru

[1]ITMO University, 49, Kronverksky pr., 197101, Saint-Petersburg, Russia

**Abstract**

The article describes the development of the speech synthesis methods from the beginnings to the present. The main approaches that have played an important role in the development of the speech synthesis, as well as modern advanced methods are considered. The extensive bibliography on this issue is also given.

**Keywords:** *synthesis of intonation speech, speech signals, emotional speech, Unit Selection, deep neural networks, prosodics, acoustic parameters.*

**Citation:** A. Kaliev and S. V. Rybin, “Speech Synthesis: Past and Present,” *Computer tools in education*, no. 1, pp. 5–28, 2019 (in Russian); doi:10.32603/2071-2340-2019-1-5-28

## References


1. B.-H. Juang and L. Rabiner, “Automatic Speech Recognition — A Brief History of the Technology Development,” UC Santa Barbara, 2004. [Online]. Available: http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/354_LALI-ASRHistory-final-10-8.pdf
2. C. G. Kratzenstein, “Sur la raissance de la formation des voyelles,” *J. Phys.*, vol. 21, pp. 358–380, 1782.
3. H. Dudley and T. H. Tarnoczy, “The Speaking Machine of Wolfgang von Kempelen,” *J. Acoust. Soc. Am.*, vol. 22, pp. 151–166, 1950.
4. C. Wheatstone, *The Scientific Papers of Sir Charles Wheatstone*, London: The Physical Society of London, 1879.
5. M. J. Jones and R.-A. Knight, eds., *The Bloomsbury companion to phonetics*, London: A&C Black, 2013.
6. A. I. Solomennik, “Tekhnologiya sinteza rechi v istoriko-metodologicheskom aspekte” [Technology speech synthesis in the historical and methodological aspect], Speech Technology, no. 1, pp. 42–57, 2013 (in Russian).
7. H. Von Helmholtz and A. J. Ellis, *On the Sensations of Tone as a Physiological Basis for the Theory of Music*, London: Longmans, Green and Company, 1875. p. 576.
8. H. Fletcher, “The nature of speech and its interpretation,” *The Bell System Technical Journal*, vol. 1, no. 1, pp. 129–144, 1922.
9. H. Dudley, “The Vocoder,” *Bell Labs Record*, vol. 17, pp. 122–126, 1939.
10. H. Dudley, R. R. Riesz, and S. A. Watkins, “A Synthetic Speaker,” *J. Franklin Institute*, vol. 227, pp. 739–764, 1939.
11. R. Hoffmann, P. Birkholz, F. Gabriel, and R. Jäckel, “From Kratzenstein to the Soviet Vocoder: Some Results of a Historic Research Project in Speech Technology,” in *International Conference on Speech and Computer*. (SPECOM 2018), Springer, Cham, 2018. pp. 215–225. doi: 10.1007/978-3-319-99579-3_23.
12. R. Hoffmann, “Zur Entwicklung des Vocoders in Deutschland,” *Jahrestagung für Akustik*, DAGA, pp. 149–150, 2011.
13. R. Hoffmann and G. Gramm, “The Sennheiser vocoder goes digital: On a German R&D project in the 1970s,” in *2nd Inernational Workshop on the History of Speech Communication Research* (HSCR 2017), 2017. pp. 35–44.
14. A. Solzhenitsyn, *The First Circle*, Moscow: INCOM NV, 1991.
15. M. R. Schroeder, *Computer speech: recognition, compression, synthesis*, Springer Science & Business Media, vol. 35, 2013.
16. D. Tompkins, *How to wreck a nice beach: The vocoder from World War II to hip-hop, The machine speak*, Melville House, 2011; doi: 10.12801/1947-5403.2012.04.02.04
17. “N. V. Kotel’nikova ob ottse” [N. V. Kotelnikov about father], in *Kotelnikov, Sud’ba, okhvativshaya vek*, N. V. Kotel’nikova and A. S. Prohorov, eds., vol. 2, Moscow: Phizmatlit, 2011 (in Russian).
18. K. F. Kolachev, *V kruge tret’em. Vospominaniya i razmyshleniya o rabote Marfinskoi laboratorii v 1948–1951 godakh* [In the third circle. Memoirs and Reflections on the Work of the Martha Laboratory in 1948–1951], Moscow, 1999 (in Russian).
19. M. R. Schroeder and E. E. David, “A vocoder for transmitting 10 kc/s speech over a 3.5 kc/s channel,” *Acta Acustica united with Acustica*, vol. 10, no. 1, pp. 35–43, 1960.
20. W. A. Munson and H. C. Montgomery, “A speech analyzer and synthesizer,” *The Journal of the Acoustical Society of America*, vol. 22, no. 5, pp. 678–678, 1950.
21. M. A. Sapozhkov, *Rechevoi signal v kibernetike i svyazi* [Speech signal in cybernetics and communication], Moscow: Svyaz’izdat, 1963.
22. W. Koenig, H. K. Dunn, and L. Y. Lacy, “The sound spectrograph,” *The Journal of the Acoustical Society of America*, vol. 18. no. 1, pp. 19–49, 1946.
23. F. S. Cooper, A. M. Liberman, and J. M. Borst, “The interconversion of audible and visible patterns as a basis for research in the perception of speech,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 37, no. 5, p. 318, 1951.
24. R. W. Young, “Review of U.S. Patent 2,432,321, Translation of Visual Symbols, R. K. Potter, assignor (9 December 1947),” *The Journal of the Acoustical Society of America*, vol. 20, no. 6, pp. 888–889, 1948; doi: 10.1121/1.1906454
25. R. W. Sproat and J. P. Olive, “Text-to-Speech Synthesis,” *AT&T Technical Journal*, vol. 74, no. 2, pp. 35–44, 1995.
26. D. H. Klatt, “Review of text-to-speech conversion for English,” *The Journal of the Acoustical Society of America*, vol. 82, no. 3, pp. 737–793, 1987.

27. J. Goldsmith and B. Laks, “Generative phonology: its origins, its principles, and its successors,” *The Cambridge History of Linguistics*, Cambridge: Cambridge University Press, 2011.
28. J. Mullennix, *Computer Synthesized Speech Technologies: Tools for Aiding Impairment*, IGI Global, 2010.
29. D. Suendermann, H. Höge, and A. Black, “Challenges in speech synthesis,” *Speech Technology*, Boston: Springer, pp. 19–32, 2010.
30. D. G. Stork, *HAL’s Legacy: 2001’s Computer as Dream and Reality*, Mit Press, 1997.
31. A. W. Black and K. A. Lenzo, “Building synthetic voices,” *Language Technologies Institute, Carnegie Mellon University and Cepstral LLC.*, vol. 4, no. 2, p. 62, 2003.
32. B. M. Lobanov and L. I. Tsirul’nik, *Komp’yuternyi sintez i klonirovanie rechi* [Computer synthesis and speech cloning], Minsk: Belarusian Science, 2008.
33. F. Charpentier and M. Stella, “Diphone synthesis using an overlap-add technique for speech waveforms concatenation,” in *ICASSP ’86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Tokyo, Japan, 1986, pp. 2015–2018; doi: 10.1109/ICASSP.1986.1168657
34. P. Taylor, *Text-to-speech synthesis*, Cambridge university press, 2009.
35. N. Campbell and A. W. Black, “Prosody and the selection of source units for concatenative synthesis,” *Progress in speech synthesis*, New York: Springer, pp. 279–292, 1997.
36. Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura, “ATR v–Talk Speech Synthesis System,” *Proc. ICSLP*, pp. 483–486, 1992.
37. A. Hunt and A. Black, “Unit selection in a concatenative speech synthesis system using large speech database,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1996, pp. 373—376.
38. M. Ostendorf and I. Bulyko, “The impact of speech recognition on speech synthesis”, in *Proc, IEEE Workshop Speech Synthesis*, Santa Monica, 2002, pp. 99–106.
39. J. Hirschberg, “Pitch accent in context predicting intonational prominence from text,” Artificial Intelligence, vol. 63, no. 1-2, pp. 305–340, 1993.
40. M. Q. Wang and J. Hirschberg, “Automatic classification of intonational phrase boundaries,” *Computer Speech & Language*, vol. 6, no. 2, pp. 175–196, 1992.
41. K. Ross and M. Ostendorf, “Prediction of abstract prosodic labels for speech synthesis,” *Computer Speech & Language*, vol. 10, no. 3, pp. 155–185, 1996.
42. P. Taylor and A. W. Black, “Assigning phrase breaks from part-of-speech sequences,” *Computer Speech & Language*, vol. 12, no. 2, pp. 99–117, 1998.
43. C. S. Fordyce and M. Ostendorf, “Prosody prediction for speech synthesis using transformational rule-based learning,” in *Proc. 5th Int. Conf. on Spoken Language Processing*, (ICSLP), Sydney, Australia, 1998.
44. K. E. A. Silverman, “On customizing prosody in speech synthesis: Names and addresses as a case in point,” in *Proc. of the workshop on Human Language Technology. Association for Computational Linguistics*, 1993, pp. 317–322.
45. M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana, “Transformation of formants for voice conversion using artificial neural networks,” *Speech communication*, vol. 16, no. 2, pp. 207–216, 1995.
46. T. Watanabe, T. Murakami, M. Namba, T. Hoya, and Y. Ishida, “Transformation of spectral envelope for voice conversion based on radial basis function networks,” in *Proc. 7th International Conference on Spoken Language Processing* (ICSLP), Denver, USA, 2002.
47. O. Karaali, “Speech synthesis with neural networks,” in *Proc. of the World Congress on Neural Networks* (WCNN’96), San Diego, USA, 1996, pp. 45–50.
48. H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing* (ICASSP), Vancouver, Canada, 2013, pp. 7962–7966; doi: 10.1109/ICASSP.2013.6639215
49. R. E. Donovan and P. C. Woodland, “Improvements in an HMM-based speech synthesizer,” in *Proc. 4th European Conf. on Speech Communication and Technology* (ESCA), Madrid, Spain, 1995.
50. N. Campbell and A. W. Black, “Prosody and the selection of source units for concatenative synthesis,” in *Progress in speech synthesis*, J. P. H. van Santen, J. P. Olive, R. W. Sproat, J. Hirschberg, eds., New York, NY: Springer, 1997, pp. 279–292; doi: 10.1007/978-1-4612-1894-4_22
51. Y. Jiang, X. Zhou, C. Ding, Y.-J. Hu, Z.-H. Ling, and L.-R. Dai, “The USTC system for Blizzard Challenge 2018,” *Blizzard Challenge Workshop*, 2018.
52. T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis” in *Sixth European Conf. on Speech Communication and Technology* (EUROSPEECH’99), Budapest, Hungary, 1999.
53. Z. H. Ling and R. H. Wang, “HMM-based unit selection using frame sized speech segments,” in *Ninth*

*Int.l Conf. on Spoken Language Processing* (INTERSPEECH 2006 — ICSLP), Pittsburgh, PA, 2006.
54. A. W. Black, "CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling, in *Ninth Int.l Conf. on Spoken Language Processing* (INTERSPEECH 2006 — ICSLP), Pittsburgh, PA, 2006.
55. H. Zen, T. Toda, M. Nakamura, and T. Tokuda, “Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005,” *IEICE transactions on information and systems*, vol. 90, no. 1, pp. 325–333, 2007.
56. K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” in *2000 IEEE Int. Conf. on Acoustics, Speech, and Signal Proc.*, 2000, vol. 3, pp. 1315–1318; doi: 10.1109/ICASSP.2000.861820
57. M. Tamura, T. Masuko, K. Tokuda, T. Kobayashi, “Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR,” in *2001 IEEE Int. Conf. on Acoustics, Speech, and Signal Proc.*, Salt Lake City, UT, 2001, vol. 2, pp. 805–808; doi: 10.1109/ICASSP.2001.941037
58. T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Speaker interpolation in HMM-based speech synthesis system,” *5th European Conf. on Speech Communication and Technology*, (EUROSPEECH ’97) Rhodes, Greece, 1997.
59. K. Shichiri, A. Sawabe, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Eigenvoices for HMM-based speech synthesis,” in *7th Int. Conf. on Spoken Language Proc.*, (INTERSPEECH 2002), Denver, Colorado, 2002, pp. 1269–1272.
60. T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, “A style control technique for HMM-based expressive speech synthesis,” *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 9, pp. 1406–1413, 2007.
61. Y. Morioka, S. Kataoka, H. Zen, Y. Nankaku, K. Tokuda, and T. Kitamura, “Miniaturization of HMM-based speech synthesis,” in *Autumn Meeting of ASJ*, 2004, pp. 325–326.
62. S. J. Kim, J. J. Kim, and M. Hahn, “HMM-based Korean speech synthesis system for hand-held devices,” *IEEE Transactions on Consumer Electronics*, vol. 52, no. 4, pp. 1384–1390, 2006; doi: 10.1109/TCE.2006.273160
63. A. Gutkin, X. Gonzalvo, S. Breuer, and P. Taylor, “Quantized HMMs for low footprint text-tospeech synthesis,” *Eleventh Annu. Conf. of the Int. Speech Communication Association*, Makuhari, Chiba, Japan, 2010, pp. 837–840.
64. J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, “Robust speaker-adaptive HMM-based text-to-speech synthesis,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1208–1230, 2009; doi: 10.1109/TASL.2009.2016394
65. M. Tachibana, S. Izawa, T. Nose, and T. Kobayashi, “Speaker and style adaptation using average voice model for style control in HMM-based speech synthesis,” in *2008 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Las Vegas, NV, 2008, pp. 4633–4636; doi: 10.1109/ICASSP.2008.4518689
66. T. Nose, M. Tachibana, and T. Kobayashi, “HMM-based style control for expressive speech synthesis with arbitrary speaker’s voice using model adaptation,” *IEICE Transactions on Information and Systems*, vol. 92, no. 3, pp. 489–497, 2009; doi: 10.1587/transinf.E92.D.489
67. H. Zen, K. Tokuda, and A. W. Black, “Statistical parametric speech synthesis,” *Speech communication*, vol. 51, no. 11, pp. 1039–1064, 2009; doi: 10.1016/j.specom.2009.04.004
68. Z.-H. Ling, S.-Y. Kang, H. Zen, A. Senior, M. Schuster, X.-J. Qian, H. Meng, L. Deng, “Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends,” *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 35–52, 2015; doi: 10.1109/MSP.2014.2359987
69. J. Dean et al, “Large scale distributed deep networks,” in *Advances in neural information processing systems* (NIPS 2012), Lake Tahoe, NV, 2012, pp. 1223–1231.
70. D. Yu, L. Deng, and G. Dahl, “Roles of pre-training and fine-tuning in context-dependent DBNHMMs for real-world speech recognition,” *Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2010.
71. G. Dahl, D. Yu, L. Deng, and A. Acero, “Large vocabulary continuous speech recognition with context-dependent DBN-HMMS,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, (ICASSP 2011), Prague, Czech Republic, 2011, pp. 4688–4691.
72. G. Dahl, D. Yu, L. Deng, and A. Acero, “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition,” *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 1, pp. 30–42, 2012; doi: 10.1109/TASL.2011.2134090
73. A. Mohamed, G. E. Dahl, and G. Hinton, “Acoustic modeling using deep belief networks,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012; doi: 10.1109/TASL.2011.2109382
74. T. N. Sainath, B. Kingsbury, H. Soltau, and B. Ramabhadran, “Optimization techniques to improve

training speed of deep neural networks for large speech tasks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2267–2276, 2013; doi: 10.1109/TASL.2013.2284378
75. L. Deng, M. L. Seltzer, D. Yu, A. Acero, A. R. Mohamed, and G. Hinton, "Binary coding of speech spectrograms using a deep auto-encoder," in *Eleventh Annual Conference of the International Speech Communication Association*, Makuhari, Chiba, Japan, 2010, pp. 1692–1695.
76. X. L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697–710, 2013.
77. Z. H. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted Boltzmann machines for statistical parametric speech synthesis," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada, 2013, pp. 7825–7829; doi: 10.1109/ICASSP.2013.6639187
78. Z. H. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis," *IEEE transactions on audio, speech, and language processing*, vol. 21, no. 10, pp. 2129–2139, 2013; doi: 10.1109/TASL.2013.2269291
79. S. Kang, X. Qian, and H. Meng, "Multi-distribution deep belief network for speech synthesis," in *2013 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada, 2013, pp. 8012–8016; doi: 10.1109/ICASSP.2013.6639225
80. R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "F0 contour prediction with a deep belief network-Gaussian process hybrid model," in *2013 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada, 2013, pp. 6885–6889; doi: 10.1109/ICASSP.2013.6638996
81. H. Lu, S. King, and O. Watts, "Combining a vector space representation of linguistic context with a deep neural network for text-to-speech synthesis," in *Eighth ISCA Workshop on Speech Synthesis*, Barcelona, Catalonia, Spain, 2013.
82. L.-H. Chen, Z.-H. Ling, Y. Song, and L.-R. Dai, "Joint spectral distribution modeling using restricted boltzmann machines for voice conversion," in *14th Annu. Conf. of the Int. Speech Communication Association* (INTERSPEECH 2013), Lyon, France, 2013, pp. 3052–3056.
83. T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, "Voice conversion in high-order eigen space using deep belief nets," in *14th Annu. Conf. of the Int. Speech Communication Association* (INTERSPEECH 2013), Lyon, France, 2013, pp. 369–372.
84. Z. Wu, E. S. Chng, and H. Li, "Conditional restricted boltzmann machine for voice conversion," in *2013 IEEE China Summit and Int. Conf. on Signal and Information Processing*, Beijing, China, 2013, pp. 104–108; doi: 10.1109/ChinaSIP.2013.6625307
85. X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *14th Annu. Conf. of the Int. Speech Communication Association* (INTERSPEECH 2013), Lyon, France, 2013, pp. 436–440.
86. B. Xia and C. Bao, "Speech enhancement with weighted denoising auto-encoder," in *14th Annu. Conf. of the Int. Speech Communication Association* (INTERSPEECH 2013), Lyon, France, 2013, pp. 3444–3448.
87. Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters*, vol. 21, no. 1, pp. 65–68, 2014; doi: 10.1109/LSP.2013.2291240
88. H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *2015 IEEE Int. Conf. on Acoustics, Speech and Signal Processing* (ICASSP 2015), Brisbane, QLD, Australia, 2015, pp. 4470–4474; doi: 10.1109/ICASSP.2015.7178816
89. H. Zen, Y. Agiomyrgiannakis, N. Egberts, F. Henderson, and P. Szczepaniak, "Fast, compact, and high quality LSTM-RNN based statistical parametric speech synthesizers for mobile devices," *arXiv preprint*, arXiv:1606.06061, 2016.
90. Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 84–96, 2018, doi: 10.1109/TASLP.2017.2761547
91. B. Bollepalli, L. Juvela, and P. Alku, "Generative adversarial network?based glottal waveform model for statistical parametric speech synthesis," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 3394—3398.
92. B. Liu, S. Nie, Y. Zhang, D. Ke, S. Liang, and W. Liu, "Boosting noise robustness of acoustic model via deep adversarial training," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing* (ICASSP), Calgary, Canada, 2018, pp. 5034–5038; doi: 10.1109/ICASSP.2018.8462093
93. J. Han, Z. Zhang, Z. Ren, F. Ringeval, and B. Schuller, "Towards conditional adversarial training for

predicting emotions from speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing* (ICASSP), Calgary, Canada, 2018, pp. 6822–6826; doi: 10.1109/ICASSP.2018.8462579
94. A. Van Den Oord et al. "WaveNet: A generative model for raw audio," SSW, 2016, vol. 125.
95. S. Ö. Arik et al., "Deep voice: Real-time neural text-to-speech," in *Proc. of the 34th Int. Conf. on Machine Learning* (ICML), Sydney, Australia, 2017, pp. 195–204.
96. Y. Wang et al., "Tacotron: Towards end-to-end speech synthesis," arXiv preprint arXiv:1703.10135, 2017.
97. J. Shen et al., "Natural tts synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing* (ICASSP), Calgary, Canada, 2018, pp. 4779–4783; doi: 10.1109/ICASSP.2018.8461368
98. Y. Yasuda, X. Wang, S. Takaki, and J. Yamagishi, "Investigation of enhanced Tacotron text-tospeech synthesis systems with self-attention for pitch accent language," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing* (ICASSP), Brighton, United Kingdom, 2019, pp. 6905–6909.
99. M. Schröder, "Emotional speech synthesis: A review," in *Proc. 7th European Conf. on Speech Communication and Technology*, Aalborg, Denmark, 2001, pp. 561–564.
100. D. Govind and S. R. M. Prasanna, "Expressive speech synthesis: a review," *International Journal of Speech Technology*, vol. 16, no. 2, pp. 237–260, 2013
101. J. E. Cahn, "The generation of affect in synthesized speech," *Journal of the American Voice I/O Society*, vol. 8. vo. 1. pp. 1–19, 1989.
102. F. Burkhardt, W. F. Sendlmeier, "Verification of acoustical correlates of emotional speech using formant-synthesis," in *ISCA Tutorial and Research Workshop on speech and emotion*, Newcastle, Northern Ireland, UK, 2000, pp. 151–156.
103. M. Schröder, "Expressive speech synthesis: Past, present, and possible futures," in *Affective information processing*, J. Tao, T. Tan, Eds. London: Springer, 2009, pp. 111–126; doi: 10.1007/978-1-84800-306-4_7
104. M. Schröder, "Can emotions be synthesized without controlling voice quality," *Phonus*, vol. 4, pp. 35–50, 1999.
105. A. Iida, and N. Campbell, "Speech database design for a concatenative text-to-speech synthesis system for individuals with communication disorders," *International Journal of Speech Technology*, vol. 6, no. 4, pp. 379–392, 2003; doi: 10.1023/A:1025761017833
106. W. L. Johnson, S. S. Narayanan, R. Whitney, R. Das, M. Bulut, and C. LaBore, "Limited domain synthesis of expressive military speech for animated characters," in *Proc. of 2002 IEEE Workshop on Speech Synthesis*, Santa Monica, CA, 2002, pp. 163-166; doi: 10.1109/WSS.2002.1224399
107. J. F. Pitrelli, R. Bakis, E. M. Eide, R. Fernandez, W. Hamza, and M. A. Picheny, "The IBM expressive text-to-speech synthesis system for American English," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1099–1108, 2006; doi: 10.1109/TASL.2006.876123
108. W. Hamza, E. Eide, R. Bakis, M. Picheny, and J. Pitrelli, "The IBM expressive speech synthesis system," in *8th Int. Conf. on Spoken Language Processing* (INTERSPEECH 2004), Jeju Island, Korea, 2004, pp. 2577–2580.
109. Y. Zhao, D. Peng, L. Wang, M. Chu, Y. Chen, P. Yu, and J. Guo, "Constructing stylistic synthesis databases from audio books," in *9th Int. Conf. on Spoken Language Processing* (INTERSPEECH 2006), Pittsburgh, PA, 2006.
110. K. Prahallad, A. R. Toth, and A. W. Black, "Automatic building of synthetic voices from large multi-paragraph speech databases," in *Proc. 8th An. Conf. of the International Speech Communication Association*, Antwerp, Belgium, 2007.
111. N. Braunschweiler, M. J. F. Gales, and S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings," in *Proc. 11th An. Conf. of the International Speech Communication Association*, Makuhari, Japan, 2010.
112. F. Eyben , S. Buchholz, N. Braunschweiler et al., "Unsupervised clustering of emotion and voice styles for expressive TTS," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing* (ICASSP), Kyoto, Japan, 2012, pp. 4009–4012; doi: 10.1109/ICASSP.2012.6288797
113. E. Székely, J. P. Cabral, P. Cahill, and J. Carson-Berndsen, "Clustering expressive speech styles in audiobooks using glottal source parameters," in *12th Annu. Conf. of the Int. Speech Communication Association*, Florence, Italy, 2011, pp. 2409–2412.

*Received 15.01.2019, the final version — 21.02.2019.*