УДК 004.934

А.А. Тарасов, В.Н. Костин

О ПРОБЛЕМЕ РАЗРАБОТКИ СИСТЕМЫ ОБРАБОТКИ РЕЧИ В РЕЖИМЕ РЕАЛЬНОГО ВРЕМЕНИ

Описаны методы разработки системы распознавания речи, ее перевода на иностранный язык, а также синтеза речи. Рассмотрены следующие методы проектирования: создание моделей существующих решений и построение на этой основе модели будущего решения (IDEF0 модели AS-IS и TO-BE), построение диаграммы потоков данных (DFD-модель), построение модели сущность-связь (ERD-модель). При описании проблемы разработки системы решался ряд нетривиальных проблем: считывание речи, ее анализ, перевод на иностранный язык, симуляция голоса говорящего. В качестве дополнительного программного обеспечения предлагается использовать интерфейс прикладного программирования "Google Translate API", который обеспечивает реализацию считывания речи и ее перевода на иностранный язык. Для реализации анализа речи применяются методы, которые используются в клонировании голоса: спектральный анализ речевого сигнала на базе быстрого преобразования Фурье (FFT-анализ), математический аппарат динамического программирования (ДП-метод), алгоритм разметки питчей речевого сигнала.

Ключевые слова: проблема международной коммуникации, система обработки речи, автоматическая обработка речи, распознавание речи, анализ голоса, перевод речи, синтез речи, клонирование голоса, клонирование речи.

Введение

В современном обществе каждый человек стремится к общению. Люди путешествуют, изучают разные страны, знакомятся с местными жителями. Также для общения люди используют интернет или социальные сети. В этом случае собеседники общаются между собой виртуально. Однако возникают ситуации, когда человек хочет общаться с людьми на иностранном языке, но он знает его на минимальном уровне, либо им не владеет [1]. Из этого следует, что есть необходимость в разработке системы, позволяющей решить подобную проблему.

В соответствии с данными исследования Фонда «Общественное мнение», проведенного 8–9 июня 2013 г., 57% опрошенных

ISSN 0236-1493. Горный информационно-аналитический бюллетень. 2017. № 2. С. 209–222.

Таблица 1

Владеете ли вы, хотя бы на минимальном уровне, каким-либо иностранным языком

Доли групп	Население в целом, %
Нет, не владею	43
Английским	38
Французским	3
Немецким	19
Испанским	1
Другим языком	6
Затрудняюсь ответить	0

респондентов владеют хотя бы одним иностранным языком (табл. 1) [2].

При этом около 33% опрошенных респондентов признались, что владеет вторым языком на низшем (начальном) уровне (табл. 2) [2].

Приведенные выше данные позволяют сделать вывод о том, что проблема международного общения в России продолжает быть актуальной.

Для обоснования актуальности синтеза речи необходимо узнать, как именно человек воспринимает информацию. Исследования, проведенные британскими учеными, показывают, что люди усваивают 55% разговора с собеседником с применением языка телодвижений (невербальных средств передачи информации): поз, жестов и контактов глазами, 38% – с использованием тона голоса (паравербальных) и лишь 7% – содержания того, о чем говорится (вербальных) (рис. 1) [3].

Все то, что было описано выше, позволяет сделать вывод о том, что для реализации стопроцентного восприятия диалога с ино-

Таблица 2

На каком уровне вы владеете этим иностранным языком

Доли групп	Население в целом (из 57%), %
Начальный уровень	33
Средний уровень	15
Продвинутый уровень	3
Свободно владею	5
Затрудняюсь ответить	1

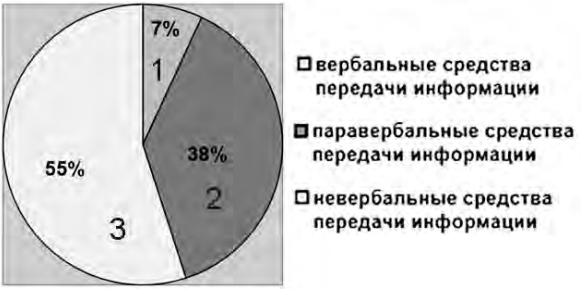

Рис. 1. Соотношения между различными средствами передачи информации

странцами, необходимо создать систему, которая предлагает возможность разговора с использованием видеотрансляции для обеспечения невербальной передачи информации, переводит исходную речь, а также синтезирует голос говорящего, произносящего полученную фразу на другом языке, что предоставит вербальное и паравербальное общение.

Общие положения

После определения актуальности разработки систем для обработки речи в режиме реального времени, необходимо описать бизнес-процессы, проходящие в сфере международного общения в настоящее время, с целью определить в них недостатки. Также необходимо описать бизнес-процессы, которые появятся в данной сфере после разработки подобной системы. Для этого были созданы модели IDEF0 AS-IS и TO-BE с точки зрения участника переговоров.

Рис. 2. Модель AS-IS. Блок A0

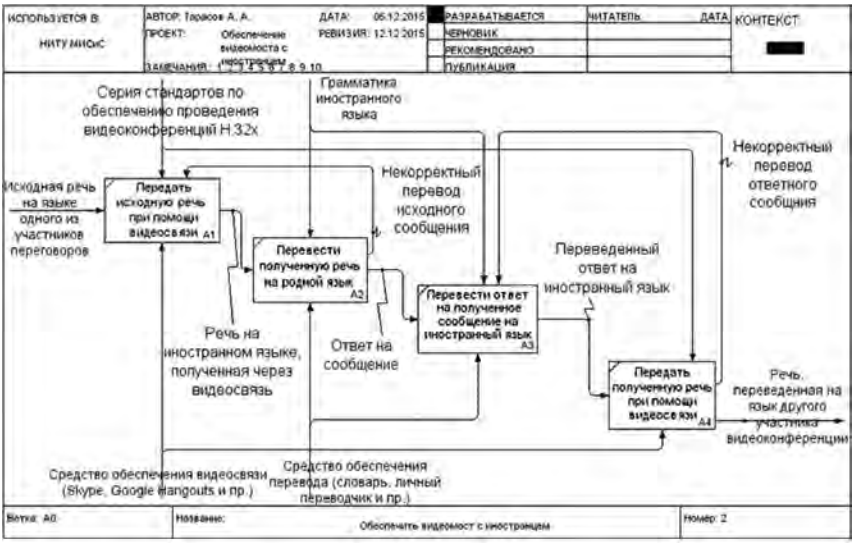

Рис. 3. Модель AS-IS. Декомпозиция блока A0

На схеме представлена модель AS-IS, целью которой является определение процессов, которые подлежат переопределению (рис. 2, 3). Она показывает протекающие бизнес-процессы, которые происходят при организации диалога с иностранцами в настоящее время.

Недостатками данной модели являются:

1. Необходимость наличия средств обеспечения перевода (словарь, личный переводчик и т. д.).

2. Воспроизведение голосовых сообщений на иностранном языке осуществляется либо с применением услуг личного переводчика, либо самим участником переговоров.

3. При использовании личного переводчика требуется оплата его услуг.

4. Затраты времени, связанные с процессом перевода.

Для того чтобы устранить подобные недостатки, необходимо разработать системы обработки речи, которая обеспечит автоматизацию перевода, а также реализует возможность озвучивания сообщений голосом говорящего. Модель TO-BE, которая представляет будущее решение с введенной в него системой, представлена на рис. 4, 5, 6. В качестве примера средства обеспечения связи был взят «Skype».

Следует отметить, что на рис. 6 блок A11 и последовательность A12-A13 должны выполняться одновременно.

Рис. 4. Модель TO-BE. Блока A0

Таким образом, подобная система устраняет недостатки, описанные в модели AS-IS, и является более выгодной для участников переговоров.

Для клонирования речи необходимо, прежде всего, создать базу данных звуковых волн аллофонов, опираясь на уже созданную эталонную базу данных аллофонов с голосом диктора.

Общая структурная схема автоматизации создания базы данных аллофонов с клонированным голосом представлена на

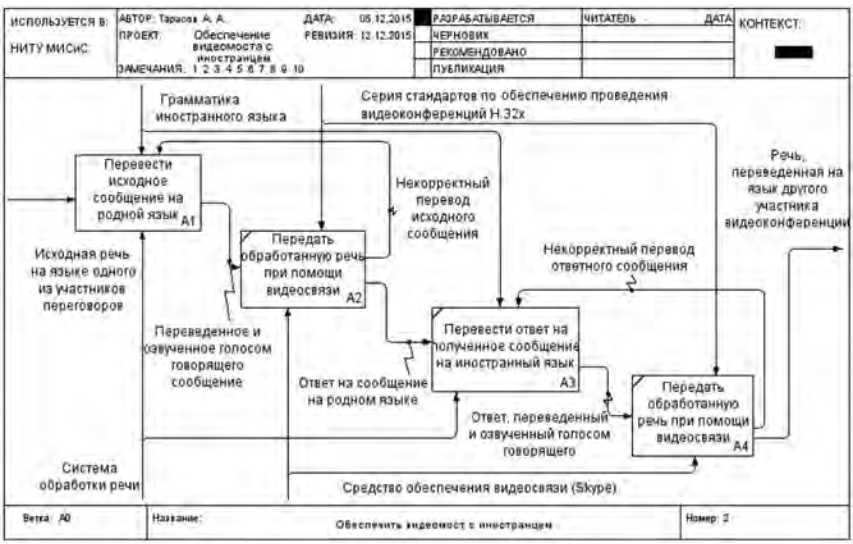

Рис. 5. Модель TO-BE. Декомпозиция блока A0

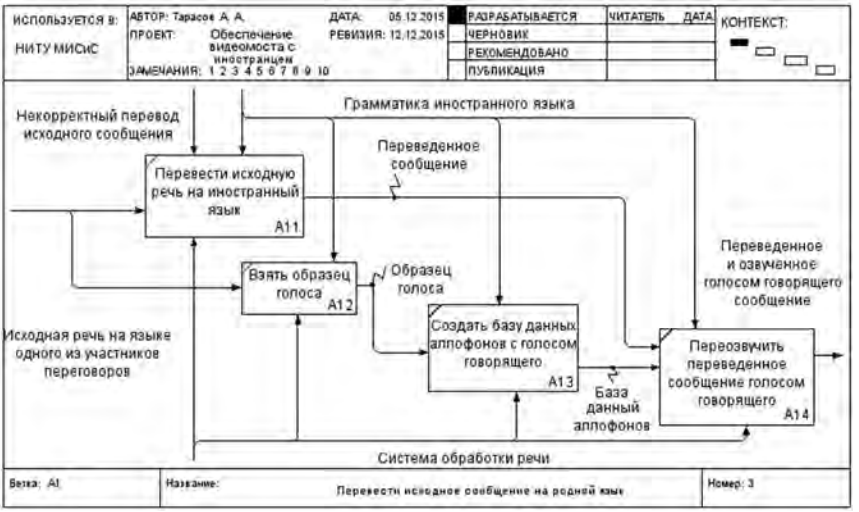

Рис. 6. Модель TO-BE. Декомпозиция блока A1

рис. 7. Далее будут рассмотрены основные этапы данного процесса.

Для анализа спектральных параметров используется спектральный анализ на базе быстрого преобразования Фурье (FFT-анализ). Быстрое преобразование Фурье – это математическая функция, позволяющая получить из временной зависимости сигнала его частотные компоненты, т.е. проводить спектральный анализ сигналов [5]:

$$x(n) = \sum_{k=0}^{N/2} X_k \cos \frac{2\pi k(n + \varphi_k)}{N},$$

где $x(n)$ – исходный дискретный сигнал, N – количество отсчетов дискретного сигнала $x(n)$ на анализируемом интервале, X_k – амплитуда k-ого колебания, φ_k – фаза k-ого колебания. Частоты этих синусоид равны $k*F/N$, где F – частота дискретизации сигнала [6]. В качестве частоты дискретизации сигнала удобно принять 8000 Гц, так как это значение используется в телефонной связи [7]. Набор коэффициентов X_k называется амплитудным спектром сигнала. Как видно из формулы, частоты синусоид, на которые раскладывается сигнал, равномерно распределены от 0 (постоянная составляющая) до $F/2$ – максимально возможной частоты в цифровом сигнале [6].

Основная идея автоматизации процессов сегментации и аллофонной маркировки заключается в реализации алгоритмов

Рис. 7. Схема автоматизации создания базы данных аллофонов с клонированным голосом

переноса меток начала и конца аллофонов с синтезированного сигнала, полученного с применением использования БД эталонных аллофонов, на речевой сигнал клонируемого пользователя.

Динамическое сопоставление (ДП-сопоставление) синтезированного и естественного сигналов осуществляется путем вычисления матрицы интегральных расстояний между векторами спектральных признаков сигналов по рекуррентной формуле:

$$D(n+1, m+1) = \max(D(n+1, m); D(n, m+1); D(n, m) + \partial_{\text{A}}(n, m)),$$

где D – матрица интегральных расстояний, n – отсчет меток синтезированного сигнала, m – отсчет меток речевого сигнала клонируемого пользователя, $\partial_{\text{A}}(n, m)$ – локальные расстояния между векторами спектральных признаков синтезированного и естественного сигналов.

Начальные условия для вычисления матрицы интегральных расстояний следующие: $D(n, 0) = 0$; $D(0, m) = 0$. Значения матрицы локальных расстояний $\partial_{\text{A}}(n, m)$ вычисляются по формуле:

$$\partial_A(n,m) = \partial\left(S(n), E(m)\right) = \frac{1}{K}\sum_{k=1}^{K}\left|S(n,k) - E(m,k)\right|,$$

где $S(n)$ – вектор спектральных признаков синтезированного сигнала в точке n; $E(m)$ – вектор спектральных признаков естественного сигнала в точке m; k – номер спектрального параметра; K – число спектральных параметров [4].

Питч – граница периодов каждого аллофона. Разметка питчей речевого сигнала (РС) осуществляется по следующему алгоритму. Для всего аллофонного сигнала по ординате минимума сдвиговой функции (рис. 8) определяется средний период основного тона T_O.

Находится начальная фаза – позиция питча в центре сигнала, от которого будут отсчитываться (влево и вправо) остальные питчи [8].

Для этого на середине сигнала берется окно размером в 3 периода T_O. Окно – весовая функция, которая используется для управления эффектами, обусловленными наличием растекания спектра [9]. В этом окне ищется участок с максимальным перепадом от положительной полуволны к отрицательной, т.е. такое место в окне, которое соответствует моменту времени закрытия голосовой щели и началу формантных колебаний.

Позицию питча определяет момент перехода через ноль от положительной полуволны к отрицательной.

1. Необходимо передвинуться вправо от центрального питча.
2. Берется окно размером в 3 периода T_O.
3. На нем определяется новый период основного тона T_O.
4. Новый период T_O ищется в диапазоне ±5% от T_O, полученного на предыдущей итерации.
5. T_O откладывается от предыдущего питча и осуществляется переход туда.
6. Затем ищется ближайший момент времени перехода через ноль от положительной полуволны к отрицательной.
7. Ставится питч, и повторяются шаги 3–8.
8. Когда достигается конец сигнала, осуществляется движение влево от центрального питча по тому же алгоритму.
9. Полученные метки питчей переносятся на исходный сигнал. Если от первого питча до начала сигнала $T_O/2 < t$, то этот участок сигнала выбрасывается. Та же процедура осуществляется для конца сигнала.

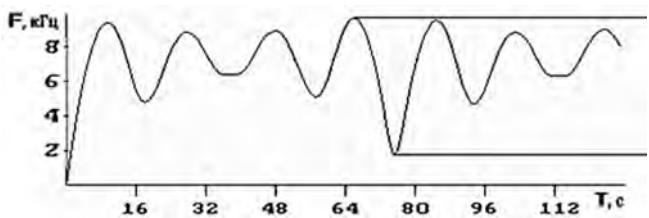

Рис. 8. Сдвиговая функция РС

10. Начало и конец сигнала сглаживаются путем добавления слева и справа по 32 отсчета и дополнения сигнала на этих отсчетах линейным участком от значения сигнала в начале (конце) сигнала до значения «0» [8].

Соответствие между синтезированным и естественным сигналами находится путем отображения наиболее эффективного пути на вычисленной матрице интегральных расстояний *D*. В результате получаются аллофонные сигналы с голосом клонируемого пользователя, которые необходимо обработать и преобразовать в базу данных.

На основе данной методики строится и процедура переозвучки речи. В качестве синтезируемого сигнала берутся аллофоны из созданной базы данных, а в качестве естественного источника – речь говорящего. Пример переноса меток границ аллофонов для слова «абракадабра» с синтезированного речевого сигнала (вертикальная ось) на естественный (горизонтальная ось) с использованием найденного пути соответствия показан на рис. 9 [4].

Далее необходимо определить, какие потоки данных протекают в системе и как требуется хранить в ней данные. Для этого были построены диаграммы потоков данных и хранения данных.

Для описания процессов, протекающих в системе, была спроектирована модель DFD (Data Flow Diagram), которая отображает внешние по отношению к системе источники и адресаты данных, потоки данных и хранилища данных, к которым осу-

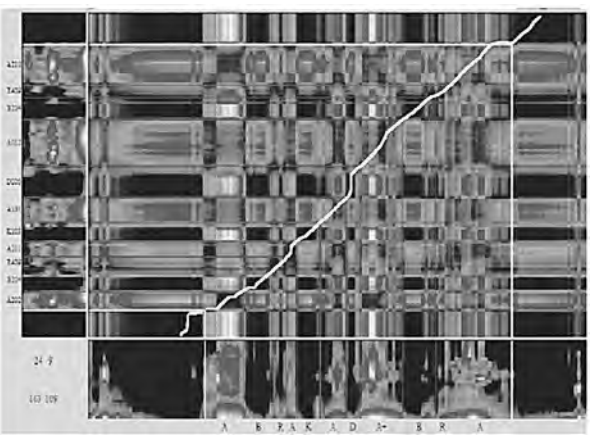

Рис. 9. Иллюстрация пути соответствия между синтезированным и естественным сигналами

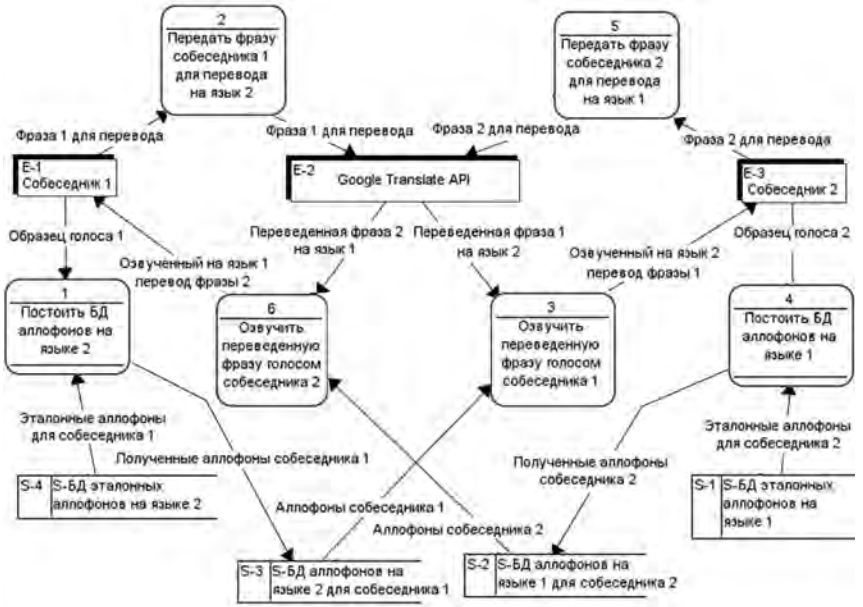

Рис. 10. Диаграмма потоков данных в нотации Гейна-Сарсона

ществляется доступ. Для упрощения модели на рисунке показана реализация диалога (количество участников равно 2) (рис. 10).

На основании данной диаграммы определяются основные процессы, на которые необходимо обратить особое внимание, такие как процессы 1, 3, 4 и 6 ввиду того, что они задействуют работу одновременно и с внешними сущностями, и с хранилищами, а также данные, которые с этими действиями связаны.

Для описания концептуальной схемы предметной области была спроектирована модель ERD (Entity-Relationship Diagram), которая позволяет выделить ключевые сущности и обозначить связи, которые могут устанавливаться между этими сущностями (рис. 11).

На основе данной концептуальной модели была также построена физическая модель баз данных для ее практической реализации (рис. 12).

Таблица «Собеседник» содержит в себе образец голоса для его передачи в таблицу «Аллофоны с голосом собеседника».

Таблица «Эталонные аллофоны» включает в себя звуковой файл эталонного аллофона, благодаря которым создается таблица «Аллофоны с голосом собеседника».

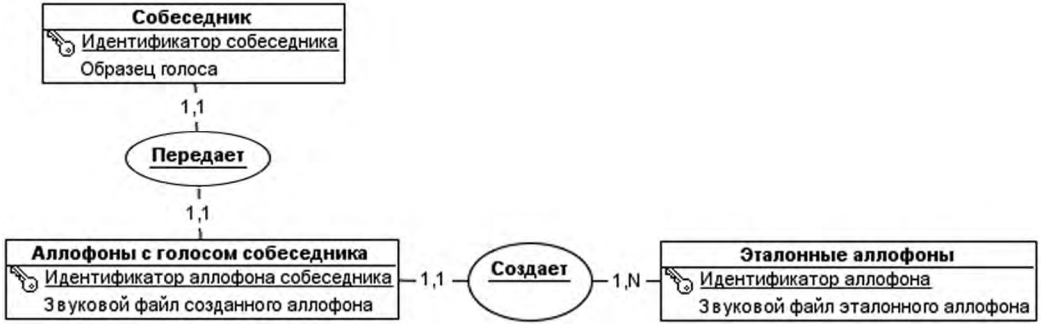

Рис. 11. Концептуальная модель

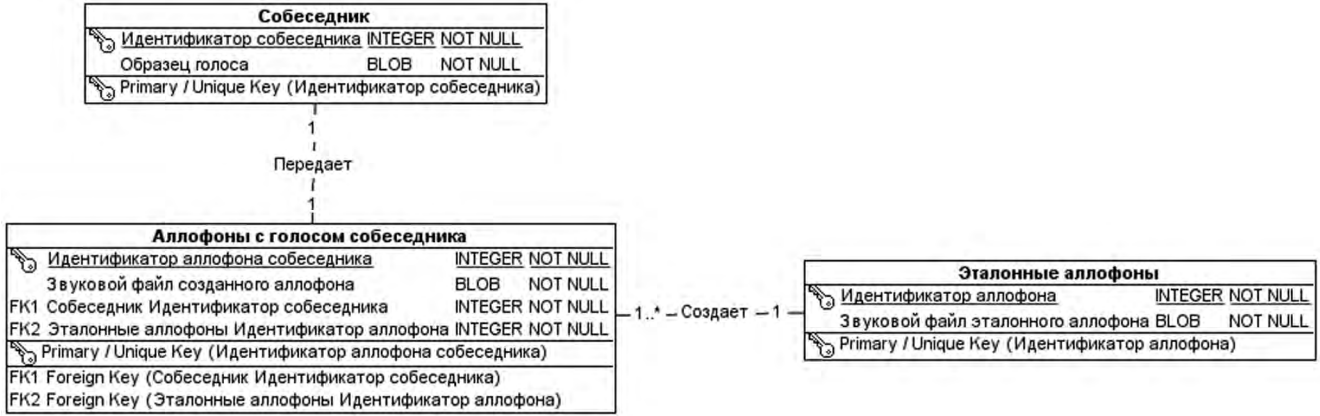

Рис. 12. Физическая модель

Таблица «Аллофоны с голосом собеседника» хранит звуковой файл созданного аллофона, с которым и работает приложение.

На основе данного проектирования появляется возможность сгенерировать СУБД на многих языках управления базами данных.

На сегодняшний день наиболее приближенный к реализации данного решения является Skype-переводчик – программа, разработанная компанией «Skype», обеспечивающая перевод голосовых звонков и видеозвонков в режиме реального времени. К сожалению, данная разработка обладает несколькими недостатками: отсутствие поддержки русского языка, что не позволяет ее использовать для жителей России и стран СНГ, а также переведенные фразы озвучиваются машинным голосом.

Исходя из полученных данных, выявляются следующие критерии к системам обработки речи:

- обеспечение считывания и обработка речи в режиме реального времени;
- синтез голоса говорящего для воспроизведения речи, переведенной на иностранный язык, на основе методов клонирования голоса;
- поддержка русского языка.

Заключение

В данной статье были проанализированы литературные источники по вопросу разработки систем обработки речи, а также проведена работа по проектированию подобной системы. Также в рамках данной статьи были определены требования к системе, а также было рассмотрено математическое обеспечение, на основе которого система должна быть реализована.

СПИСОК ЛИТЕРАТУРЫ

1. *Казанчева А.Ф.* Актуальность проблем межкультурной коммуникации в современном поликультурном пространстве. – Пятигорск: ПГЛУ, 2012. – 1 с.

2. *Владение* иностранными языками [Электр. ресурс] // Владение иностранными языками / ФОМ. – режим доступа: http://fom.ru/Nauka-i-obrazovanie/10998 (дата обращения: 13.11.2015).

3. *Mehrabian, Albert; Ferris, Susan R.* Inference of Attitudes from Nonverbal Communication in Two Channels. Journal of Consulting Psychology, 1967, 31 (3): 248–252.doi:10.1037/h0024648.

4. *Автоматическая* сегментация и маркировка речевого сигнала [Электр. ресурс] // БЛОГ Web Программиста. – режим доступа: http://juice-health.ru/archive/38-kompyuternyj-sintez-i-klonirovanie-rechi/184-avtomaticheskaya-segmentatsiya (дата обращения: 25.03.2016).

5. *БПФ* (Быстрое преобразование Фурье) [Электр. ресурс] // Контрольно-измерительные приборы и системы. – режим доступа http://www.kipis.ru/info/index.php?ELEMENT_ID=40417 (дата обращения: 27.03.2016).

6. *Спектроанализатор* – мы на нем видим? [Электр. ресурс] // ProSound.iXBT.com. – режим доступа: http://prosound.ixbt.com/education/spektr-analys.shtml (дата обращения: 27.03.2016).

7. *Частота* дискретизации [Электр. ресурс] // Статья из Википедии – Свободной энциклопедии. – режим доступа https://ru.wikipedia.org/wiki/%D0%A7%D0%B0%D1%81%D1%82%D0%BE%D1%82%D0%B0_%D0%B4%D0%B8%D1%81%D0%BA%D1%80%D0%B5%D1%82%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D0%B8 (дата обращения: 27.03.2016).

8. *Лобанов Б. М., Киселев В. В.* Автоматизация клонирования персонального голоса и дикции для систем синтеза речи по тексту // Международная конференция Диалог-2003. Сборник научных трудов. – М., 2003. – С. 417–424.

9. Окно (весовая функция) [Электр. ресурс] // Статья из Википедии – Свободной энциклопедии. – режим доступа: https://ru.wikipedia.org/wiki/%D0%9E%D0%BA%D0%BD%D0%BE_(%D0%B2%D0%B5%D1%81%D0%BE%D0%B2%D0%B0%D1%8F_%D1%84%D1%83%D0%BD%D0%BA%D1%86%D0%B8%D1%8F) (дата обращения: 03.04.2016). ГИАБ

КОРОТКО ОБ АВТОРАХ

Тарасов Александр Александрович[1] – студент,
e-mail: tarasov258@gmail.com,
Костин Виталий Николаевич[1] – кандидат технических наук,
доцент, e-mail: iitem1@yandex.ru,
[1] НИТУ «МИСиС».

Gornyy informatsionno-analiticheskiy byulleten'. 2017. No. 2, pp. 209–222.

UDC 004.934

A.A. Tarasov, V.N. Kostin

ON THE PROBLEM OF DEVELOPING A SYSTEM OF SPEECH PROCESSING IN REAL TIME

This article describes the methods of developing a system of speech recognition, speech translation into a foreign language, and speech synthesis. Objective: to examine the relevance of developing such systems, to consider the business processes in the field of international communication, to identify system requirements, to describe the methods of developing such systems and analysis of existing solutions.

This article discusses the following design methods: analysis of models of existing solutions and building a model of the future solutions on this basis, building a data flow diagram (DFD-model), building an entity-relationship diagram (ERD-model).

In describing of the problem of developing of the system have been solved several non-trivial problems, such as reading the speech, analysis, translation into a foreign language, simulation of the speaker's voice.

For the implementation of speech analysis methods are applied, which are used in the voice cloning: Spectral analysis of the speech signal based on the fast Fourier transform, mathematical dynamic programming unit (DP-method), the algorithm of a marking of pitches of the speech signal.

Discussed in this article a system of speech processing allows to simplify the communication between different cultures.

Key words: problem of international communication, system of speech processing, automatic speech processing, speech recognition, speech analysis, speech translation, speech synthesis, voice cloning, speech cloning.

AUTHORS

Tarasov A.A.[1], Student,
e-mail: tarasov258@gmail.com,
Kostin V.N.[1], Candidate of Technical Sciences,
Assistant Professor, e-mail: iitem1@yandex.ru,
[1] National University of Science and Technology «MISiS»,
119049, Moscow, Russia.

REFERENCES

1. Kazancheva A. F. *Aktual'nost' problem mezhkul'turnoy kommunikatsii v sovremennom polikul'turnom prostranstve* (The urgency of the problems of intercultural communication in modern multicultural space), Pyatigorsk, PGLU, 2012, 1 p.

2. *Vladenie inostrannymi yazykami. Vladenie inostrannymi yazykami. FOM*, available at: http://fom.ru/Nauka-i-obrazovanie/10998 (accessed 13.11.2015).

3. Mehrabian, Albert; Ferris, Susan R. Inference of Attitudes from Nonverbal Communication in Two Channels. *Journal of Consulting Psychology*, 1967, 31(3), pp. 248–252. doi: 10.1037/h0024648.

4. *Avtomaticheskaya segmentatsiya i markirovka rechevogo signala. BLOG Web Programmista*, available at: http://juice-health.ru/archive/38-kompyuternyj-sintez-i-klonirovanie-rechi/184-avtomaticheskaya-segmentatsiya (accessed 25.03.2016).

5. *BPF (Bystroe preobrazovanie Fur'e). Kontrol'no-izmeritel'nye pribory i sistemy*, available at: http://www.kipis.ru/info/index.php?ELEMENT_ID=40417 (accessed 27.03.2016).

6. *Spektroanalizator my na nem vidim? ProSound.iXBT.com*, available at: http://prosound.ixbt.com/education/spektr-analys.shtml (accessed 27.03.2016).

7. *Chastota diskretizatsii*, available at: https://ru.wikipedia.org/wiki/%D0%A7%D0%B0%D1%81%D1%82%D0%BE%D1%82%D0%B0_%D0%B4%D0%B8%D1%81%D0%BA%D1%80%D0%B5%D1%82%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D0%B8 (accessed 27.03.2016).

8. Lobanov B. M., Kiselev V. V. *Mezhdunarodnaya konferentsiya Dialog-2003*. Sbornik nauchnykh trudov (International conference Dialog-2003. Collection of scientific papers), Moscow, pp. 417–424.

9. *Okno (vesovaya funktsiya)*, available at: https://ru.wikipedia.org/wiki/%D0%9E%D0%BA%D0%BD%D0%BE_(%D0%B2%D0%B5%D1%81%D0%BE%D0%B2%D0%B0%D1%8F_%D1%84%D1%83%D0%BD%D0%BA%D1%86%D0%B8%D1%8F) (accessed 03.04.2016).