УДК 004.934

Н.Н. Сажок

Речевые информационные технологии и системы

Обобщены результаты многолетних исследований в области распознавания речи, описаны основные результаты теоретических исследований, примеры разработок, определивших ключевые направления практического использования речевых информационных технологий, а также знаковые тенденции и основные направления дальнейших исследований.

Ключевые слова: распознавание речевого сигнала, смысловая интерпретация речи, синтез речи по тексту, системы речевого диалога, генеративная модель.

Узагальнено результати багаторічних досліджень в області розпізнавання мовлення, описано основні результати теоретичних досліджень, приклади розробок, які визначили ключові напрями практичного використання мовленнєвих інформаційних технологій, а також знакові тенденції та основні напрями подальших досліджень.

Ключові слова розпізнавання мовленнєвого сигналу, смислова інтерпретація мовлення, синтез мовлення за текстом, системи усного діалогу, генеративна модель.

Введение. Речевые технологии и системы практически стали неотъемлемой частью современного информационного общества, во многом способствуя переходу к обществу знаний. Между тем, основы теории и ряд практических применений обработки речевого сигнала заложены в Украине в разные периоды развития научной и инженерной мысли. Генеративная модель, впервые предложенная для распознавания речи, способствовала становлению теории распознавания образов в целом [1]. Пройден долгий и нелегкий путь от построения экспериментальных систем речевого диалога на большой электронной счетной машине (БЭСМ) до портативных устройств речевого управления, от систем, ограниченных словарями в несколько десятков или сотен слов, до практически полного охвата лексикона.

Сегодня в Украине ведущую роль в развитии речевых технологий занимает Международный Центр, где исследования проблематики речевых информационных технологий и систем ведутся традиционно, их высокий уровень признается как на национальном, так и на международном уровнях. В статье обобщены результаты многолетних исследований в области распознавания речи, коротко описаны основные результаты теоретических исследований, приведены примеры разработок, определивших ключевые направления практического использования речевых информационных технологий, описаны некоторые тенденции и направления будущих исследований.

Сферы применения речевых технологий (актуальность)

С каждым новым продвижением в области речевых технологий возрастают ожидания пользователей, вызванные как опытом эксплуатации прототипов систем, так и расширением сфер применения, связанных с новыми источниками получения речевой информации и многоязычностью.

Ввод информации голосом позволяет освободить руки, а синтез речи по тексту – разгрузить глаза. Объединение этих двух технологий позволяет создать различные системы речевого диалога, в том числе системы голосового управления техническими системами.

Речевое общение человека и компьютера постепенно входит в нашу повседневную жизнь. Например, подсистема *Siri* в устройствах на базе *iOS* и подсистема *Cortana* в *MS Windows* и *Android* помогают находить всевозможную информацию как в Интернете, так и на самом устройстве. В основе моделирования такого общения лежат технологии анализа, распознавания, интерпретации и синтеза речевого сигнала в пределах предметных областей и устного диалога [1]. Развиваются сервисные службы, с использованием диалоговых систем, позволяющие получать информацию о расписании движения транспорта, текущее состояние банковского счета, бронировать билеты и т.д.

Ввод текстовой информации под диктовку предусматривает выход за пределы предметных областей, размеры рабочих словарей составляют сотни тысяч слов.

Еще одним применением речевых технологий есть обработка медийной информации, в частности теле- и радиовещания. Такого рода системы, в отличие от ранее рассмотренных задач, не предусматривают обратной связи с лицом, чья речь распознается. Здесь также нет жесткого требования реального времени. В мире существует ряд преимущественно экспериментальных и вспомогательных систем, в которых автоматизированы генерирование субтитров и поиск информации для английского и ряда других языков, в основном европейских [2, 3]. В основе таких систем лежит технология распознавания речи, предназначенная для преобразования в текст сигнала, который принимается из определенных источников вещания и соответствует определенному набору телерадиопрограмм (новости, интервью, телешоу, трансляции заседаний парламента и др.). Полученный в результате преобразования текст должен соответствовать содержанию, а пользователь системы должен иметь возможность прослушивать запись передачи, параллельно следя за текстом и, при необходимости, корректируя его. При этом важно сократить задержку получения ответа распознавания, одновременно учитывая ограничения доступных вычислительных ресурсов.

Генеративная модель распознавания речи

За последние 20 лет в составе Международного научно-учебного центра информационных технологий и систем получила дальнейшее развитие ИКДП-теория (подход, нынешнее название – Генеративная модель) автоматического распознавания, понимания, синтеза и компрессированной передачи речевых сигналов.

Эта теория, предложенная в 1966 году, основывается на экономном описании (задании, композиции (К)) модельных речевых сигналов с помощью иерархически (И) организованных стохастических автоматных порождающих грамматик и на сравнении модельных сигналов с теми, которые распознаются с помощью динамического программирования (ДП) [1]. Эти теоретические разработки Центра признаны пионерными в мире и такими, которые повлияли решающим образом на развитие речевой информатики [4].

С 1975 года модификация этой теории известна под названием Скрытые Марковские Модели (*HMM – Hidden Markov Model*), а в начале 2000-х теория *HMM* была эффективно переформулирована в терминах конечных взвешенных трансдьюсеров [5].

В задачах преобразования речи в текст входной речевой сигнал преобразуется в последовательность акустических векторов фиксированного измерения $\mathbf{Y}_{1:T} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T)$ в результате препроцессинга. Иными словами, происходит переход в пространство первичных признаков. Затем декодер пытается найти последовательность слов $\boldsymbol{w}_{1:T} = (w_1, w_2, \ldots, w_L)$, которая наиболее вероятно соответствует наблюдаемому **Y**, т.е. декодер пытается найти

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}}\, P(\mathbf{w}|\mathbf{Y}). \qquad (1)$$

Несмотря на сложность, ряд дискриминантных моделей пытаются оперировать с этим выражением напрямую [6]. Впрочем, наиболее продуктивна – генеративная модель, рассматривающая эквивалентную задачу, возникающую в результате применения правила Байеса к (1):

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}}\, p(\mathbf{Y}|\mathbf{w})P(\mathbf{w}). \qquad (2)$$

Мера схожести $p(\mathbf{Y}|\mathbf{w})$ определяет акустическую составляющую, а вероятность $P(\mathbf{w})$ – лингвистическую составляющую генеративной модели распознавания речевого сигнала. Параметры модели оцениваются на основе речевых и текстовых корпусов [7].

Рассмотрим подробнее акустическую составляющую или акустическую модель (АМ). Каждое произнесенное слово *w* разлагается на последовательность L_w базовых звуков, т.е. фонем из некоторого алфавита базовых фонем. Эта последовательность есть произношением слова или его фонетической транскрипцией. Чтобы учесть множественность вариантов произношения слова, степень сходства вычисляется по многим допустимым фонемным транскрипциям:

$$p(\mathbf{Y}|\mathbf{w}) = \sum_{\mathbf{Q}} p(\mathbf{Y}|\mathbf{Q})P(\mathbf{Q}|\mathbf{w}), \qquad (3)$$

где сумма, обычно замещаемая максимумом, берется по всем допустимым последовательностям произношения для **w**, **Q** – некоторая последовательность фонемных транскрипций,

$$P(\mathrm{Q}|\mathrm{w}) = \prod_{l=1}^{L} P(\mathbf{q}^{(w_l)}|w_l), \qquad (4)$$

где $\mathbf{q}^{(w_l)}$ – допустимое произношение слова w_l.

Чтобы построить фонемные транскрипции слов, разрабатывается процедура графемно-фонемного преобразования. Это преобразование моделирует особенности, присущие конкретному языку. Между символами из алфавита языка и звуками из базового алфавита фонем должна моделироваться связь. Ряд языков при этом имеют свои характерные признаки. Так, в отличие от английского, для украинского языка в алфавит фонем включены как ударные, так и безударные гласные. Из этого следует, что необходимо предусмотреть также прогнозирование ударения в словах. Кроме того, для всех языков возникает необходимость расшифровки цифр и символов. Эти аспекты проработаны в рамках многозначной многоуровневой генеративной модели [8].

Каждая фонема q представляется в виде порождающей модели, как проиллюстрировано на рис. 1, *а*, где $\{a_{ij}\}$ – статистические параметры перехода между состояниями i и j, $\{b_j(\mathbf{y}_t)\}$ – распределения в пространстве первичных признаков для рабочих состояний. Эти распределения фактически аппроксимируют в пространстве первичных признаков те области, через которые проходят траектории, соответствующие фонеме q. Такой общий вид имеет базовая неявная (или скрытая) марковская модель *HMM*.

Технически, переход от рабочего состояния генеративной модели к одному из состояний, с которым рабочее состояние связано, осуществляется за единицу отсчета времени.

Допустимая последовательность состояний

$$\Theta_{1:T} = (\theta_1, \theta_2, \ldots, \theta_T), \qquad (5)$$

по которой генерируется эталонный (модельный) сигнал, является некоторой акустической транскрипцией наблюдаемого сигнала. Согласно генеративной модели, эти состояния связаны условными зависимостями как между собой, так и с отсчетами наблюдаемого сигнала. На рис. 1, *б* эти зависимости для базовой *HMM* представлены в виде динамической байесовской сети (ДБС) [7]. В принятой здесь нотации дискретные переменные изображаются в квадратах, непрерывные переменные – в кругах, наблюдаемые переменные затенены, а скрытые оставлены светлыми.

Этот вид удобный для иллюстрации расширений базовой генеративной модели, в частности, для введения дополнительных параметров и зависимостей, например, между соседними отсчетами наблюдаемого сигнала. Кроме того, ДБС довольно иллюстративна при объяснении дискриминантных моделей.

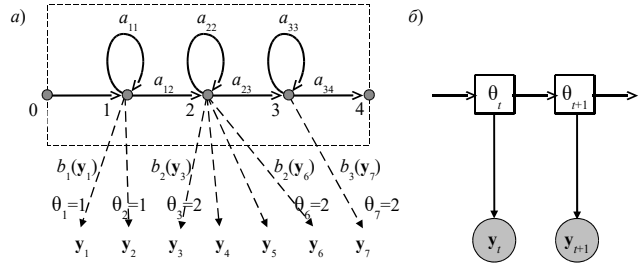

Рис. 1. Базовая генеративная модель *HMM* фонемы: *а* – в виде свернутого графа динамического программирования; *б* – в виде динамической байесовской сети

Недавнее переформулирование *HMM* в терминах конечных взвешенных трансдьюсеров и разработка соответствующих инструментальных средств [5, 9] позволили применить математический аппарат трандьюсеров для оптимизации построения графа ДП. При этом снимается проблема масштабирования акустической и лингвистической составляющих, и появляется возможность более прозрачного описания и обоснования конструкций при построении генеративных моделей.

Моделирование в пространстве первичных признаков

Модель аппроксимации областей наблюдения фонем в пространстве первичных признаков сигнала на основе нормального закона распределения демонстрировала эффективность на протяжении многих десятилетий. Для лучшего качества аппроксимации в пространстве первичных признаков сигнала областей пребывания фонемы вместо одного нормального закона (гауссоида) $G(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ вводится смесь гауссоидов:

$$b_j(\mathbf{y}) = \sum_{m=1}^{M} \gamma_{jm} G\left(\mathbf{y}; \boldsymbol{\mu}^{(jm)}, \boldsymbol{\Sigma}^{(jm)}\right), \qquad (6)$$

где γ_{jm} – априорная вероятность пребывания в m-м гауссоиде j-го состояния, удовлетворяю-

щая условиям функции вероятности, в частности, $\gamma_{jm} \geq 0$ и $\sum_{m=1}^{M} \gamma_{jm} = 1$.

Смесью нормальных законов моделируются асимметричные распределения и распределения со многими модами. Это позволяет более точно отразить многообразие сигнала, обусловленое индивидуальными особенностями дикторов (диалект, пол, тембр) и взаимовлиянием звуков в потоке речи или коартикуляцией. На самом деле, более стандартным приемом учета коартикуляции есть введение контекстной зависимости фонем, например, использование фонемно-трифонной модели [10]. В более общем случае, $\{ b_j(\mathbf{y}_t) \}$ принимают вид

$$p(s_i, c_j|\mathbf{y}_t) = p(s_i|\mathbf{y}_t)p(c_j|s_i, \mathbf{y}_t), \qquad (7)$$

где c_j – один из кластеризированных контекстных классов $C = \{c_1, \ldots, c_J\}$, а s_i – контекстно-независимая фонема или состояние в некоторой контекстно-независимой фонеме. К контекстным классам могут относится не только прилегающие справа или слева фонемы, но и прочие лингвистические признаки, относящиеся, например, к границам слов или морфем, движение интонации и т.д.

Особенно важным при применении смеси нормальных законов есть обоснованное обеспечение диагональности каждой ковариационной матрицы. Для этого, при необходимости, проводится декореляция пространства первичных признаков путем применения дискретного косинус-преобразования. Таким образом, аппроксимация областей пребывания фонем осуществляется объединением эллипсоидов, вытянутых вдоль осей координат.

На рис. 2 изображена проекция на двумерное пространство траектории движения реализации слова *оса* в многомерном пространстве первичных признаков. Отсчеты наблюдаемого сигнала $\mathbf{y}_{t=1:72}$ проходят через области пребывания соответствующих фонем: # (фонема–пауза), *о*, *с*, *А* (а отмечена), #. Фонема–пауза # аппроксимируется эллипсоидом, что соответствует одному гауссоиду в едином состоянии модели этой фонемы # 1. Предполагается, что вероятность аппроксимации гауссоидом некоторой точки внутри соответствующего эллипсоида больше некоторого значения, например, 0,1. Модели фонем *о* и *А* содержат по три состояния: *О*1, *О*2, *О*3 и *А*1, *А*2, *А*3, распределение каждого из них аппроксимируется двумя компонентами смеси нормальных законов. Гауссоиды, соответствующие одному и тому же состоянию в пределах фонемы, имеют одинаковое штрихование. Модель фонемы *с* содержит также три состояния, но для аппроксимации распределения каждого из состояний используется только один гауссоид.

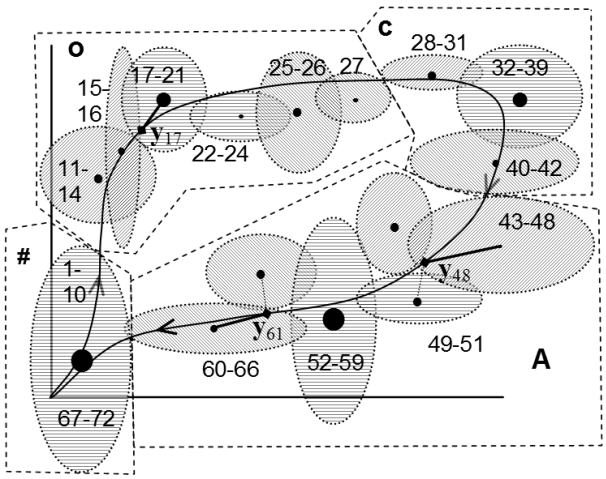

Рис. 2. Проекция на двумерное пространство траектории движения акустической реализации слова *оса́* в пространстве первичных признаков

В процессе распознавания всех допустимых акустических транскрипций ищется такая, что наилучшим образом аппроксимирует траекторию сигнала. Изображенная на рис. 2 траектория сигнала лучше аппроксимируется акустической транскрипцией вида (5), переменные которой имеют значения: $\theta_{1:10} = $ #1, $\theta_{11:16} = $ о1, $\theta_{17:24} = $ о2, $\theta_{25:27} = $ о3, $\theta_{28:31} = $ с1, $\theta_{32:39} = $ с2, $\theta_{40:42} = $ = с3, $\theta_{43:48} = $ А1, $\theta_{49:59} = $ А2, $\theta_{60:66} = $ А3, и $\theta_{67:72} = $ #1.

Другим способом оценивать априорные вероятности вида (7) есть применение многослойного перцептрона в рамках подхода, известного как *Deep Learning* [11]. Хотя такой метод и не дает прямой возможности проводить процедуры адаптации на голос диктора, он имеет ряд теоретических преимуществ (например, не подвержен локальности), а системы с его использованием демонстрируют заметное улучшение надежности.

Многоязычие

Речевые сигналы в разных языках могут быть представлены последовательностями символов

из некоторого универсального (международного) фонетического алфавита. Действительно, независимо от языка, все люди имеют одинаковую анатомию и физиологию для генерирования и восприятия речи. Значит, может быть предложена единая артикуляторно-фонемная модель, позволяющая описывать речевые сигналы с помощью фонетико-акустических символов из некоторого общего алфавита. Хотя языки разные, однако артикуляционные жесты, характерные для одного языка, выбираются из некоторого общего для всех языков алфавита. Так, все языки имеют гласные и согласные, ударные и безударные звуки, звонкие и шипящие согласные и др.

Более того, все люди говорят об одних и тех же событиях в одном и том же мире, выражая сходные эмоции. Вот почему мы выдвигаем гипотезу о создании единого фонетического кода из около 100 базовых фонем (артикуляторных движений) и предлагаем использовать этот код для описания речевых сигналов для различных языков и народов. В каждом языке в среднем около 50 фонем – артикуляционных жестов, и поэтому можно использовать около C_{100}^{50} (гигантское количество!) различных языков. Следовательно, при обработке многоязыковых речевых сигналов должен быть создан фонетико-акустический транскриптор, превращающий речевой сигнал в последовательность акустико-фонетических кросс-лингвистических символов. В основу этого объединенного аудиопроекта положен межязыковый фонетический код.

Таким образом, сначала речевой сигнал описывается последовательностью фонетических симолов в международном фонетическом коде, а затем полученная последовательность превращается в последовтельность символов в национальном фонетическом коде. В дальнейшем обработка последовательности символов ведется с учетом лексики, синтаксиса и другой априорной информации в терминах национального языка.

Общая структура системы распознавания речи

Общая структура системы преобразования текста в речь, показанная на рис. 3, разделяется на компоненты, одна из которых, компонента реального времени (он-лайновая) – *Распознаватель,* – обращается к *Базе данных и знаний*, формируемой в отложенном режиме (офф-лайн).

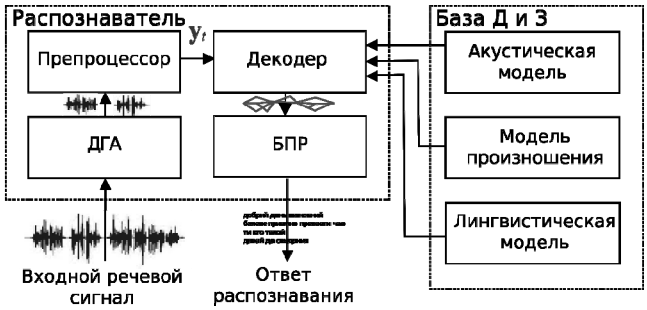

Рис. 3. Общая структура базовой системы преобразования текста в речь

Компонента реального времени *Распознаватель* получает *Входящий речевой сигнал* из некоторого источника (микрофон, файловая система или компьютерная сеть). *Детектор голосовой активности* (ДГА) обнаруживает предполагаемые начала речевых сегментов, чтобы начать передачу сигнала в *Препроцессор*, который извлекает первичные акустические признаки. *Декодер* сравнивает входящий сегмент с гипотезами модельного сигнала, которые генерируются на основе акустической и лингвистической моделей, отбрасывая неперспективные гипотезы. Результат декодирования, представленный в виде последовательностей слов или сети несовпадений, дополненных оценками длительностей и доверительной мерой, передается в *Блок принятия решений* (БПР), который формирует окончательный *Ответ распознавания* с возможным учетом предыстории и доверительных интервалов.

Прикладные разработки

Ранние прикладные разработки на основе речевых технологий выполнены на базе как наиболее мощных в свое время компьютеров, например, БЭСМ, так и с применением специализированных процессорных систем в 1970–1990 гг. Они характеризировались необходимостью настройки на голос пользователя с произнесением последним всего словаря. Наиболее успешный пример ранних прикладных разработок – система речевого диалога «Речь–121» (рис. 4), созданная в 1986–1991 гг. по контрактам с ЮНЕСКО. Это многоязычная система устного диалога для ПК (использовалось семь европейских языков).

Рис. 4. Система речевого диалога «Речь–121» (1986)

В рамках Государственной научно-технической программы «Образный компьютер» (2001–2011 гг.) были разработаны первые портативные устройства речевой информатики. Цифровой диктофон с голосовым управлением «Вокофон» (рис. 5) обеспечивал голосовые функции записи и воспроизведения аудиоинформации, именования и разметки записанной информации, поиск информации по ключевым словам, произнесенных пользователем. Были созданы макетные устройства на основе микропроцессора цифровой обработки сигналов *ADSP*–2188*N* и *BF*–561. Разработанный на той же электронной базе портативный устный словарь–переводчик «Вокопретер» выполнял устный перевод в пределах выбранной предметной области. После произнесения пользователем на родном языке слова или фразы результат автоматического распознавания и перевода озвучивался на иностранном языке. Предварительно от пользователя требовалось проговорить все фразы на родном языке для настраивания параметров системы распознавания.

На основании теоретических исследований разработан ряд интеллектуальных речевых информационных технологий и экспериментальных компьютерных систем, ориентированных на различные сферы применения. Среди них особое место занимает автоматический фонетический стенограф, предназначенный для преобразования речи в последовательность фонем без учета знаний о лексике, синтаксисе и семантике. Являясь реализацией базовой технологии ввода информации голосом, фонетический стенограф постоянно совершенствуется, начиная с 2004 года.

Рис. 5. Модификации цифрового диктофона с голосовым управлением (2001–2002 гг.)

Пример диктовальной машины – система «Диригент» [12] (первая версия – в 2012 г.). Сегодня она покрывает до 95 процентоя лексики трех языков и позволят вводить текст под диктовку на ноутбуке. Еще одна система, которая непосредственно кооперируется с диктором, – *vOKopreter* – осуществляет устный перевод в пределах предметных областей без необходимости предварительной настройки на голос диктора (рис. 6). В основе системы лежит генеративная модель смысловой интерпретации речевого сигнала [13]. Дальнейшее развитие этой системы предполагает построение модели общения пользователя с кибернетическими системами естественной речью [14, 15].

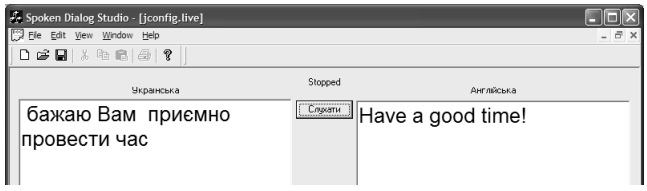

Рис. 6. Результат смыслового перевода фразы, произнесение которой демонстрируется в различных вариантах (например, переставляются слова), но с сохранением единого смысла

Для обработки медийной информации, в частности теле- и радиовещания, в мире существует ряд преимущественно экспериментальных и вспомогательных систем, в которых автоматизированы генерирование субтитров и поиск информации для английского и ряда других языков, в основном европейских [2, 3]. Разработанная в Украине система *WebSten* обеспечивает преобразование в текст спонтанной речи, полученной по каналам телерадиовещания без

предусмотрения взаимодействия с диктором, а система *MediaAudit* проводит мониторинг телерадиоэфира по ключевым словам (рис. 7). Демонстрируется высокая релевантность найденных сюжетов. Поддерживается использование шаблонов при поиске. Создан веб-сервис с возможностью редактирования распознанного текста. В обеих системах пользователь имеет возможность прослушивать отобранные или найденные медийные фрагменты синхронно с распознанным текстом (рис. 7) [15, 16].

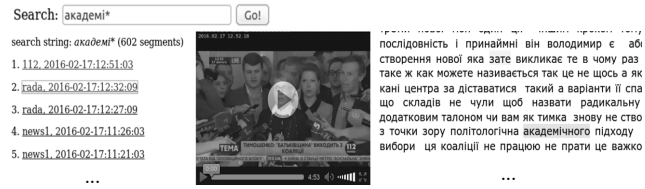

Рис. 7. Результат поиска по семейству ключевых слов, начинающихся на «академі»

Пример адаптации речевых технологий для специфической сферы деятельности человека – средства автоматизации составления судебных протоколов в коммерческой системе *SRS* «*Femida*». На систему распознавания речи возложены: определение стадий судебного разбирательства, поиск ключевых слов в фонограммах и формирование протокола под диктовку. В рамках этих разработок был создан первый украинский корпус эфирной речи [17].

Технология обратная распознаванию речи – синтез речи по тексту – реализована в системе «Текстофон». Она обеспечивает озвучивание произвольных украиноязычных текстов. Доступны мужской и женский голоса, возможно регулирование скорости воспроизведения речи. Высокая натуральность и разборчивость достигаются путем обработки больших объемов записанной речи. Устанавливается на ПК или сервер [18].

Перспективы развития

Для повышения уровня жизни и ее безопасности необходимы новые компьютерные средства (интеллектуальные помощники человека) для анализа, распознавания, понимания, кодирования и генерирования речевой и звуковой информации. Эти средства должны не только повысить надежность распознавания и понимания речевых сигналов, повысить натуральность и разборчивость синтезированной речи, а и обеспечить такие же показатели для беглой речи, в том числе в условиях помех и для шепотной речи. Отдельную группу задач составляют задачи распознавания индивидуальной речи человека, идентификации и верификации личности по голосу говорящего, улучшение качества речи в условиях помех, разделение смесей и сумм сигналов по источнику их происхождения.

Заключение. За последние десятилетия в речевых технологиях осуществлен значительный прогресс, и не в последнюю очередь благодаря увеличению мощности вычислительных ресурсов. Достигнутое повышение робастности для ряда задач в распознавании значительно приблизило уровень восприятия речи компьютером к человеческому.

В общих теоретических подходах наблюдается определенная сбалансированность генеративных и дискриминативных моделей. Трансдьюсерное представление позволило обобщить методы комбинирования генеративных моделей и их оптимизации, что привело к более простым и гибким конструкциям систем распознавания.

Моделирование разного рода контекстных зависимостей и более качественная аппроксимация в математическом описании пространства первичных акустических признаков дали ощутимый эффект, продолжение работ в этих направлениях имеет значительную перспективу. Предстоит раскрыть потенциал моделирования структуры лингвистической составляющей генеративной модели, особенно для спонтанной речи и многоязычия.

1. *Винцюк Т.К.* Анализ, распознавание и смысловая интерпретация речевых сигналов. Киев: Наук. думка, 1987. – 264 с.
2. *http://voxalead.labs.exalead.com/*
3. *http://tech.ebu.ch/docs/events/metadata*15/*Petr* Vitek and Pavel Ircing_CT_UWB.pdf
4. *Sadaoki Furui*. 50 years of progress in speech and speaker recognition / Proc. of 10th Int. Conf. «Speech and Computer». – Patras, Greece, 2005. – P. 1–9.
5. *Mohri M.*, *Pereira F.*, *Riley M.* Speech recognition with weighted finite-state transducers. Springer Handbook on Speech Processing and Speech Communication. – Berlin, Heidelberg: Springer, 2008. – P. 559–584.
6. *Gales M.* Discriminative models for speech recognition, ITA Work-shop, Univ. San Diego, USA, Feb. 2007.

7. *Gales M.*, *Young S.* The Application of Hidden Markov Models in Speech Recognition // Foundations and Trends in Signal Processing. – 2007. – 1(3). – P. 195–304.
8. *Робейко В.В.*, *Сажок М.М.* Багатозначна багаторівнева модель перетворення орфографічного тексту на фонемний // Штучний інтелект. – 2011. – № 4. – С. 117–125.
9. *The Kaldi* Speech Recognition Toolkit / D. Povey, A. Ghoshal, G. Boulianne et. al. // IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.
10. *Vintsiuk Taras*, *Sazhok Mykola*. Multi-Level Multi-Decision Models for ASR // Proc. of the 10th Int. Conf. on Speech and Computer – SpeCom'2005, Patras, Greece, 17–19 Oct. 2005. – P. 69–76.
11. *Context*-Dependent Pre-Trained Deep Neural Networks for Large Vocabulary Speech Recognition / G. Dahl, Yu. Dong, Li. Deng et al. // IEEE Trans. Speech and Audio Proc., Special Issue on Deep Learning for Speech Processing, 2011.
12. *Робейко В.*, *Сажок М.* Розпізнавання спонтанного мовлення на основі акустичних композитних моделей слів у реальному часі // Штучний інтелект. – 2012. –№ 4. – С. 253–263.
13. *Сажок М.*, *Яценко В.* Система усного перекладу на основі інтерпретації мовленнєвого сигналу в межах предметних областей // Пр. Міжнар. конф. УкрОбраз'2010, 25–29 жовт. 2010. – К.: – С. 103–106.
14. *Построение* сценариев формализованного устного диалога на примере заказа билетов на железнодорожные поезда / В.В. Пилипенко, С.А. Биднюк, Р.А. Селюх и др. // УСиМ. – 2013. – № 4. – С. 71–75.
15. *Васильева Н.Б.*, *Сухоручкина О.Н.*, *Яценко В.В.* Особенности построения модели речевого общения пользователя с многофункциональным сервисным мобильным роботом // Там же. – 2015. – № 6. – С. 16–22, 28.
16. *Система* преобразования телерадиовещания в текст для украинского языка / Н.Н. Сажок, В.В. Робейко, Д.Я. Федорин и др. // Там же. – С. 66–73.
17. *Створення* акустичного корпусу українського ефірного мовлення / Н.Б. Васильєва, В.В. Пилипенко, О.М. Радуцький та ін. // Обробка сигналів і зображень та розпізнавання образів: X Міжнар. конф. УкрОбраз'2010, 25–29 жовт. 2010. – К.: – С. 55–58.
18. *Автоматичний* озвучувач українських текстів на основі фонемно-трифонної моделі з використанням природного мовного сигналу / Т.К.Вінцюк, Т.В. Людовик, М.М. Сажок та ін. / Оброблення сигналів і зображень та розпізнавання образів: Пр. 6-ї Всеукраїнської міжнар. конф. УкрОбраз'2002, Київ, 8–12 жовт., 2002 р. – К.: УАсОІРО, 2002. – С. 79–84.

Поступила 30.04.2017
E-mail: *sazhok@gmail.com*

UDC 004.934

N. N. Sazhok

Speech Information Technologies and Systems

Keywords: speech signal recognition, speech understanding, text-to-speech, spoken dialog systems, generative model.

Introduction. Speech technology and systems became the part of the contemporary world, which helps to transform the society. The Generative Model proposed in Ukraine in 1960s became a base of modern and most productive techniques for speech recognition as well as for pattern recognition in general.

The **purpose** is to analyze and generalize the theoretical and applied progress in order to characterize the state-of-the-art, shape trends and suggest further research and development.

Scope of application. Speech technology scope extends continuously by introducing new speech signal sources and multilingualism, growing user expectations and IT progress. Voice input and text-to-speech systems allow relieving human hands and eyes that leads to the natural man-machine communication where a user can co-operate with the cybernetic system. An opposite example, where no such co-operation provided, is the automatic broadcast monitoring system.

Methods. The contemporary formulation of Generative Model is presented with the focus on its acoustic component. The recently adapted mathematical tools, which allow the context dependency and pattern hierarchy effective modelling, are referred. For decades, the feature space areas, where a basic speech segment is observed, have been successfully approximated by Gaussian Mixture Model. The further applied Deep Learning technique improves the approximation quality. The way to model the multilingual aspects are described. The general model of the speech recognition system is presented and the key applications are described.

Conclusion. The gap between computer and human speech processing has been significantly reduced for the certain tasks. The extended (e.g. structural) context dependency and feature space modelling showed their effectiveness and is promising for the further research.

●