



Речевые технологии

2/2012

Главный редактор — Харламов А.А., доктор технических наук

Состав редколлегии:

Потапова Р.К., доктор филологических наук, профессор,
заместитель главного редактора

Ронжин А.Л., доктор технических наук, доцент

Женило В.Р., доктор технических наук, профессор

Жигулёвцев Ю.Н., кандидат технических наук, доцент

Кривнова О.Ф., доктор филологических наук, профессор

Кушнир А.М., кандидат психологических наук

Лобанов Б.М., доктор технических наук (Беларусь)

Максимов Е.М., доктор технических наук

Голенков В.В., доктор технических наук, профессор (Беларусь)

Ромашкин Ю.Н., кандидат технических наук

Петровский А.А., доктор технических наук, профессор (Беларусь)

Хитров М.В., кандидат технических наук

Чучупал В.Я., кандидат физико-математических наук

Шелепов В.Ю., доктор физико-математических наук (Украина)

Кушнир Д.А., кандидат технических наук, ответственный секретарь

Содержание

<i>Селиух Р.А., Юхименко А.А.</i>	
Адаптация акустических моделей фонем к голосу диктора на основе метода MLLR	3
<i>Васильева Н.Б., Пилипенко В.В., Радуцкий А.М., Робейко В.В., Сажок Н.Н.</i>	
Корпус украинской эфирной речи	12
<i>Яценко В.В.</i>	
Параметризация типов предложений предметной области для системы устного фразаря-переводчика	22
<i>Робейко В.В., Сажок Н.Н</i>	
Преобразование между орфографическими и фонемными текстами для моделирования спонтанного произношения	33

<i>Крак Ю.В., Загваздин А.С.</i>	
Система распределённого автоматизированного документирования речевых сигналов	43
<i>Савенкова О.А.</i>	
Система сегментно-слогового распознавания изолированных слов из больших словарей	54
<i>Харламов А.А., Ермоленко Т.В., Дорохина Г.А., Гнитъко Д.С.</i>	
Метод выделения главных членов предложения в виде предикативных структур, использующий минимальные структурные схемы	75
<i>Васильева Н.Б.</i>	
Распознавание речевых образов суб-словного уровня в слитной украинской речи	86
<i>Вальтер Каспер</i>	
Sentiment Analysis for Hotel Reviews	96
Памяти Тараса Климовича Винценко	110

Адаптация акустических моделей фонем к голосу диктора на основе метода MLLR

Селюх Р.А., младший научный сотрудник

Юхименко А.А., научный сотрудник

В статье рассказывается об адаптации диктора для пофонемного распознавания изолированных слов украинского языка. Описывается метод максимальной правдоподобности линейной регрессии. Оцениваются матрицы линейных преобразований для корректирования начальных акустических фонемных моделей. Обсуждаются результаты экспериментальных исследований; особенно анализируется количество слов в адаптационном примере.

- *Распознавание речи • адаптация к голосу диктора • метод MLLR • фонемные модели • линейные преобразования.*

The paper deals with speaker adaptation for phoneme recognition of Ukrainian isolated words. The method of Maximum Likelihood Linear Regression (MLLR) is described. The matrixes of linear transformation are estimated in order to correct initial acoustic phoneme models. Results of experimental research of the adapted recognition system are discussed; particularly the amount of words in the adaptation sample is analyzed.

- *Speech recognition • adaptation • Maximum Likelihood Linear Regression (MLLR) • phoneme models • linear transformation.*

Введение

Пофонемное распознавание речевого сигнала предусматривает формирование речевого паспорта диктора, который включает в себя акустические модели фонем (вероятностные параметры моделей) [1]. Оценка этих параметров моделей фонем проводится с использованием данных обучающей выборки, которая должна содержать всё фонемное разнообразие речи. Опыт формирования таких выборок показывает, что их объём должен быть достаточно большим. Диктору необходимо потратить не один час для записи речевых сигналов с целью создать систему распознавания с приемлемой надёжностью при пофонемном распознавании изолированных слов из больших словарей [2]. Такая система распознавания будет давать неплохие результаты для диктора, на обучающей выборке которого происходило обучение распознаванию (оценка параметров). Этот диктор будет называться опорным. Но для другого, нового диктора эта же система распознавания будет выдавать неважные результаты. Напрашивается выход — провести точно такое же обучение для нового диктора, как и для опорного, с такой же большой обучающей выборкой. Но предполагается гипотетическая ситуация — либо новый диктор совершенно не имеет никакой возможности наговаривать большую обучающую выборку, либо не имеет ни малейшего желания этого делать. Резонно возникает вопрос: нельзя ли новому диктору произнести относительно небольшую обучающую вы-



борку из изолированных слов, а потом с помощью определённых методов адаптировать её к уже существующей системе распознавания, обученной на опорного диктора, и при этом получить приемлемую надёжность распознавания? Такая возможность должна существовать. Сравнение видеоспектрограмм, полученных при анализе речи разных дикторов, показывает, что при всём разнообразии проявления индивидуальных особенностей голоса видеоспектрограммы одних и тех же слов достаточно похожи. Таким образом, необходимо преобразовать речевые сигналы одного диктора в сигналы другого.

Следовательно, задача адаптации на голос диктора предусматривает предварительное проведение обучения на голос определённого опорного диктора или базового кооператива дикторов. Потом осуществляется корректирование параметров акустических моделей фонем для нового диктора на относительно небольшой адаптационной выборке. Также адаптация может применяться и к смене условий распознавания, например, при переходе на иной канал получения речевой информации (другой микрофон, телефонная линия).

Цель работы — исследование и применение к украинской речи одного из наиболее распространённых подходов в адаптации на голос диктора при пофонемном распознавании отдельно произносимых слов.

В предыдущих исследованиях по адаптации на голос диктора проводилось корректирование акустических моделей целых слов [3]. На нынешнем этапе мы переходим к адаптации на фонемном уровне.

Постановка задачи адаптации и пути её решения

Пусть имеются параметры акустических генеративных моделей фонем, вычисленные на основании итерационных процедур для опорного диктора или базового кооператива дикторов [3,4]. В частности, для каждой из трёх фаз-состояний фонемы φ (рис. 1) известны вектор математического ожидания $\mu = [\mu_1, \mu_2, \dots, \mu_n]^T$ и ковариационная матрица Σ , размерностью $n \times n$, где n — размерность вектора первичных признаков речевого сигнала.

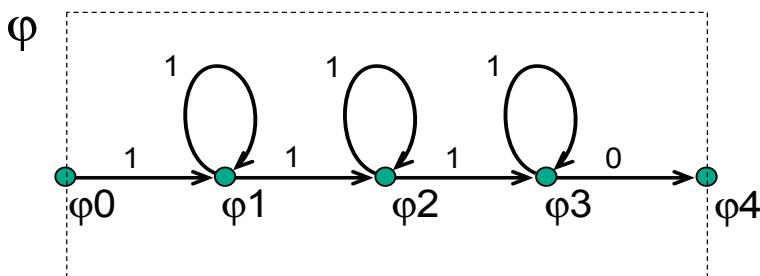


Рис. 1. Генеративная модель фонемы φ с тремя фазами-состояниями $\varphi_1, \varphi_2, \varphi_3$

Начальное состояние φ_0 и конечное φ_4 служат для соединения с другими моделями фонем в словах. Число рядом со стрелкой обозначает количество временных отсчётов, за которое осуществляется переход по стрелке.

Предполагается, что существует линейное преобразование, которое переводит векторы математического ожидания опорного диктора или базового кооператива дикторов в векторы математического ожидания нового диктора.

Это линейное преобразование представляет собой матрицу размерностью $n \times (n + 1)$. Действие этого преобразования — смещение средних значений параметров моделей фонем и изменение дисперсий этих параметров в начальной системе распознавания таким образом, что каждое состояние в системе акустических моделей фонем будет более точно генерировать данные адаптации, полученные от нового диктора.

Линейное преобразование для вектора математического ожидания записывается в виде:

$$\hat{\mu} = W\xi, \quad (1)$$

где $\hat{\mu}$ — вектор математического ожидания нового диктора, W — матрица преобразований размерностью $n \times (n + 1)$, ξ — вектор расширенного матожидания,

$$\xi = [1, \mu_1, \mu_2, \dots, \mu_n]^T \quad (2)$$

В свою очередь, матрица W представляется в виде:

$$W = [b \ A], \quad (3)$$

где A — матрица линейных преобразований размерностью $n \times n$, а b — вектор смещения в n -мерном пространстве.

В такой форме преобразование удобнее вычисляется в условиях непрерывного распределения по нормальному закону.

В отличие от исследований, представленных в [4], в этой работе рассматривается линейное преобразование и ковариационных матриц, которое представляется в виде:

$$\hat{\Sigma} = H \Sigma H^T, \quad (4)$$

где H — матрица преобразования ковариационной матрицы Σ размерностью $n \times n$.

Матрицы линейных преобразований получают путём оптимизации значения критерия распознавания. Один из таких оптимизационных алгоритмов — линейная регрессия максимальной правдоподобности (Maximum Likelihood Linear Regression — MLLR) [4]. Состояния всех моделей фонем разделяются на определённое количество классов регрессии методом векторного квантования. Затем для каждого класса регрессии вычисляются свои две матрицы преобразований — для матожидания и для ковариационной матрицы. В случае, когда состояния моделей фонем описываются смесью нормальных законов распределения — гауссианами (каждое состояние будет иметь несколько векторов матожиданий и такое же количество ковариационных матриц), тогда в классы регрессии входят отдельные гауссианы.

Ниже приведена стандартная вспомогательная функция, которая используется для вычисления преобразований:

$$Q(M, \hat{M}) = -\frac{1}{2} \sum_{r=1}^R \sum_{m_r=1}^{M_r} \sum_{t=1}^T L_{m_r}(t) \left[K^{(m_r)} + \ln(\hat{\Sigma}_{m_r}) + \right. \\ \left. + \ln(\hat{\Sigma}_{m_r}) + (o(t) - \hat{\mu}_{m_r})^T \cdot \hat{\Sigma}_{m_r}^{-1} \cdot (o(t) - \hat{\mu}_{m_r}) \right]$$

где M — множество моделей фонем,

\hat{M} — адаптированное множество моделей фонем,

R — количество классов регрессии,

M_r — количество гауссианов в r -м классе регрессии,



T — количество n -мерных векторов наблюдения из адаптационной выборки,

$o(t)$ — n -мерный вектор наблюдения из адаптационной выборки в дискретный момент времени $t, 1 \leq t \leq T$,

$L_{m_r}(t)$ — вероятность того, что вектор наблюдения $o(t)$ был «сгенерирован» гауссианом с номером m_r ,

$K^{(m_r)}$ включает все константы гауссиана m_r .

Для нахождения матрицы преобразования, например, векторов матожидания, вводится замена в выражение для MLLR адаптации матожидания

$$\hat{\mu}_{m_r} = W_r \xi_{m_r}, \quad \hat{\Sigma}_{m_r} = \Sigma_{m_r}$$

во вспомогательную функцию, и, принимая во внимание, что ковариационные матрицы — диагональные, получаем:

$$Q(M, \hat{M}) = -\frac{1}{2} \sum_{r=1}^R \sum_{m_r=1}^{M_r} \sum_{t=1}^T L_{m_r}(t) [K^{(m_r)} + \ln(\|\Sigma_{m_r}\|) + (o(t) - W_r \xi_{m_r})^T \cdot \Sigma_{m_r}^{-1} \cdot (o(t) - W_r \xi_{m_r})]$$

После ряда преобразований получаем формулу в виде:

$$Q(M, \hat{M}) = K - \frac{1}{2} \sum_{r=1}^R \sum_{i=1}^n [w_i G_r^{(i)} w_i^T - 2 w_i k_r^{(i)}],$$

где

w_i — i -я строка матрицы W_r ,

$$G_r^{(i)} = \sum_{m_r=1}^{M_r} \frac{1}{\sigma_{m_r, i}^2} (\xi_{m_r} \xi_{m_r}^T) \sum_{t=1}^T L_{m_r}(t)$$

$$k_r^{(i)} = \sum_{m_r=1}^{M_r} \sum_{t=1}^T L_{m_r}(t) \frac{1}{\sigma_{m_r, i}^2} \xi_{m_r} o_i(t)$$

Дифференцируя вспомогательную функцию относительно преобразования W_r , а потом максимизируя относительно к преобразованному среднему, получаем формулы для вычисления матрицы преобразования:

$$w_i = k_r^{(i)} G_r^{(i)-1}, \quad i = 1 : n, r = 1 : R.$$

Экспериментальная база

Были проведены экспериментальные исследования. В экспериментах задействовали 67 дикторов (25 мужчин и 42 женщины). Поскольку общеизвестен тот факт, что надёжность распознавания женских голосов ниже, количество женщин-дикторов больше чем мужчин. Каждый диктор наговаривал свою определённую обучающую выборку (далее ОВ). Поскольку этих определённых ОВ было 10, то разные дикторы могли наговаривать одинаковые слова. Всего этими дикторами было наговорено 2 416 разных слов. В алфавит фонем вошло 55 элементов.

В базовый кооператив дикторов было отобрано 53 диктора. Остальные 14 дикторов вошли в контрольную группу. Дикторы из контрольной группы наговаривали один и тот же набор слов (241 слово). Реализации этих слов не входят в базовый кооператив.

Результаты экспериментальных исследований

Результаты первого эксперимента отображены в таблице 1. В ней приведена усреднённая надёжность распознавания до и после адаптации к базовому кооперативу дикторов каждого диктора из контрольной группы отдельно на 30, 60, 100 и 150 слов.

Таблица 1

Результаты распознавания тестовых выборок слов для контрольной группы дикторов после адаптации на разное количество слов – 30, 60, 100 и 150 слов

Дикторы	Количество слов на адаптацию	30	60	100	150
1. Аня	93.78	95.13	95.30	95.32	97.07
2. Анна	91.29	92.76	93.19	93.90	94.51
3. Богдан	80.50	89.71	90.98	92.62	95.24
4. Валентина	95.02	95.26	96.13	96.03	94.87
5. Дмитрий	92.12	95.60	96.96	97.73	97.80
6. Катерина	79.25	86.60	87.66	90.21	90.48
7. Елена	90.46	94.11	95.40	95.32	96.34
8. Олеся	92.53	96.82	97.79	98.01	97.80
9. Руслан	89.21	93.23	94.57	95.46	95.24
10. Сергей	95.81	96.41	96.60	97.45	97.80
11. Слава	89.21	93.09	92.81	93.62	93.77
12. Татьяна	87.14	91.33	93.00	94.33	96.33
13. Юрий	89.21	93.16	93.93	96.31	96.70
14. Юрий В.	92.53	96.07	96.04	96.31	97.07
В среднем по группе	89.86	93.52	94.31	95.19	95.79

Количество гауссианов в смесях моделей фонем — 16.

Группа дикторов из контрольной группы (14 дикторов) из разных городов, все наговаривали один и тот же набор слов (241 слово).

Результаты, приведённые в таблице 1, показывают, что после адаптации на голос нового диктора надёжность распознавания в среднем выросла на 3,66% для адаптационной выборки объёмом в 30 слов, на 4,45% — для 60 слов, на 5,33% — для 100 слов, на 5,93% — для 150 слов. На рис. 2 изображён график надёжности распознавания в среднем по контрольной группе дикторов до и после адаптации.

Обучение распознаванию проводилось на основе базы данных для 53 дикторов из пяти городов Украины.

При адаптации вычислялись матрицы перехода для среднего и дисперсии.

Второй эксперимент заключался в том, чтобы базовый кооператив разбить на два по гендерному признаку. По такому же признаку контрольная группа разбивалась на две — женщин-дикторов и мужчин-дикторов. В данном случае женщины-дикторы адаптировались к женскому кооперативу, а мужчины-дикторы — к мужскому, соответственно. Предполагалось, что из-за существенной разницы женских и мужских голосов это даст повышение надёжности распознавания после адаптации.

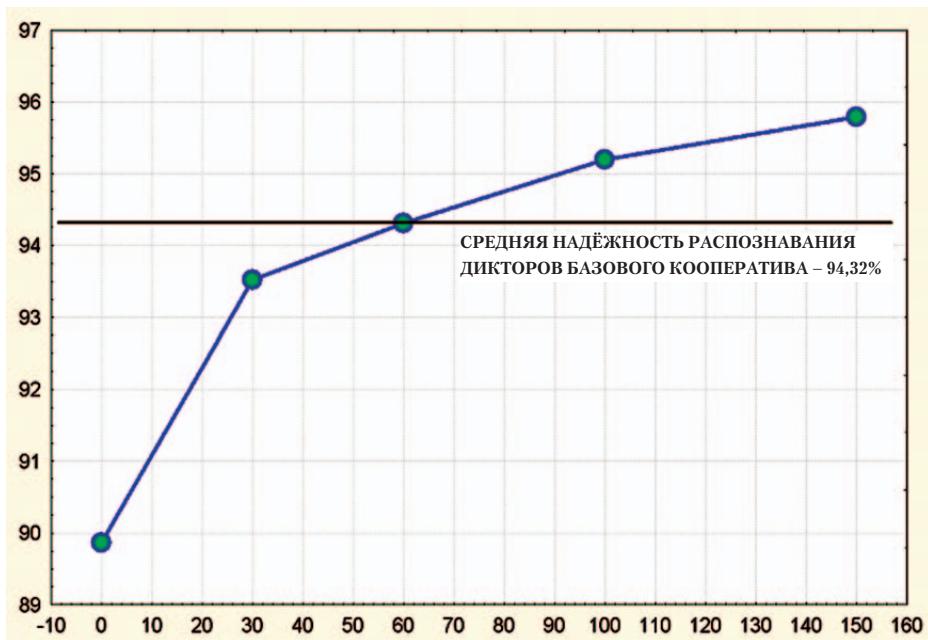


Рис. 2. Средняя надёжность распознавания дикторов контрольной группы до и после адаптации

В таблице 2 приведены усреднённые результаты надёжности распознавания для контрольной группы дикторов женского пола до адаптации и после адаптации к кооперативу женщин-дикторов на разное количество слов. Контрольная группа (7 женщин-дикторов) из разных городов, все наговаривали один и тот же набор слов (241 слово). Из таблицы видно, что после адаптации на голос нового диктора надёжность распознавания в среднем выросла на 2,41% для адаптационной выборки объёмом 30 слов, на 2,95% — для 60 слов, на 3,76% — для 100 слов, на 4,46% — для 150 слов. На рис. 3 изображены сравнительные графики надёжности распознавания в среднем по контрольной группе женщин-дикторов без учёта гендерности и с её учётом.

Обучение распознаванию проводилось на основе базы данных для 36 дикторов женского пола из нескольких городов Украины.

Таблица 2

Результаты распознавания тестовых выборок слов для контрольной группы женщин-дикторов после адаптации на разное количество слов к кооперативу женщин-дикторов — 30, 60, 100 и 150 слов

Дикторы	Количество слов на адаптацию		30	60	100	150
1. Аня	95.85	96.21	96.60	97.30	98.90	
2. Анна	92.95	93.64	94.02	94.61	96.33	
3. Катерина	84.65	89.37	89.78	92.20	93.04	
4. Елена	93.36	96.07	96.41	96.45	97.44	
5. Олеся	92.95	97.16	97.61	98.15	97.80	
6. Валентина	94.19	94.51	95.58	96.45	95.97	

7. Татьяна	88.80	92.62	93.37	93.90	94.51
В среднем по группе	91.82	94.23	94.77	95.58	96.28
Без учёта гендерности	89.92	93.14	94.07	94.73	95.34

Таблица 2

Результаты распознавания тестовых выборок слов для контрольной группы мужчин-дикторов после адаптации на разное количество слов к кооперативу мужчин-дикторов – 30, 60, 100 и 150 слов

Дикторы	Количество слов на адаптацию		30	60	100	150
1. Богдан	84.65	89.37	89.96	90.64	92.31	
2. Дмитрий	92.95	94.18	95.21	96.31	97.07	
3. Руслан	87.97	94.04	94.67	96.31	95.60	
4. Сергей	96.34	96.55	96.04	98.01	96.70	
5. Слава	91.70	93.57	94.01	94.61	94.87	
6. Юрий	90.04	91.81	93.46	93.48	93.41	
7. Юрий В.	91.29	96.47	95.96	97.02	97.43	
В среднем по группе	90.71	93.71	94.19	95.20	95.34	
Без учёта гендерности	89.80	93.90	94.56	95.64	96.23	

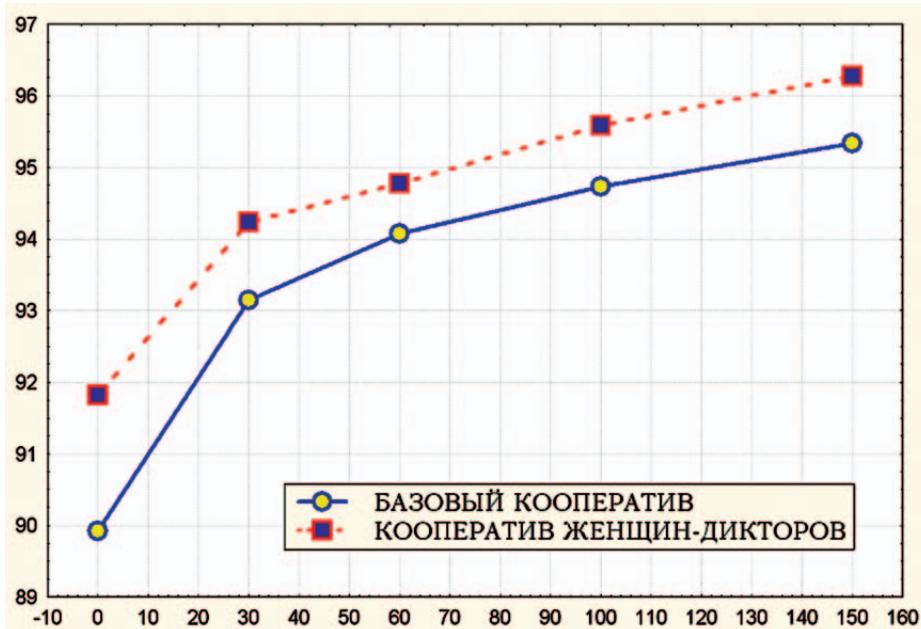


Рис. 3. Средняя надёжность распознавания дикторов женского пола

На рисунке 3 приведены усреднённые результаты надёжности распознавания для контрольной группы дикторов мужского пола до адаптации и после адаптации к кооперативу мужчин-дикторов на разное количество слов. Контрольная группа (7 мужчин-дикторов) из разных городов, все наговаривали один и тот же набор слов (241 слово). Из таблицы видно, что после адаптации на голос нового диктора надёжность распознавания



в среднем выросла на 3% для адаптационной выборки объёмом 30 слов, на 3,48% — для 60 слов, на 4,49% — для 100 слов, на 4,63% — для 150 слов. На рис. 4 изображены сравнительные графики надёжности распознавания в среднем по контрольной группе мужчин-дикторов без учёта гендерности и с её учётом.

Обучение распознаванию проводилось на основе базы данных для 17 дикторов мужского пола из нескольких городов Украины.

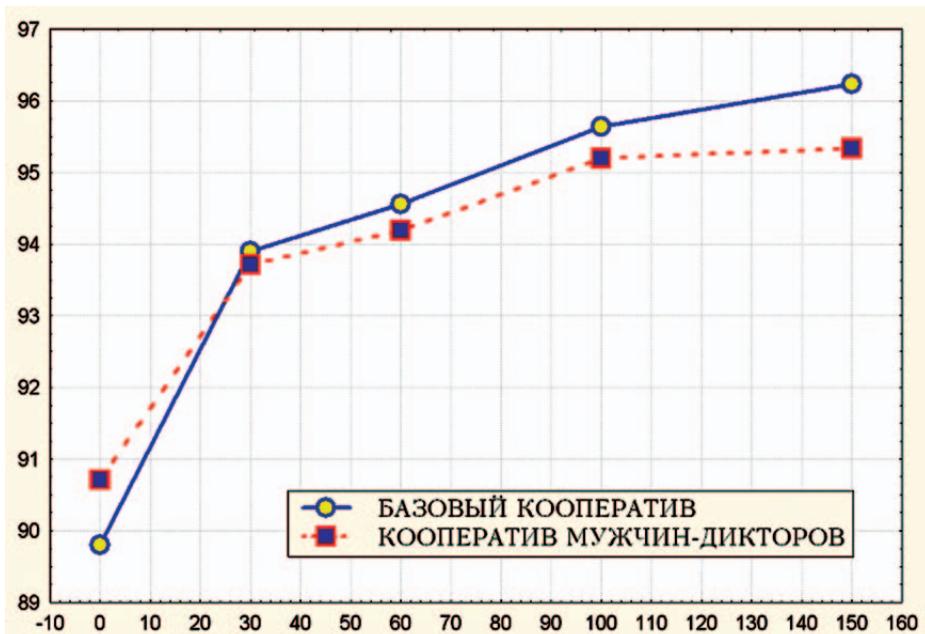


Рис. 4. Средняя надёжность распознавания дикторов мужского пола

Выводы

Результаты гендерозависимого распознавания показывают уменьшение количества ошибочно распознанных слов до 10–20% в сравнении с распознаванием на акустических моделях, сформированных на базовом кооперативе дикторов. Необходимо отметить, что средняя надёжность распознавания самих дикторов из базового кооператива составляет 94,32%. Фактически, в контрольной группе дикторов уже при адаптации на 60 слов достигается эта надёжность (в среднем, разумеется), не говоря уже о большем количестве слов на адаптацию. При этом дикторы базового кооператива заговорили свыше 12 000 слов в общей обучающей выборке. В такой ситуации преимущество адаптации очевидно.

Дальнейшая адаптация к голосу диктора на базе гендерозависимых акустических моделей показала такую же динамику уменьшения ошибок для дикторов женского пола. Этот эффект не наблюдался для мужского пола, очевидно, по причине меньшего количества дикторов-мужчин в базовом кооперативе.

Дальнейшие работы будут направлены на повышение качества адаптации, в том числе использования оценки длины голосового тракта диктора. Будут также исследованы другие пространства первичных характеристик сигнала. Планируется работа не только с изолированными словами, но и со слитной речью.

Література

- 1.** *Vintsiuk T. Speaker Voice Passport for a Spoken Dialogue System / Taras Vintsiuk, Mykola Sazhok //Proceedings of the 3rd International Workshop «Speech and Computer» — Specom'98. St.-Petersburg, 1998. P. 275–278.*
- 2.** *Vasylyeva N. Text Selection for Training Procedures under Phoneme Units Variety / N. Vasylyeva, M. Sazhok //Proceedings of the 10th International Conference on Speech and Computer — SpeCom'2005. Patras, 2005. P. 69–76.*
- 3.** *Винцюк Т.К. Аналіз, распознавание и смысловая интерпретация речевых сигналов.* Київ: Наукова думка, 1987.
- 4.** *Young S.J. HTK Book, version 3.1. Cambridge University, 2002.*
- 5.** *Olsen P. and Dharanipragada S. «An efficient integrated gender detection scheme and time mediated averaging of gender dependent acoustic models,» Eurospeech, 4. P. 2509–2512, September 1–4, 2003, Geneva Switzerland.*
- 6.** *Сажок М., Селюх Р., Юхименко О. Адаптація акустичних моделей фонем до голосу диктора для пофонемного розпізнавання ізольованих слів української мови // Штучний інтелект. Донецьк, 2009. № 4. С. 230–233.*

Сведения об авторах

**Международный научно-учебный центр информационных технологий и систем
Национальной академии наук Украины.**

Селюх Руслан Анатольевич —

младший научный сотрудник Международного научно-учебного центра информационных технологий и систем. Занимается проблематикой обучения, распознавания и адаптации речи. selyukh@uasoiro.org.ua

Юхименко Александр Анатольевич —

научный сотрудник Международного научно-учебного центра информационных технологий и систем. Занимается проблематикой обучения, распознавания и адаптации речи. yukhymenko@uasoiro.org.ua



Корпус украинской эфирной речи

Васильева Н.Б., научный сотрудник

Пилипенко В.В., старший научный сотрудник

Радуцкий А.М., кандидат технических наук

Робейко В.В., научный сотрудник

**Сажок Н.Н., кандидат технических наук,
старший научный сотрудник**

В статье описывается основанный на эфирных записях корпус украинской разговорной речи. Информация, содержащаяся в корпусе, используется для создания акустической и языковой модели в технологии распознавания речи. Изложен краткий обзор текущего состояния разработок речевых корпусов в мире. Даны характеристики созданных в Украине речевых корпусов, имеющих практическое или историческое значение. Обсуждается роль и особенности использования текстовых и речевых корпусов в современных речевых технологиях, указываются наиболее характерные ошибки и противоречия, возникающие при формулировании концепции корпуса. Представлены концепция большого репрезентативного корпуса разговорной речи и требования к его свойствам. Описаны фундаментальные понятия и технические решения, используемые при записи и аннотировании речевого материала. Введенная система обозначений обеспечивает подробное описание зарегистрированных данных. Определены первоочередные задачи, стоящие на пути расширения корпуса. Экспериментальные исследования текущей версии речевого корпуса показали устойчивое повышение надежности распознавания речи по сравнению с более ранними стадиями разработки корпуса. Представленный материал рассматривается как шаг к созданию национального речевого корпуса, применимого для разработки широкого диапазона речевых технологий.

- корпус спонтанной речи • распознавание речи • сегментирование
- аннотирование.

In this paper we describe a media-based speech corpus for spoken Ukrainian language. Information contained in the corpus is aimed to develop acoustic and language models for speech recognition technology. We give an overview of the current state of the art in speech corpora all over the world. Developed in Ukraine speech databases both historical and available today are listed and summarized. Nowadays role and specific features of text and speech corpora are investigated as well as the most frequent mistakes and misunderstandings of the corpus concept are discussed. The concept of a large representative corpus of spoken language and its desired properties are presented. Basic concepts and technical solutions used for speech corpus recording and annotation are described. The introduced mark-up system provides a detailed description of the recorded data. The most significant problems standing in the way of building a huge speech corpus are pointed out. A current version of the speech corpus has been validated with HTK tools that showed steady progress of speech recognition accuracy comparing to early stages of corpus development. We consider the presented corpus as a step to creation the national speech corpus applicable for entire range of speech technology.

- spontaneous speech • speech recognition • segmentation • annotation.

Вступление

Речевые корпуса играют большую роль при разработке речевых информационных технологий. Информация, которая содержится в таких корпусах, используется для построения акустических и лингвистических моделей для построения как систем наговаривания речи, так и моделей диалога человека с машиной, а также моделей предметных областей для смысловой интерпретации речи. Особые требования предъявляются к корпусам, которые разрабатываются для высококачественных систем автоматического распознавания речи и озвучивания текстов. Каждый корпус создаётся с определённой целью, которая учитывает определённую специфику научных исследований или разрабатываемых прикладных систем.

Создание данного корпуса длится уже около двух лет. Результат этой работы — пилотная версия корпуса эфирной речи.

Цель данной работы — описание структуры корпуса, средств формирования корпуса, первых конкретных результатов анализа и использования речевого материала, а также перспектив дальнейших исследований.

Опыт создания речевых корпусов в Украине

Речевой корпус состоит из структурированного множества речевых фрагментов, описания этих фрагментов, а также компьютерных средств для оперирования со всем множеством данных корпуса.

Речевой фрагмент как базовая единица корпуса — это оцифрованный фрагмент речевого сигнала, который сопровождается ассоциированной информацией определённого типа (типов). Такая информация называется аннотацией речевого фрагмента [1].

Создание акустических корпусов — достаточно сложная научная и технологическая задача, которая требует значительных ресурсов. В 90-е гг. ХХ в. во многих странах были созданы координационные центры для сбора, хранения и распространения общедоступных и стандартизованных корпусов, в том числе и речевых [2]. Создание акустических корпусов становится самостоятельным направлением речевых технологий.

В Украине первые корпуса речи были созданы в 70-е гг. прошлого столетия для тестирования и оценки показателей систем распознавания речи на одинаковом стандартном речевом материале. Корпус из 1 тыс. отдельных слов использовался для тестирования системы распознавания на основе ЭВМ БЭСМ-6, при этом была достигнута точность распознавания в 96% при словаре в 1 тыс. слов [3]. Также для тестирования кооперативной (многодикторной) системы распознавания была накоплена выборка из 1600 реализаций из словаря в 100 слов для 6 дикторов. Было показано, что метод кооперативного обучения позволяет достичь 92% точности распознавания речи диктора, не входящего в кооператив [4].

В 90-е гг. ХХ в. для тестирования распознавания ключевых слов была записана английская слитная речь 11 дикторов длительностью в 3500 слов и размечена экспертами на отдельные слова, а часть материала — для обучения на отдельные фонемы [5].

Толчком в развитии пофонемного распознавания украинской речи послужил однодикторный корпус, содержащий более 6 тыс. изолированных слов, в значительной мере покрывающих фонетическое разнообразие языка. Акустическая модель, созданная на основе этого корпуса, позволила превысить надёжность 95% на словаре 3 тыс. слов. Также, начиная с 2004 года, успешно демонстрировалась базовая технология фонетического стенографа, как одно из достижений Государственной научно-технической программы «Образный компьютер».

Создание алгоритма распознавания речи из сверхбольших словарей (до 2 млн слов) потребовало накопление корпуса речи в 14 тыс. слов и сочетаний слов. Была достигнута



точность распознавания в 99,9% для словаря в 1 тыс. слов, а также точность в 85% для словаря в 2 млн слов при среднем времени распознавания в 7 сек. [6].

Многодикторный корпус «UkReco» содержит более 30 тыс. реализаций фонетически сбалансированных слов и фраз, записанных от около 100 дикторов из разных регионов Украины. Этот корпус используется для распознавания изолированных слов, адаптации на голос диктора, а также для построения акустических моделей для словаря-переводчика [7].

Другой размеченный корпус речи, записанной через телевизионную сеть, состоит из выступлений около 330 депутатов Верховной Рады Украины. Речь депутатов отличается быстрым темпом, спонтанностью и эмоциональностью. Объем обучающей выборки — 54 часов речи, контрольной — 11 часов речи. Средняя точность распознавания для контрольной выборки составила 71% [8].

Для исследования методов послогового и морфемного распознавания речи был накоплен корпус из более 35 часов читаемой речи одного диктора [9].

Интересное направление использования корпусов речи — их использование для синтеза речи. Такие корпуса предъявляют особые требования к качеству записи и подробности описания речевого сигнала. Для озвучивания украиноязычных текстов был записан женский голос профессионального диктора в студийных условиях [10].

Опыт, накопленный в предыдущих разработках, стал неоценимым при создании концепции данного корпуса эфирной речи.

Структура и состав акустического корпуса

Акустический корпус украинской эфирной речи (Акустичний корпус українського ефірного мовлення — АКУЕМ) — общий по цели своего применения акустический корпус, который содержит читаемую, подготовленную и спонтанную речь (последнее составляет самую большую часть корпуса). Все материалы корпуса по типу речевого сигнала разделяются на теле- и радиовещание, также присутствуют небольшие вкрапления записи публичной речи и речи в естественной среде. Основные языки материалов корпуса — украинский и русский.

В АКУЕМ вошли материалы разной тематики и жанров, но основу корпуса составили звуковые записи рубрик: новости и интервью (политика, культура, образование, общество и т.д.), телепередачи и телетрансляции (судебные заседания, политические дебаты, публичные выступления и др.). В целом корпус должен отображать полную картину речи украинского теле- и радиоэфира, поэтому работы над его пополнением будут вестись и в дальнейшем. В настоящее время количественное распределение звуковых записей по жанрам неравномерно. Это связано с первоочерёдностью отбора речевого материала определённой тематики, необходимой для работ по созданию системы распознавания речи (см. рис. 1).

На данный момент корпус украинской эфирной речи характеризуется следующими количественными показателями: более 260 часов аннотированной речи, словарь корпуса содержит почти 45 000 слов украинского языка и почти 50 000 слов русского языка, более 1500 тыс. дикторов. Среди записей встречается речь дикторов разного возраста, пола, социального положения и профессий, что отражает состав дикторов телевизионного эфира.

Кроме общеупотребительных слов, был создан словарь суржика (более 1700 слов), словарь территориальных и социальных диалектов (более 800 слов).

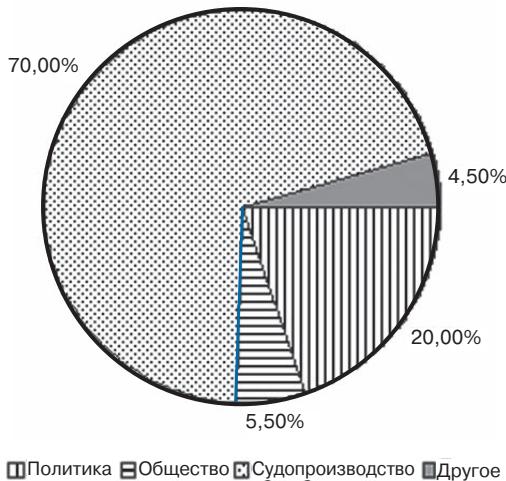


Рис. 1. Соотношение записей разных тематик (по продолжительности)

Разметка корпуса

Одна из основных черт, которые отличают акустический корпус от обычной коллекции звуковых записей или текстов, его разметка (аннотирование) — описание дополнительной информации о речевом сигнале.

Разметка АКУЕМ проводилась экспертами на основании предварительной автоматической разметки. Фактически, эксперт перепроверяет предварительную аннотацию, исправляя ошибки, делая необходимые дополнения, а также добавляя информацию о дикторах.

Разметка заключается в сегментировании речевого материала и детальном описании определённых лингвистических и экстралингвистических явлений в речевых фрагментах. Для внесения дополнительной информации в текст используются специфицированные теги, которые отделяются от текста знаком *. На данный момент используется 74 таких метабозначения.

Все обозначения можно разделить на несколько групп:

- обозначения языка;
- обозначения нелитературных слов;
- обозначения способа произношения слов;
- обозначения фона;
- обозначения неинформативных слов и звуков, которые произносит диктор;
- обозначения диалогов и хоров;
- обозначения шума.

Обозначения языка касаются всех последующих слов до альтернативного обозначения, и ставятся перед первым словом соответствующего языка. В АКУЕМ в настоящее время встречаются записи девяти языков, хотя основной объём (более 97%) составляет украинский и русский языки.

Следующая группа обозначений предназначена для слов, отсутствующих в литературных словарях:

- суржик — смесь украинского и русского языков;
- социальные диалекты (жаргон, арго) — языки людей, связанных определённой общностью профессиональных или социальных интересов;
- территориальные диалекты — языки лиц, связанных между собой территориальной общностью;
- аббревиатуры и сокращения.



Обозначения способа произношения слов включают обозначения дефектов речи (заикание, картавость и др.), речевых сбоев (обрывы и оговорки), специфического произношения слов (например, послогового, с редуцированием или с растягиванием). Все эти обозначения касаются только одного слова и ставятся перед соответствующим словом.

Обозначения неинформативных слов и звуков, которые произносит диктор, включают обозначения заполненных пауз, звуков-паразитов и подобных явлений. К этой же группе относятся неинформативные звуки, например покашливание, шмыганье носом, смех, плач, громкий вдох или выдох диктора. Такие обозначения ставятся на месте соответствующего звука и обозначают соответствующие звуки в записи, которые произносит диктор. Эта группа обозначений самая большая.

Обозначения диалогов соответствуют местам в звуковых записях, где так или иначе сливаются речь нескольких дикторов. Диалог — места, где во время разговора двух дикторов конец фразы первого диктора накладывается на начало фразы другого диктора. Хор — полное наложение речи нескольких дикторов.

Важная группа обозначений, которые описывают звуковые сегменты корпуса, — обозначения фона, на котором говорит диктор, и разнообразных шумов, которые присутствуют в сегментах. Такие обозначения касаются целого сегмента речи.

Примеры обозначений и частота их использования приведены в таблице 1.

Таблица 1

Некоторые обозначения, которые используются для описания сегментов АКУЕМ

Обозначение	Значение	Частота использования
y	украинский язык	3299
p	русский язык	3240
p-ак	русский язык с сильным иностранным акцентом	90
c	суржик	6300
ж	жаргон, арго	943
об	оговорка	3388
карт	картавость (неправильное произношение звуков «р» и «л»)	1191
см	смех	369
е	экание	15030
хор	хор	6066
пт	шелест бумаги (фон)	4695
мт	музыка (фон)	11718
опл	аплодисменты (фон)	1281
стук	стук	979
вул	шум улицы	2

Целевая аудитория АКУЕМ

Целевая аудитория проекта в первую очередь — разработчики систем автоматического распознавания украинской и русской речи. АКУЕМ предназначен для обучения и тестирования таких систем распознавания речи. Современ-

ным статистическим системам распознавания речи необходим большой объём акустических материалов для построения акустических и лингвистических моделей (далее АМ и ЛМ) речи, а также для тестирования надёжности распознавания речи.

На материалах корпуса проводятся многочисленные научные эксперименты в области распознавания речи, например, выявление и классификация экстралингвистических речевых явлений, исследование реальных акустических условий речи, исследование различных вариантов произношения дикторов, изучение специфики устной спонтанной речи на разных уровнях и много других.

АКУЕМ отображает современную языковую ситуацию в Украине, включает как литературный, так и разговорный стиль речи. Поэтому корпус может служить основой для широкого спектра исследований в области лингвистики, диалектологии, речевой акустики, психоакустики, фонетики, фонологии и других областях науки.

Программное обеспечение корпуса

Эффективное создание корпуса невозможно без развитого инструментария. К этому инструментарию относятся программные средства для стенографирования звуковых записей, дальнейшего их сегментирования и аннотирования (транскрибирования), автоматического исправления транскрипций, статистического анализа результатов сегментирования, а также подготовки материала к обучению распознавания.

Средства стенографирования звукозаписей

Стенографирование производится средствами протоколирования событий SRS-Femida [11].

Стенографист создаёт транскрипцию звукозаписи с уровнем детализации, которая включает признаки языка и говорящего. С помощью ножной педали осуществляется предварительное разделение на речевые сегменты, которые отвечают смене говорящего. Видеоряд, который сопровождает звукозаписи, облегчает определение диктора.

Кроме этого, указываются участки сигнала, где речь неразборчивая, перекрывается шумами или отсутствует.

Для обеспечения орфографической правильности набранного текста используются стандартные средства проверки орфографии, адаптированные к специфике стенограмм: учитываются обозначения языковых признаков и добавляются признаки отклонения от нормативов литературного языка для соответствующих слов.

Средства сегментирования

Сегментирование выполняется средствами программного обеспечения с открытым кодом *Transcriber 1.5.1* [12] (см. рис.2), адаптированного к кириллице. Подготовленный специалист с соответствующим уровнем лингвистического и компьютерного образования углубляет детализацию транскрипции, полученной в результате стенографирования звукозаписей. Проводится тщательное разбиение по паузам речевых сегментов, синхронизация их с соответствующим текстом. Кроме этого, в текст вставляются детальные признаки-теги, которые касаются как отдельных слов и звуков, так и речевого сегмента в целом.

Дальнейший анализ сегментирования состоит в выявлении и исправлении типичных ошибок и внесении некоторых регулярных изменений, обусловленных как непрерывным развитием концепции корпуса, так и появлением различных версий использования корпуса.

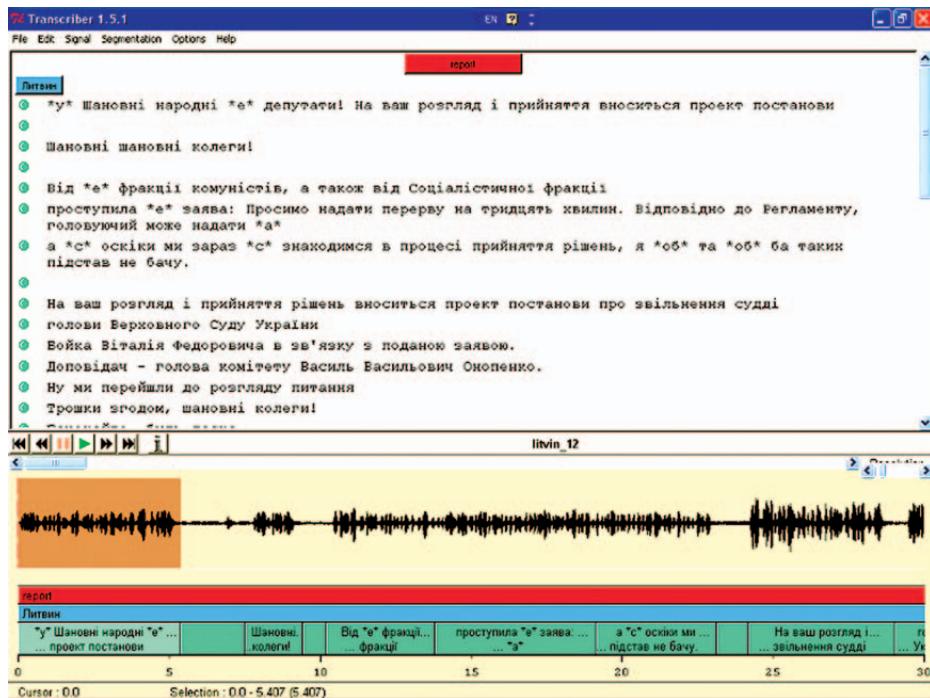


Рис. 2. Диалоговое окно эксперта в программе Transcriber

Средства статистического анализа

Для анализа накопленного материала производится подсчёт различных статистик, в частности формируются:

- частотные словари для разных языков, которые встречаются в корпусе;
- частотные словари суржика, социальных и региональных диалектов, абривиатур, редуцированных слов и др.;
- статистика длин речевых сегментов для каждого звукового файла, а также общая статистика;
- статистика длин речевых сегментов для каждого диктора в отдельности.

Средства подготовки к обучению распознавания речи

Кроме указанного выше, производится формирование звуковых файлов, применимых для обучения и распознавания речи. При этом каждому звуковому фрагменту соответствует текстовая запись и имя диктора.

Словарь системы распознавания дополнен словами, которые отвечают неинформационным звукам (например, заполненные паузы) и, соответственно, во время формирования текста фразы эти звуки рассматриваются как отдельные слова. Были проведены эксперименты по обучению таким звукам-словам, и результаты показали высокую точность их определения (около 80%).

Обозначения, которые характеризуют целый сегмент, например, *стук*, *вул*, предлагается использовать для построения моделей гауссовских смесей (GMM) для того, чтобы система распознавания определяла такие сегменты и относила их к соответствующему классу.

Информацию о дикторе предлагается использовать для настройки системы распознавания на кластеры дикторов. Это позволит повысить точность рас-

познавания за счёт предварительного определения кластера дикторов и использования индивидуальной акустической модели для данного кластера.

Предварительные эксперименты по распознаванию слитной речи

1. Речевой материал

Для экспериментов по распознаванию речи, относящейся к судебной тематике, использовалась только часть аудиофайлов. Это в основном записи телепередач «Судові справи» («Судебные дела»). Речь, звучащую в этих телепередачах, можно назвать спонтанной по форме, но не по содержанию, поскольку дикторы говорили в рамках соответствующих ролей. Кроме этого, часть аудиофайлов содержит записи реальных судебных заседаний, в которых присутствует как спонтанная речь судьи, так и неподготовленное (и, таким образом, приближенное к спонтанному) чтение протоколов.

Речевой материал, использованный для построения АМ, состоял из аудиозаписей (длительностью около 52 часов), в которых содержится речь около 1500 дикторов. Распределение неравномерное: большинство дикторов представлено короткими записями, однако, у 150 дикторов длительность записей составляет более 10 минут.

2. Текстовый материал

Текстовый материал, использованный для построения лингвистических моделей, состоит из текстов, загруженных из Интернета (400 Мбайт). Загруженный текст был модифицирован для того, чтобы убрать служебную информацию, записать числа в текстовом виде, а также отделить тексты на разных языках. В дополнение к этим текстам использовались также расшифровки звукового материала из обучающей выборки АКУЕМ.

3. Контрольная выборка

Для распознавания использовались записи длительностью 3,74 часа, в которых встретилось 29 500 слов. Всего в контрольной выборке присутствовала речь 34 дикторов. Темп произнесения — средний и быстрый.

4. Система распознавания речи

Для исследований использовался инструментарий HTK [13]. На его основе была создана многодикторная система распознавания речи.

В качестве АМ используются скрытые Марковские модели, обученные на обучающей выборке. 56 украинских контекстно-независимых фонем (включая фонему-паузу) моделируются тремя состояниями Марковской цепи без пропусков. Используется диагональный вид Гауссовых функций плотности вероятности. Редко встречающиеся фонемы моделируются 64 смесями Гауссовых функций плотности вероятности, более часто встречающиеся фонемы моделируются большим числом смесей, наиболее часто встречающиеся фонемы используют 1024 смесей.

В качестве лингвистической модели языка использовалась биграммная статистическая модель.

Словарь распознавания, используемый наряду с уже обученными ЛМ и АМ, насчитывал 42 598 словоформ. Произнесение каждой словоформы было представлено транскрипцией, несколько отличающейся от канонической (литературной). А именно, односложные словоформы представлены двумя транскрипциями (ударный и безударный варианты), а также упрощены некоторые сочетания согласных в соответствии со спонтанным произнесением (например, «дч» → «чч» вместо канонического «джч»).

Результаты распознавания приведены в таблице 2. Заметим, что в контрольной выборке наряду с записями телепередач присутствуют записи реального судьи Ш.



Таблица 2

Результаты распознавания речи

Дикторы	Профессия	Надёжность распознавания (%)
Окис	актёр в роли судьи	73,47
Калинская	актриса в роли судьи	58,65
Ш.	судья	59,47
Антонюк	актриса в роли прокурора	63,90
Наум	актёр в роли прокурора	59,10
Бойко	актёр в роли прокурора	57,76
Бевз	актёр в роли адвоката	55,93
Жуковская	актриса в роли адвоката	66,38
Бабич	актриса в роли адвоката	51,64
Бузаджи	актёр в роли адвоката	60,28
Солодко	актёр в роли адвоката	46,95
Сологуб	актриса в роли судебного секретаря	81,26
В среднем		59,61

Выводы

Разработанная пилотная версия АКУЕМ позволяет строить акустические и дополнять лингвистические модели для исследования по автоматическому транскрибированию звуковых сигналов, для поиска ключевых слов, а также для распознавания дикторов.

Дальнейшие исследования предусматривают построение информационно-поисковой системы на основе веб-интерфейса, который позволит пользователям ориентироваться в речевом материале и находить в нём нужную информацию более эффективно. Также полезными могут оказаться средства для синхронизации текстовых и речевых материалов.

Несмотря на сложность и трудоёмкость, мы надеемся создать полноценный ресурс, который станет основой для многих речевых технологий и систем, которые могут использоваться во многих сферах экономики, образования, права и повседневной жизни. Материал корпуса состоит из разнообразных звуковых записей вместе с их расшифровкой и может стать частью Национального корпуса украинского языка.

Литература

1. Кривнова О.Ф. Речевые корпуса на новом технологическом витке // Речевые технологии. 2008. № 2. С. 13–24.
2. Кривнова О.Ф., Захаров Л.М., Строкин Г.С. Речевые корпусы (опыт разработки и использование) // Труды семинара Диалог'2001. Москва, 2001.
3. Винцюк Т.К., Шинкаж А.Г. Распознавание 1000 слов // Автоматическое распознавание слуховых образов. Тбилиси, Мецниереба, 1978.
4. Винцюк Т.К., Кулляс А.И., Людовик Е.К., Шинкаж А.Г. Кооперативная система распознавания речи // Автоматическое распознавание слуховых образов. Ереван, 1980.
5. Вінцюк Т., Біднюк С., Кулляс А., Пилипенко В., Дослідження з розпізнаванням ключових слів у потоці зв'язного мовлення // Праці першої всеукраїнської конференції УкрОБРАЗ 92. Київ, 1992. С. 125–128.

6. Pylypenko V. Information Retrieval Based Algorithm for Extra Large Vocabulary Speech Recognition // Proc. of the 13th International Conference «Speech and Computer: SPE-COM'2006». St. Petersburg, Russia, 2006. P. 67–69.
6. Сажок М., Селюх Р., Юхименко Ю. Адаптація акустичних моделей фонем до голосу диктора для пофонемного розпізнавання ізольованих слів української мови // Штучний інтелект. Донецьк, 2009. № 4. С. 230–233.
7. Pylypenko V., Robeiko V. Experimental System of Computerized Stenographer for Ukrainian Speech. // Proc. of the 13th International Conference «Speech and Computer: SPE-COM'2009». St. Petersburg, Russia, 2009. P. 67–70.
8. Васильєва Н., Сажок М. Порівняння пофонемного та поскладового розпізнавання мовленнєвого сигналу для української мови // Праці десятої всеукраїнської міжнародної конференції УкрОБРАЗ, Київ, 2010. С 49–54.
9. Lyudovskyk T., Brozinski S., Noner M., Robeiko V., Sazhok M. Speech Synthesis Applied to SMS reading // Proc. of the 13th International Conference «Speech and Computer: SPE-COM'2009». St. Petersburg, Russia, 2009. P. 300–305.
10. Радуцький О., Богданов Л. Технічна фіксація судових процесів: системний підхід до розвитку комп’ютерних технологій та інформаційних ресурсів // Юридичний журнал. 2002. № 2. <http://www.justinian.com.ua/article.php?id=431>
11. Barras C., Geoffrois E., Wu Z., Liberman. Transcriber: a free Tool for Segmenting, Labeling and Transcribing Speech. In: Proc. First Int. Conf. on Language Resources and Evaluation (LREC 98), Granada, Spain, M., 1998. P. 1373–1376.
12. Young S. et al. The HTK Book (for HTK Version 3.4) // Cambridge University Engineering Department: Cambridge, UK, 2009. <http://htk.eng.cam.ac.uk/>

Сведения об авторах

Васильєва Ніна Борисовна —

научный сотрудник отдела распознавания и синтеза звуковых образов Международного научно-учебного центра информационных технологий и систем, г. Киев, Украина. n.vassilleva@gmail.com

Пилипенко Валерій Васильевич —

старший научный сотрудник отдела распознавания и синтеза звуковых образов Международного научно-учебного центра информационных технологий и систем, г. Киев, Украина. valeriy.pylypenko@gmail.com

Радуцкий Александр Михайлович —

кандидат технических наук, директор ООО «Специальные регистрирующие системы», г. Киев, Украина. alex@srs.kiev.ua

Робейко Валентина Васильевна —

научный сотрудник отдела распознавания и синтеза звуковых образов Международного научно-учебного центра информационных технологий и систем, г. Киев, Украина. valya.robeiko@gmail.com

Сажок Николай Николаевич —

кандидат технических наук, старший научный сотрудник отдела распознавания и синтеза звуковых образов Международного научно-учебного центра информационных технологий и систем, г. Киев, Украина. sazhok@gmail.com



Параметризация типов предложений предметной области для системы устного фразаря-переводчика

Яценко В.В., младший научный сотрудник

В статье рассматриваются подходы построения системы перевода устного сигнала в рамках предметных областей. Блок интерпретации получает произнесённое предложение в виде последовательности слов, распознанной декодером. На выходе системы принимается решение о принадлежности распознанной последовательности слов типу предложения, задающего тип смысла. Распознавание выполняется с учётом параметров, которые описывают множество возможных вариантов высказываний. Проанализированы альтернативные подходы моделирования ограничений на допустимые последовательности слов. Надёжность распознавания НММ-декодера в условиях сформированных акустической и лингвистической моделей позволила получить приемлемую интерпретацию распознанного сигнала. Это легло в основу разработки демонстрационной системы устного фразаря-переводчика.

- распознавание и интерпретация речи • словарь-переводчик • предметная область • тип смысла • тип предложения • украинская разговорная речь.

In this paper we describe approaches to build the spoken translation system within a subject area. The decoded sequence of words enters to the interpretation subsystem, which finally makes decision concerning the sentence type and the respective meaning type for the pronounced sentence. Possible variations of date, time, place etc. that may occur in sentences are parameterized and integrated to the language model. Strict, free and phonetic word based word grammars for speech decoder are analyzed. Acoustic and language models created for the HMM-based decoder shows such performance that allows for understanding response accuracy sufficient for practical application. The demonstration version of the spoken interpreter has been developed and presented.

- speech recognition and understanding • spoken phrasebook • subject area • meaning type • sentence type • ukrainian spoken language.

Среди важных практических задач, связанных с распознаванием речи, к которым относятся системы надиктовывания текстов, справочные системы, системы речевого управления оборудованием, системы речевого диалога и т.д., мы выделяем систему устного перевода. Актуальность задачи, в частности, отображается востребованностью автоматизации всем известного бумажного разговорника, в котором пользователь вынужден искать необходимую фразу и озвучивать её перевод. Вместо этого пользователю предлагается

только произнести фразу на родном языке в выбранной теме. Далее система делает всё самостоятельно. Дополнительного внимания требует вопрос моделирования параметров в предложениях, т.е. необходимо предусмотреть все возможные варианты значений для определённого предложения. Например, в вопросе о путешествии в конкретный город параметром будет название города.

Такие системы актуальны в свете использования их при разговоре с носителем другого языка. Пользователь не только получает перевод фразы, но и её озвучивание, что существенно облегчает общение в неродной языковой среде.

При построении систем устного перевода в рамках предметных областей возникает ряд проблем, общих с проблемами задачи понимания речевого сигнала. Необходимо построить модели всех возможных предложений языка диалога, которые выражают один и тот же смысл, смоделировать параметры слов в типах предложений, сгенерировать и найти наиболее правдоподобные эталонные сигналы, учитывая параметры.

Для исследования и спецификации ограничений на допустимые последовательности слов во фразах использовались LISP-структуры [1, 2]. На основе этих структур генерируется большое количество предложений, которые имеют одинаковый смысл с точностью до параметров. Впрочем, существует ряд ограничений на использование этой технологии, связанных как с субъективным фактором при построении LISP-структур, так и с увеличением количества вычислений, обусловленных существенным усложнением графа распознавания.

В качестве альтернативы LISP-структур предлагается способ оценивания принадлежности последовательности слов типам предложений, которые характеризуют смысл [3]. Этот подход требует развития, в частности, с целью учёта возможных ошибок распознавания.

Для моделирования ограничений на порядок следования слов использовались грамматические знания [2]. Для моделирования параметров в типах предложений использовались базы данных и базы знаний. Была сформулирована лингвистическая модель интерпретации распознанного сигнала с учётом параметров.

Общая структура системы устного перевода в пределах предметных областей

Распознавание и смысловая интерпретация слитной речи выполняются во взаимосвязанном процессе, конечная цель которого — перевод смысла сообщения на другой язык.

Рассмотрим задачи распознавания и интерпретации слитной речи [1, 2] и их взаимосвязь. Распознавание речи — процесс автоматической обработки сигнала с целью определения последовательности слов, которые передаются этим сигналом. Смысловая интерпретация языка — процесс автоматической обработки речевого сигнала с целью выявления смысла, передаваемого сигналом, и представление этого смысла в определённой канонической форме, удобной для дальнейшего использования в системе устного перевода.

Очевидно, что смысловая интерпретация языка является более высокой степенью обобщения информации, чем распознавание. Поскольку каждую мысль можно выразить различными предложениями в языке диалога без изменения содержания, то следует определить некоторые ограничения на допустимые последовательности слов в предложениях. Поэтому, при интерпретации смысла речи различные предложения, которые передают одну и ту же мысль, должны отражаться в один и тот же результат, т.е. ответ распознавания не должен противоречить синтаксису, семантике и прагматике предметной области.

Ввиду этого, предлагается рассмотреть структуру системы устного перевода в рамках предметных областей (рис.1). Задача смысловой интерпретации слитной речи с целью дальнейшего перевода основывается на том, что сначала пользователь должен задать предметную область (далее ПО), с которой он хочет работать. Для этого нужно назвать эту ПО. Вообще рассматривается 15 ПО, с которыми может работать пользователь.

Активатор выбирает названную ПО и загружает подсловари ПО с соответствующими этой области типами предложений и грамматику, по которой моделируются допустимые ограничения на последовательности слов в предложениях.

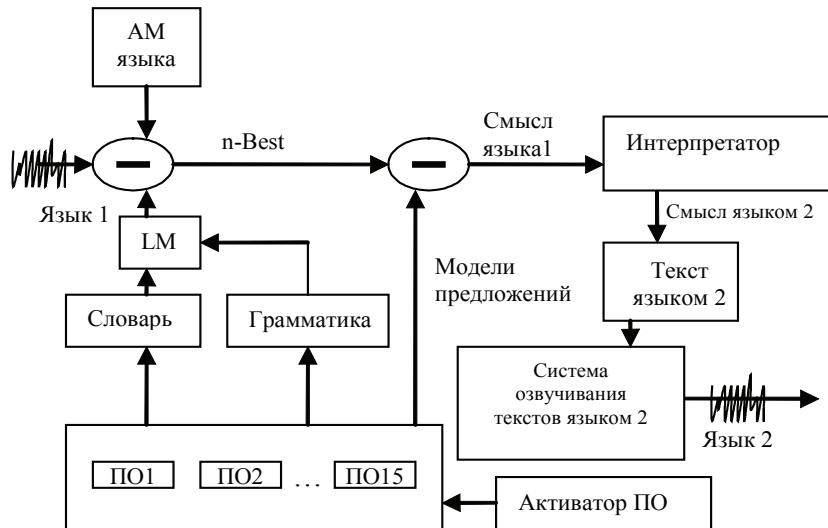


Рис.1. Структура системы устного перевода в рамках предметных областей

Диктор произносит на языке 1 предложение, которое распознаётся с учётом акустической модели и, построенной согласно словарю соответствующей ПО и грамматики, лингвистической модели (LM). Затем выбирается n лучших последовательностей слов и сравнивается с нагенерированными моделями предложений, которые могут задавать соответствующий тип предложения (далее ТП). Используя вероятностное оценивание, принимается решение о принадлежности распознанной последовательности слов к ТП. По этому ТП определяется тип смысла (далее ТС) и интерпретатор находит соответствующий ТС на другом языке. На выходе мы должны получить текст на языке 2, который озвучивается соответствующей системой озвучивания текстов на языке 2.

В описанной структуре перевода остаётся достаточно сложная задача интерпретации распознанного сигнала. Пути решения этой задачи, описанные в [2], основываются на том, чтобы научиться экономно задавать все допустимые предложения в языке диалога.

Таким образом, автоматический перевод фразы, произнесённой на языке 1, на язык 2 с озвучиванием результата, с помощью предлагаемой структуры устного перевода будет заключаться в том, чтобы сначала для сигнала, который произносится диктором, найти наиболее правдоподобный, возможно, параметризованный, ТП среди всех ТП, задающих ТС. Затем определить сам ТС произнесённого содержательного высказывания и найти для него подходящий ТС в языке 2 с учётом параметров. Наконец, предложение, полученное на языке 2, озвучивается.

Моделирование типов предложений с учётом параметров

Поскольку структура перевода должна работать в рамках ПО, то предлагается рассмотреть определённую иерархию речевых сигналов [2]. Подразумевается, что вся деятельность человека разбивается на ПО по аналогии с бумажным разговорником. Каждая ПО состоит из конечного множества ТС.

В каждый ТС входит множество эквивалентно содержательных ТП. ТП — конструкция, экономно задающая множество предложений, полученных из одного предложения независимыми допустимыми заменами и допустимыми перестановками или выпадением слов и словосочетаний.

В рамках задачи распознавания, интерпретации и перевода речевого сигнала немаловажен вопрос описания параметров слов во фразах, где могут быть разные варианты имён собственных, времени, адресов и т.д. Значение термина «параметр» может иметь разную интерпретацию в зависимости от контекста. В общем параметром называют величину, значения которой служат для различия элементов некоторого множества между собой.

Рассмотрим пример ТП «просьба разбудить человека в определённое время», из ПО «Гостиница». Базовая структура будет иметь вид:

$$\left(\text{(разбудите)} \left(\begin{array}{c} \text{меня} \\ \text{нас} \\ * \end{array} \right) \left(\text{пожалуйста} \right) \left(\begin{array}{c} \text{в } \$time : \text{app} \\ \left[\begin{array}{c} \text{в} \\ \text{через} \end{array} \right] \$time \end{array} \right) \right)$$

В круглых скобках () указаны подсловари, которые можно переставлять местами, а в квадратных [] — которые нельзя переставлять. Символ * означает пустое слово.

Этой структурой можно сгенерировать много предложений с учётом параметров. Среди этих предложений будут, например, и такие:

Разбудите меня, пожалуйста, в семь часов.

Пожалуйста, нас разбудите в пять утра.

В семь тридцать разбудите меня.

Разбудите нас в шесть.

Разбудите меня через шесть часов.

Стоит отметить, что предложения разговорной речи тоже необходимо учитывать.

В этом примере параметр — «временное предназначение»: \$time:app, \$time. Рассмотрим первый параметр \$time:app. Он описывает любое время с точностью до, например, минут, в контексте определённого события. Чтобы предусмотреть все варианты и значения этих параметров вводится специально разработанная и описанная параметрическая грамматика словаря на основе формы Бекуса — Наура (BNF). Такую грамматику можно подать в развёрнутом виде (таблицы 1–2).

В приведённом примере базовая структура задаёт $4! \cdot 1 \cdot 3 \cdot 2 \cdot 3 = 432$ параметризованных предложений, допустимых в языке диалога. Если учесть, что каждый параметр содержит большое количество вариантов, то количество предложений значительно увеличится.

Таблица 1
Базовые структуры параметров временного предназначения

Обозначения	Пример	Параметризация для русского языка
\$time:app	в шесть	\$hour:nadj-at
	в шесть тридцать	\$hour:nadj-at [\$teen:n \$dec:5max]
	в шесть часов	\$hour:nadj-at \$hour-i
	в шесть часов утра	\$hour:nadj-at \$hour-i \$time:post
	в шесть тридцать утра	\$hour:nadj-at [\$teen:n \$dec:6max] \$time:post
	в шесть час. тридцать мин.	\$hour:nadj-at \$hour-i \$min:n-u
	в шесть часов тридцать минут утра	\$hour:nadj-at \$hour-i \$min:n-u \$time:post



Таблица 1 (окончание)

\$min:n-u	одна минута	\$min:n1 \$min1
	две минуты; 53 минуты	\$min:n2 \$min2
	5 минут; 37 минут	\$min:n5 \$min5
\$min:n5	20; 45	\$dec:5max [\$digit5]
	5; 7	\$digit5
	12; 15	\$teen:n

Все предложения языка диалога можно задавать с помощью ТС и соответствующих им ТП, используя структуру, приведённую в примере. С помощью LISP-структур генерируется огромное количество предложений, имеющих одинаковый смысл. Поскольку построение LISP-структур довольно громоздкое, требует много ручной работы, то был разработан автоматизированный спецификатор ПО.

Таблица 2

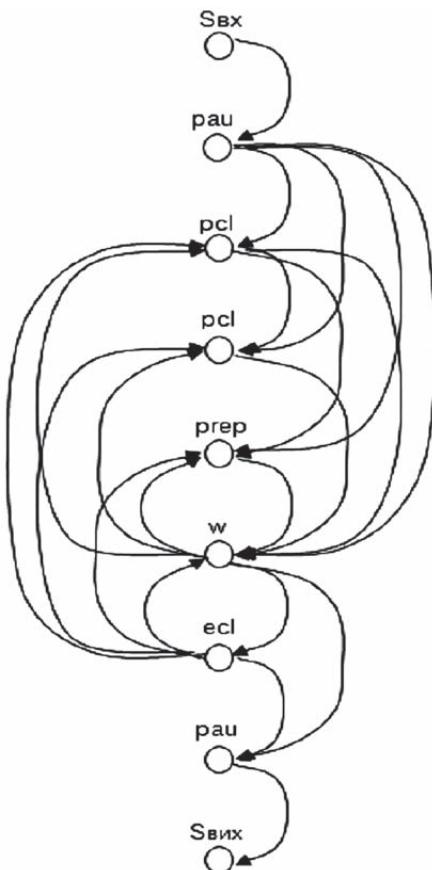
Описание значений параметров для временного предназначения

\$hour:nadj-at	первый	one	\$dec:50max	двадцать	twenty
	второй	two		тридцать	thirty
	...			сорок	forty
	двадцать третий	twenty three		пятьдесят	fifty
\$hour-i	часа	o'clock	\$digit5	пять	five
\$min:n1	одна	one		...	
	...			девять	nine
	пятьдесят одна	fifty one		\$min1	минута
\$min:n2	две	two	\$min2	минуты	minute
	три	three		\$min5	минут
	...				утра
	пятьдесят три	fifty three		\$time:post	a.m.
\$teen:n	десять	ten		дня	p.m.
		вечера	p.m.
	девятнадцать	nineteen		ночи	a.m.

Для построения всех возможных предложений языка устного диалога можно использовать так называемую ориентированную семантическую сеть (далее ОСС) [1, 2], одновременно задающую ограниченную грамматику порядка следования слов.

Альтернатива этой грамматики — грамматика свободного порядка следования слов. Между этими противоположными, по сути, грамматиками может быть построено множество других относительно свободных или относительно ограниченных грамматик. Мы предлагаем несколько ограничить свободную грамматику за счёт лингвистического понятия о фонетическом слове [2].

Под «фонетическим словом» понимаем слово с неотделимыми от него сопутствующими словами. Например, неотделимыми являются предлог от существительного или прилагательного, частица «не» перед глаголом и частица «б» после него. Предлагаемая нами, относительно свободная грамматика, представлена в виде графа (рис. 2), где *rau* — слово-пауза в начале и в кон-



це фразы, *pcl* — проклитик, *prep* — предлог, *w* — нейтральное слово, *ecl* — энклитик.

Впрочем, при такой грамматике принятие решений относительно смысла предложения неочевидно.

Рис. 2. Граф относительно свободной грамматики на основе понятия про лингвистическое слово

Статистическое оценивание принадлежности последовательности слов к типу предложения

При распознавании в условиях грамматики, которая не задаёт строгих ограничений на последовательности слов, очевидно, могут быть получены ответы распознавания, не входящие во множество предложений, которые сгенерированы определённым ТП. Это может быть обусловлено как ошибками при распознавании, так и при формировании ТП экспертом. Кроме того, сам пользователь может произнести предложение с различного рода отклонениями или аграмматизмами, например, повторить некоторое слово дважды.

Поэтому предлагается оценивать вероятность типа предложения *ST* с ОСС при распознанной последовательности слов (w_1, w_2, \dots, w_n) и объявлять ответом интерпретации тот тип предложений *ST**, для которого эта вероятность является наибольшей:

$$ST^* = \underset{ST}{\operatorname{argmax}} P(ST / w_1, w_2, \dots, w_n), \quad (1)$$

Вероятность в левой части (1) может быть записана также по формуле Байеса в следующем виде:

$$P(ST / w_1, w_2, \dots, w_n) = \frac{P(ST)}{P(w_1, w_2, \dots, w_n)} P(w_1, w_2, \dots, w_n / ST), \quad (2)$$

Рассматривая последовательность (w_1, w_2, \dots, w_n) как Марковский процесс, отображаем каждый из множителей условной вероятности в правой части (2) в виде:



$$P(w_1, w_2, \dots, w_n / ST) = \prod_{k=1}^n P(w_k / ST, w_{k-m}, \dots, w_{k-1}), \quad (3)$$

$$P(w_1, w_2, \dots, w_n) = \prod_{k=1}^n P(w_k / w_{k-m}, \dots, w_{k-1}), \quad (4)$$

где $m \geq 0$ — порядок процесса.

Оценивание каждого из множителей правой части выражений (3) и (4) может производиться различными способами в зависимости от выбранного порядка процесса.

Мы рассматривали наиболее простой случай, когда $m = 0$. Тогда, учитывая формулу Байеса, выражение (2):

$$P(ST / w_1, w_2, \dots, w_n) = P(ST) \prod_{k=1}^n P(ST / w_k). \quad (5)$$

Логично сделать предположение относительно равной вероятности всех типов предложений. В действительности, некоторые смыслы встречаются чаще других. Это зависит от предыдущего смысла (контекста). Остается рассчитать выражение вида $P(ST / w_k)$. Для этого рассмотрим $ST(w_k)$ — множество типов предложений, в которых встречается слово w_k . Тогда:

$$P(ST/w_k) = \begin{cases} |ST(w_k)|^{-1}, & \text{если } ST(w_k) \cap ST \neq \emptyset, \\ \alpha(ST, w_k), & \text{иначе.} \end{cases} \quad (6)$$

Выражение $\alpha(ST, w_k)$ отображает смысл вероятности того, что слово w_k распознано ошибочно вместо некоторого слова w : $ST(w) \neq \emptyset$. Эту вероятность можно оценить на основе некоторой меры минимальной редакторской правки $d(w_k, w)$, например, расстояния Левенштейна. При вычислении этой меры штрафуются вставки, удаления и замены символов фонемного текста сравниваемых слов. Таким образом, выражение $\alpha(ST, w_k)$ предлагается оценивать как:

$$\alpha(ST, w_k) = \max_{ST(w) \neq \emptyset} \left(\max \left\{ 1 - \frac{d(w_k, w)}{L(w)}, 0 \right\} \times |ST(w)|^{-1} \right), \quad (7)$$

где $L(w)$ — количество фонем в слове w .

Решение относительно принадлежности распознанной последовательности слов некоторому ТП принимается на основании (1) — (7).

В случае, когда распознанная последовательность слов при таком оценивании совпадает с определённым ТП, принятие решения очевидно. Но может быть так, что некоторые слова распознались ошибочно. Такое предложение можно отбросить, не найдя для него соответствующий ТП. А можно попробовать оценить, к какому ТП ближе распознанная последовательность слов. И определить гипотетический ТП, т.е. который можно объявить ответом интерпретации.

Рассмотрим это на примере. Допустим последовательность распознанных слов.

$(w_1, w_2, w_3) = \text{Допоможіть було маска}$

Обозначим ST_1 — ТП, к которому эта последовательность будет ближе всего.

$ST_1 = \text{Допоможіть мені будь ласка}$

Оценим вероятность принадлежности распознанной фразы (w_1, w_2, w_3) к ТП ST_1 . Воспользуемся формулами (5) — (6).

$$\begin{aligned} P(ST_1 / \text{допоможіть}, \text{було}, \text{маска}) &= P(ST_1)P(ST_1 / w_1)P(ST_1 / w_2)P(ST_1 / w_3) \cong \\ &\cong 1 \cdot \frac{1}{|ST(w_1)|} \cdot \alpha(ST, w_2) \cdot \alpha(ST, w_3). \end{aligned}$$

Поскольку мы предположили, что все ТП имеют одинаковую вероятность, то $P(ST_1) = 1$. Далее нужно подсчитать значение каждого множителя. Слово $w_1 = \text{«допоможіть»}$ распозналось правильно, поэтому

$$P(ST_1 / w_1) = \frac{1}{|ST(w_1)|} = 1.$$

Слова «було» и «маска» явно не принадлежат ST_1 . Мы предполагаем, что эти слова являются гипотетическими ошибками распознавания.

Чтобы подсчитать вероятность принадлежности распознанной последовательности слов (w_1, w_2, w_3) ТП ST_1 , сформируем множество слов из ST_1 , которые там остались без учёта правильно распознанного слова $w_1 = \text{«допоможіть»}$. Получится следующее множество: $w(ST_1) \setminus w_1 = \{\text{менi, будь, ласка}\}$.

Как было упомянуто выше, оценивать вероятности ошибочно распознанных слов w_k будем на основе меры Левенштейна — d , а именно используя формулу (7). Нам нужно посимвольно сравнить фонемный текст каждого гипотетически ошибочно распознанного слова $(w_2, w_3) = \{\text{було, маска}\}$ с множеством слов $w(ST_1) \setminus w_1 = \{\text{менi, будь, ласка}\}$. Вводится следующая система штрафования: вставки, удаления, замены символов фонемного текста штрафуются одним балом, а совпадение фонем не штрафуется. На рис. 3 представлены результаты сравнения фонемного текста и подсчитаны минимальные расстояния меры Левенштейна.

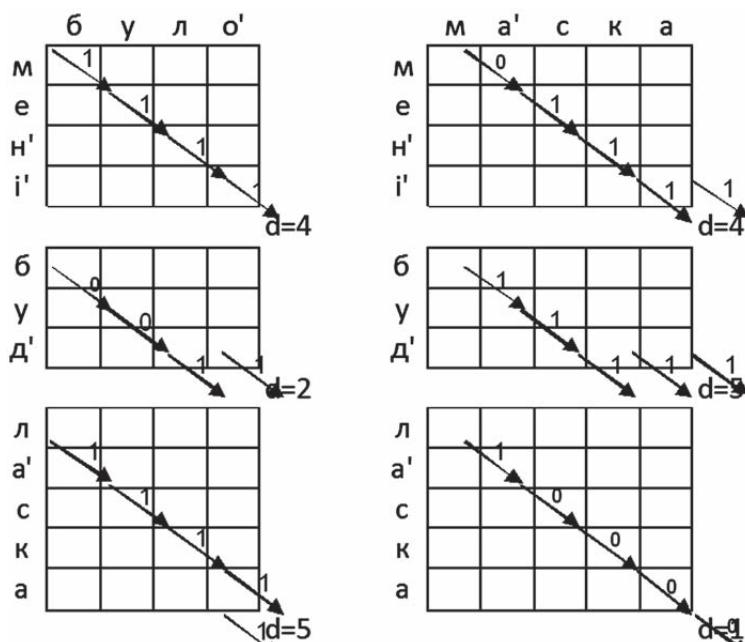


Рис. 3. Сравнение символов фонемного текста



Далее для слова $w_2 = \text{«було»}$ в развернутом виде описан подсчёт вероятности по формуле (7):

$$\begin{aligned} P(ST_1 / w_2) &= \max_{ST(w) \neq \emptyset} \left(\max \left\{ 1 - \frac{d(w_2, w)}{L(w)}, 0 \right\} \times |ST(w)|^{-1} \right) = \\ &= \max_{ST(w) \neq \emptyset} \left\{ \max \left\{ 1 - \frac{d(w_2, \text{'мені'})}{L(\text{мені})}, 0 \right\} \times |ST(\text{мені})|^{-1}, \max \left\{ 1 - \frac{d(w_2, \text{'будь'})}{L(\text{будь})}, 0 \right\} \times |ST(\text{будь})|^{-1}, \right. \\ &\quad \left. \max \left\{ 1 - \frac{d(w_2, \text{'ласка'})}{L(\text{ласка})}, 0 \right\} \times |ST(\text{ласка})|^{-1} \right\} = \\ &= \max_{ST(w) \neq \emptyset} \left\{ 0, \frac{1}{3} \times |ST(\text{будь})|^{-1}, 0 \right\}. \end{aligned}$$

Аналогично для слова $w_3 = \text{«маска»}$:

$$P(ST_1 / w_3) = \max_{ST(w) \neq \emptyset} \left\{ 0, 0, \frac{4}{5} \times |ST(\text{ласка})|^{-1} \right\}.$$

Мы видим, что для каждого из этих слов существует ненулевая вероятность того, что они могут принадлежать ТП ST_1 .

Таким образом, вероятность того, что распознанная фраза $(w_1, w_2, w_3) = \text{«Допоможіть було маска»}$ принадлежит ТП ST_1 будет:

$$P(ST_1 / \text{допоможіть, було, маска}) \equiv 1 \cdot \frac{1}{3|ST(\text{будь})|} \cdot \frac{4}{5|ST(\text{ласка})|} \leq \frac{4}{15}.$$

Подсчитав окончательно по формуле (5) вероятность принадлежности распознанной фразы к предполагаемому ТП, мы видим, что гипотеза данного ТП не отбрасывается и при отсутствии других гипотез может быть ответом интерпретации.

Экспериментальные результаты

Предложенные в работе методы оценивания принадлежности последовательности слов к ТП были экспериментально проверены на фразах из обычного разговорника. В работе для примера рассматривались три ПО: «Повседневные фразы», «Путешествие», «Гостиница». Эти ПО содержат $47 + 102 + 68 = 217$ ТС. В среднем на ТС приходится 4,17 базовых предложений.

Акустические модели для декодера разработаны на основе речевого корпуса отдельно произнесённых слов, в котором принимали участие 60 дикторов [2]. Средствами [3] проведено обучение 55 скрытых Марковских моделей фонем. Максимальное количество нормальных законов в смеси — 20.

Для эксперимента произвольным образом было выбрано 500 фраз. Смысловая интерпретация проводилась на основе результата пофонемного распознавания речевых сигналов [4] в условиях свободной и относительно свободной (на основе фонетических слов) грамматик относительно слов [2]. Из результатов проведённого эксперимента (таблица 3) следует, что для двух типов грамматик отклонение смысловой интерпретации не превышает 5%, что является приемлемым для прикладной системы.

В условиях ограниченной грамматики скорость распознавания в 10 раз превышает реальное время, а в условиях свободной и относительно свободной грамматик распознавание происходит быстрее реального времени на ресурсах нетбука.

Таблица 3

Результаты распознавания и смысловой интерпретации 500 предложений из двух предметных областей

Тип грамматики	Надёжность распознавания (%)		
Ограниченнaя	96,7	94,1	98,3
Свободная пословная	53,4	4,2	86,1
Относительно свободная	79,1	20,8	96,2

На основе проведённых исследований разработана демонстрационная программная модель для перевода произнесённых предложений с русского языка на английский (рис. 4). При этом последовательность слов в русском предложении может быть любой из допустимых. Предложению, произнесённому на русском языке, ставится в соответствие англоязычный ТС или ТП, а первое предложение этого ТС объявляется результатом перевода.

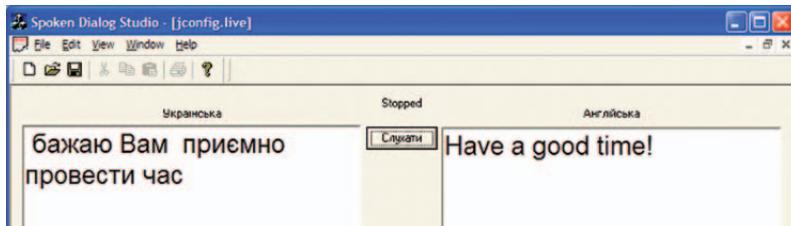


Рис. 4. Демонстрационное программное обеспечение модели устного фразаря-переводчика

Выводы

Рассматриваемая в работе система устного перевода является электронным аналогом бумажного разговорника, взаимодействие с которым происходит наиболее естественным способом — голосом. При распознавании произнесённой пользователем фразы используются лингвистические и семантические знания по выбранной ПО. Введённые при этом «мягкие» ограничения на порядок следования слов позволяют повысить надёжность распознавания, не повышая требований к вычислительным ресурсам. Разработанное программное обеспечение даёт возможность формировать грамматики следования слов для распознавания слитной речи как на основе ТП, так и на основе лингвистического понятия «фонетическое слово».

Использование параметрических моделей ТП даёт возможность пользователю более свободно и разнообразно общаться, что расширяет сферу использования системы. Предположение, что наблюдаемые последовательности слов — Марковский процесс, дало возможность сформулировать более гибкий способ формирования результата смысловой интерпретации.

На основе экспериментальной модели разработана программная модель устного словаря-переводчика, для перевода с русского языка на английский в рамках предметной области, которая работает в режиме реального времени на ограниченных вычислительных ресурсах.

Одни и те же фразы, произнесённые с различной интонацией, могут выражать как вопросительное предложение, так и повествовательное. Итак, в дальнейшей работе следует исследовать возможность распознавания интонации и ритма (просодики) с целью автоматической расстановки знаков препинания в распознанных фразах.



Яценко В.В.

ПАРАМЕТРИЗАЦИЯ ТИПОВ ПРЕДЛОЖЕНИЙ ПРЕДМЕТНОЙ ОБЛАСТИ ДЛЯ СИСТЕМЫ УСТНОГО ФРАЗАРЯ-ПЕРЕВОДЧИКА

В дальнейшем также планируется ставить в соответствие русскоязычной фразе более точный англоязычный аналог среди ТП по ТС.

Литература

1. Vintsiuk T.K. Analysis, Recognition and Understanding of Speech Signals, Kyiv: Naukova Dumka, 1987.
2. Sazhok M., Yatsenko V. Spoken translation system based on speech understanding in subject area // All-Ukrainian Int. Conference on Signal/Image Processing and Pattern Recognition UkrObraz'2010. Kyiv, 2010. P. 103–106.
3. Lee, Kawahara T. and Shikano K. Julius — an open source real-time large vocabulary recognition engine. — In Proc. European Conference on Speech Communication and Technology (EUROSPEECH). 2001.P.1691–1694.
4. Young S.J. et al. HTK Book, version 3.1, Cambridge University. 2002.

Сведения об авторах

Яценко Валентина Витальевна —

работает в Международном научно-обучающем центре информационных технологий и систем в отделе распознавания и синтеза речевых сигналов. Занимается формированием словарей, фраз для словарей-разговорников и интерпритацией распознанных фраз и переводом их на другой язык. Киев. val-yatsenko@yandex.ru

Преобразование между орфографическим и фонемным текстами для моделирования спонтанного произношения

Робейко В.В., научный сотрудник

*Сажок Н.Н., кандидат технических наук,
старший научный сотрудник*

В статье рассматривается актуальный вопрос порождения последовательностей фонем по орфографическому тексту (транскриптор) и обратное преобразование. Основная цель работы — разграничение оперативного кода (команд) и данных, что позволит превратить транскриптор в гибкий и удобный инструмент исследователя. Предложена модель, в которой заложены закономерности связи между фонетическими и орфографическими символами. Многозначные преобразования, осуществляемые согласно модели, эквивалентны построению направленного графа. При переходе вдоль стрелок графа осуществляется генерирование последовательностей фонетических символов, которые соответствуют исходному тексту. Параметры модели задаются таблично в виде контекстно-зависимых правил, которые формирует эксперт. В модели предусмотрено введение дополнительных уровней транскрибирования, что позволяет эксперту строить сложные преобразования, работая с относительно простыми правилами. Разработана система, на вход которой подаётся орфографический текст с обозначениями ударений, а на выходе получаем последовательности фонем, соответствующих различным вариантам произношения исходного текста. Практическое применение системы для автоматического распознавания спонтанной украинской речи позволяет увеличить надёжность распознавания для ряда случаев. В заключение обсуждаются сопутствующие задачи и перспективные исследования.

• *графема • фонема • преобразование • транскрибирование • спонтанная речь.*

In this paper we consider the actual problem of grapheme-to-phoneme (GTP) conversion for forward and backward directions. The main purpose is to separate the operational code (instructions) and the data that is the way to obtain the flexible and convenient GTP tool for the researcher. We propose a model describing regularities of relations between the phonetic spelling and symbols. Multi-decision transformations carried out according to the model, are equivalent to building a directed graph. Moving along the arrows we can generate multiple phoneme sequences corresponding to the input text. The model parameters are specified in tabular form as a set of context-dependent rules formed by the expert. We introduce additional model levels allowing the expert to build complex transformations, still working with relatively simple set of rules. The developed system takes the orthographic text with the pointed accentuation and produces phoneme sequences corresponding to the different pronunciation ways and manners of the input text. Practical applications of the system for automatic recognition of Ukrainian spontaneous speech showed increasing accuracy for certain cases. Finally, we discuss related tasks and further research.

• *grapheme • phoneme • conversion • transcription • spontaneous speech.*



Порождения последовательностей фонем по орфографическому тексту и последовательностей слов по фонемному тексту — актуальная проблема в области речевой информатики. Графемно-фонемные преобразования нужны для транскрибирования текстов в системах озвучивания информации, для формирования словарей произношения при оценке параметров акустической модели в различных схемах декодирования речевого сигнала и т.п. [1–4].

Преобразование орфографического текста (графем) в фонемный текст невозможно без изучения закономерностей связи между фонетическими и орфографическими символами.

Долгое время графемно-фонемное преобразование реализовывалось в виде определённого алгоритма — транскриптора, воспроизводящего в программном коде правила чтения из учебника или справочника [6].

Нужно отметить, что, в отличие от орфографического, фонемный текст для каждого определённого слова не является устойчивым как по субъективным, так и по объективным причинам.

Во-первых, до сих пор среди исследователей в области фонетики и фонологии продолжаются дискуссии относительно алфавитов фонем для языков и диалектов (несмотря на принятый стандарт Международной фонетической ассоциации — IPA). Во-вторых, алфавит фонем для системы распознавания или синтеза речи зависит не только от свойств языка, но и от того, какое фонетическое явление (ассимиляция, редукция, палатализация и др.) учитывается на уровне символов, а какое — на уровне акустической модели путём введения контекстной зависимости (CD-фонемы, фонемы-тригоны), использование смеси гауссиандов (GMM) или изменения топологии акустической модели HMM. В-третьих, при разработке речевых технологий должны учитываться индивидуальные, ситуативные особенности произношения, произношение слов в потоке речи, словарные и смысловые ударения. Это влечёт введение многозначности при переходе к фонемному тексту.

Отсутствие разграничения между оперативным кодом (командами) и данными — серьёзное препятствие для модификации и усовершенствования транскриптора и делает невозможным его превращение в гибкий инструмент исследователя.

Заметим, что в мире для преобразования графем в фонемы распространены статистические методы, которые анализируют созданный экспертами словарь произношений [5]. Это оправдано для языков, в которых орфографическое написание слов существенно отличается от их произношения (например, для английского или французского). К недостаткам такого подхода можно отнести игнорирование возможной коартикуляции на стыках слов и субъективизм экспертов.

Задача обратного перехода от фонемного текста к орфографическому является относительно новой. Она приобрела актуальность с развитием многозначной многоуровневой модели понимания речевого сигнала [4]. Реализованному в то время алгоритму подавалось на вход менее 30 правил преобразования «фонема — графема» для украинского языка, в результате чего генерировались гипотетические варианты орфографического текста для многозначного ответа распознавания свободного порядка следования фонем. Такое незначительное количество правил объясняется не только свойствами украинской орфографии, в которой правила написания главным образом основываются на фонетическом принципе (пишем то, что слышим). На выходе фонемно-графемный преобразователь дополняется

лексическим фильтром, который отбрасывает недопустимые гипотезы орфографического написания.

Дальнейшие исследования алгоритма заключались в его использовании при обратном процессе — преобразовании орфографического текста в фонемный. При этом ставилась цель сохранения прозрачности и относительной простоты правил для эксперта и одновременного расширения подхода на языки, у которых отличие написания от произношения весьма существенно, и таким образом получить универсальный транскриптор, который можно сравнительно легко модифицировать на уровне структур данных без дополнительной компиляции программного кода.

Модель многозначного преобразования последовательностей символов

Пусть задана некоторая конечная последовательность символов

$$(a_1, a_2, \dots, a_n, \dots, a_N) \equiv a_1^N, a_n \in \mathbf{A}, \quad (1)$$

где \mathbf{A} — алфавит входящих символов. Сконструируем отображение этой последовательности на множество последовательностей выходящих символов из некоторого иного алфавита \mathbf{B} .

Рассмотрим функцию f , отображающую последовательность a_1^N , начиная с её n -го символа, в символ алфавита \mathbf{B} или пустое множество:

$$f : a_n^N \rightarrow b, b \in \mathbf{B} \cup \emptyset, 1 \leq n \leq N. \quad (2)$$

Заметим, что (2) справедливо лишь в случае, когда входящая последовательность принадлежит области определения f , т.е. $a_n^N \in \text{Def}(f)$. Множество последовательных применений таких функций переводит a_n^N в последовательности символов из алфавита \mathbf{B} , и таким образом мы конструируем мультифункцию:

$$F(a_n^N) = \left\{ f_1^k(a_n^N), f_2^k(a_n^N), \dots, f_{L_k}^k(a_n^N) \right\} \in \mathbf{B}^{L_k} \cup \emptyset, 1 \leq k \leq K_F, \quad (3)$$

где L_k — длина k -й выходящей последовательности, общее количество которых K_F , своё для каждой $F \in \mathcal{F}$.

Определим аналог прямого произведения над множествами, полученными вследствие действий мультифункций из \mathcal{F} , как перебор всех вариантов объединения конечных последовательностей символов алфавита \mathbf{B} , т.е. опуская аргументы мультифункций:

$$F \otimes G = \left\{ (f_1^u, f_2^u, \dots, f_{L_u}^u, g_1^v, g_2^v, \dots, g_{L_v}^v), 1 \leq u \leq K_F, 1 \leq v \leq K_G \right\}. \quad (4)$$

Допускаем по определению, что если результат действия F или G является пустым множеством, то результатом их произведения будет пустое множество. В отличие от декартового произведения для определённого нами аналога выполняется свойство ассоциативности.

Рассмотрим упорядоченное множество $\tilde{\mathbf{F}}$ мультифункций $F \in \mathcal{F}$, которые сопроводим дополнительными параметрами:

$$\tilde{\mathbf{F}} = \left(F_{i, d_i, \delta_i} \right), 1 \leq i \leq |\tilde{\mathbf{F}}|, d_i > 0, \delta_i = \{0, 1\}, \quad (5)$$

где i — индекс мультифункции в упорядоченном множестве $\tilde{\mathbf{F}}$; параметр d_i — ширина шага анализа, δ_i — условие исключительности. Через эти параметры формулируем ограничения при вычислении произведения

$$\bigotimes_{i,n} F_{i, d_i, \delta_i} (a_n^N), 1 \leq i \leq |\tilde{\mathbf{F}}|, 1 \leq n \leq N. \quad (6)$$

Предположим, что мы уже вычислили выражение (6) на некоторых упорядоченных индексных множествах J и M и получили некоторое непустое множество

$$G_{J, M} = \bigotimes_{u \in J, v \in M} F_{u, d_u, \delta_u} (a_v^N). \quad (7)$$



Пускай j и m являются последними элементами индексных множеств J и M соответственно. Тогда при рассмотрении следующего компонента произведения, $F_{i,d_i,\delta_i}(a_n^N)$, проводим вычисления согласно с определением (4), если выполняются такие условия:

$$\begin{cases} m + d_j = n; \\ \delta_r \neq 1, 1 \leq r < i; \\ \bigotimes_{u \in J, v \in M} F_{u,d_u,\delta_u}(a_v^N) \otimes F_{r,d_r,\delta_r}(a_n^N) \neq \emptyset, 1 \leq r < i, \text{ если } \delta_i = 1. \end{cases} \quad (8)$$

В противном случае, при поступлении следующего компонента произведения получаем пустое множество.

Выражением (6) порождаются последовательности выходящих символов по некоторой последовательности входящих символов. Если исходный алфавит совпадает с алфавитом букв определённого языка, а выходящий алфавит состоит из фонем, то получаем многозначный транскриптор орфографического текста. И наоборот, если на входе — фонемный алфавит, а на выходе — алфавит букв, то получаем многозначное преобразование из фонемного текста в орфографический. Возможны промежуточные варианты.

Пример порождения реализаций фонемного текста по орфографическому тексту приведён на рис. 1. Рассматривается орфографический текст слова «снег». С целью обобщения стандартный алфавит дополнен символом «_», который разделяет слова. Соответствующие ударным гласным буквы переводятся в верхний регистр, все остальные буквы — в нижний. Позиция ударения определяется по орфоэпическому словарю с учётом омографии [10] или же учитываются все допустимые позиции ударения.

Таким образом, имеем на входе последовательность из шести символов $a_1^N = (\langle _ \rangle, \langle c \rangle, \langle h \rangle, \langle E \rangle, \langle \varepsilon \rangle, \langle _ \rangle)$, $N = 6$. На графе отображены все допустимые мультифункции $F_{i,d_i,\delta_i}(a_n^N)$, $1 \leq n \leq N$. Производя переход по стрелкам, получаем произведения вида (6), генерирующие четыре последовательности фонем или фонемных текста:

$\langle _ c h' E \varepsilon _ \rangle; \langle _ c h' E \kappa _ \rangle; \langle _ c' h' E \varepsilon _ \rangle; \langle _ c' h' E \kappa _ \rangle$.

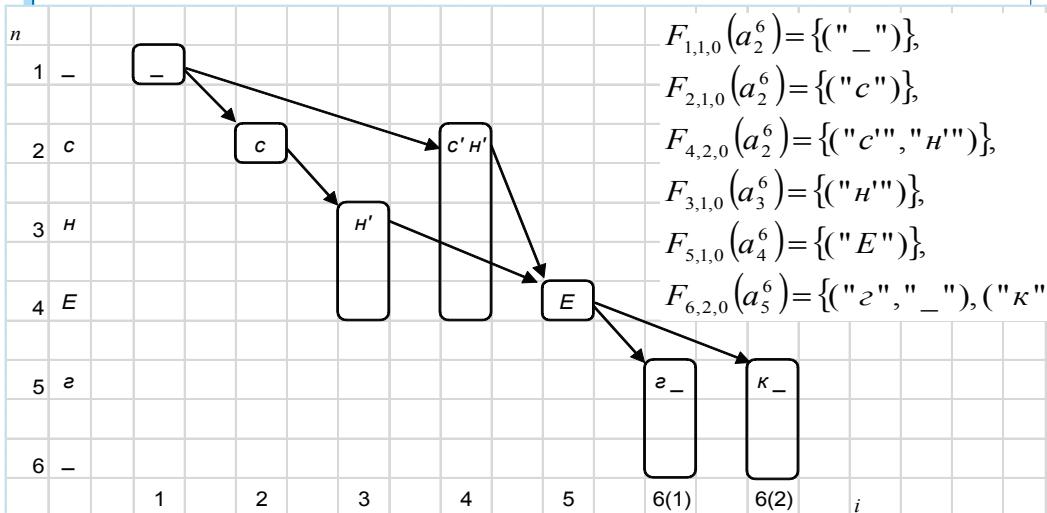


Рис. 1. Граф многозначного преобразования орфографического слова «снег» в фонемную транскрипцию

Заметим, что на практике используется не вся подпоследовательность вида a_n^N , а лишь её часть, a_n^{n-1+T} , где положительный параметр T зависит от конкретной мультифункции и определяет ширину контекста, в котором рассматриваются входящие символы. На рис. 1 высота прямоугольников в узлах графа соответствует ширине контекста.

Результат преобразования может быть многократно подвергнут описанным преобразованиям с теми же или другими параметрами. Таким образом мы можем вводить дополнительные уровни транскрибирования, постепенно переходя к выходным символам. Это позволяет существенно упростить спецификацию параметров, что важно для языков, в которых написание во многом отличается от произношения. Также появляется возможность применить идентичный алгоритм с другим набором правил к (многозначной) расшифровке чисел и сокращений и транслитерации.

Система многоуровневого многозначного транскрибирования орфографических текстов

Предлагается система, на вход которой подаётся орфографический текст, содержащий только символы из алфавита букв, включая символы границы между словами и морфемами и обозначениями ударения. Такой текст получаем вследствие автоматической обработки произвольного текста или списка слов, которая заключается в вербализации (расшифровке) символов, чисел и сокращений, расстановке ударений и разделения на синтагмы, слова и морфемы [6]. На выходе системы получаем последовательности фонем, соответствующие различным вариантам произношения входного текста.

Эта система транскрибирования используется для распознавания украинской речи [8, 11].

Разработанная система многозначного транскрибирования орфографических текстов использует модель (1) — (8), которая предусматривает возможность таблично задавать контексто-зависимые правила преобразований одних обобщённых последовательностей символов на другие. При этом в каждом правиле задаётся ширина шага, по которому происходит переход к следующей последовательности символов. Применение многих правил позволяет генерировать сразу несколько вариантов транскрипции одного и того же слова или генерировать нужный вариант из нескольких возможных, например, описывая спонтанную речь говорящего или группы дикторов (таблица 1).

Таблица 1
Примеры построения правил транскриптора для украинской речи

Входящая обобщённая подпоследовательность символов	Выходящие символы	Ширина шага	Пояснения
[зсц] [жшч]	[жшч]	1	з, с, ц перед ж, ш, ч переходят соответственно в ж, ш, ч
[тс][дтзснц][іюяїЇЮЯ]	т'	1	т и с перед мягкими д, т, з, с, н, ц смягчаются
с т [лн]	с	2	т между с и л или н выпадает

В целом для транскрибирования украинской речи (для литературного варианта произношения) достаточно ввести чуть более 30 подобных правил. Для литературного варианта русской речи уже потребовалось вводить уровни, постепенно отходя от орфографического написания к фонемному тексту. Предложено 5 уровней и около 120 правил.

Последующее развитие системы транскрибирования для украинского языка происходило на основе анализа спонтанной речи нескольких сотен дикторов [8, 11]. Для всех дикто-



ров был создан общий вариант транскрибирования на основе литературного произношения. Кроме этого, все дикторы были распределены на группы, для которых разработаны свои правила индивидуализированного транскрибирования, дополняющие или заменяющие основной вариант [9]. Также для одной из групп дикторов был разработан общий вариант транскрибирования спонтанной слитной речи [12].

Результаты изучения спонтанного вещания многих дикторов свидетельствуют о том, что никто из них не соблюдает орфоэпических правил произношения в полном объёме. Прежде всего, это касается запрещённых литературной нормой регressiveйной асимиляции по глухости в паре фонем «звонкая + глухая» и оглушение согласных перед паузой (**тобто → т о п т о; підтримати → п' і т т р И м а т и; робив → р о б И ф**). Дикторы с такими особенностями произношения были собраны в отдельную группу. Было также установлено много других характерных черт произношения разных дикторов: редукция окончаний некоторых слов (прилагательных, глаголов) в слитной речи (**шановний → ша н О в н и; доброго → д О б р о**), «аканье» (**робити → р а б И т и**), твёрдое произношение мягких согласных (**синього → с И н о г о**) и пр.

Для некоторых слов (например, служебных частей речи, слов с разными ударениями) задаётся несколько вариантов транскрипций — с ударением на различных слогах или вообще без ударения: **коли → к о л И; к О л и; к о л и.**

Такие тенденции моделируются путём изменения правил перехода от одних последовательностей символов к другим и расширением действующих правил.

Все правила индивидуализированной модификации транскрипций можно разделить на несколько групп (на основе обработки звуковых записей).

К позиционным, зависящим от общих фонетических условий — позиции в составе / слове, ударности / безударности т.д. [7] нелитературным изменениям звуков в потоке речи относим:

- 1) ослабленное произношение *o* как *a* в безударной позиции, реже встречается редукция безударных гласных до полного исчезновения (**тепер → т и п Е р, зозуля → з у з У л' а, боротьба → б а р а д' б А или б р а д' б А**);
- 2) оглушение звонких согласных перед паузой (**брід → б р' I т, зараз → з А р а с**);
- 3) редукцию в терминальных частях слов в процессе произношения — исчезновений согласного звука в окончаниях — *ого*, — *их*, — *ич*, — *ий*, — *их*, — *ий*, — *иї*, — *ої*, — *еї*, — *ою*, — *єю*, — *ити* и подобных (**коротший → к о р О ч ш и, Іванович → і в А н о в и, синіх → с И н' i, безпекою → б е с п Е к о у**); исчезновений конечного гласного звука в окончаниях — *ою*, — *єю*, — *ю* и подобных (**доброю → д О б р о й, землею → з е м-л Е й**) и пр.

К комбинаторным (качественные и количественные изменения соседних звуков [7]) нелитературным изменениям относим:

- 1) полную регressiveйную асимиляцию по глухости в сочетании „звонкий+глухой” на границе любых морфем в слове и на границе слов (**без причини → б е с п р и ч И н и, розсунути → р о с с У н у т и, книжка → к н И ш к а, сядьте → с' А т' т е**);
- 2) асимиляцию по мягкости свистящих и шипящих согласных, губных и заднеязычных согласных (**злі → з' л' I, шлях → ш' л' А х, квітка → к' в' I т к а**);
- 3) произнесение удлинённого согласного звука как обычного неудлинённо-

го, произнесение двух гласных как одного звука (**віддати** → **в' і д А т и**, **знання** → **з н а н' А**, **зоопарк** → **з о п А р к**, **аэропорт** → **а р о п О р т**);

4) неполное упрощение в группах согласных, его отсутствие (**чесний** → **ч Е с т н и й**) и пр.

Для создания индивидуализированных словарей произношения все исследуемые дикторы были распределены на группы. Это происходило в несколько этапов: первый предусматривал выделение группы дикторов с литературной речью и дикторов с отклонениями от литературной нормы; следующим участком работы было выделение общих черт речи во второй группе и создание соответствующих транскрипций для каждой из подгрупп. Таким образом было сгенерировано 18 видов транскрипций словаря для тех или иных дикторов. Адекватность индивидуализированных транскрипций проверялась с помощью распознавания речи (улучшение / ухудшение результатов распознавания для конкретного диктора по сравнению с результатами распознавания для этого же диктора с использованием литературной транскрипции). Каждая из транскрипций была проверена на всех дикторах (таблица 2).

Таблица 2
Примеры индивидуализированных словарей спонтанной речи

Литературная транскрипция	Индивидуализированный словарь	devocal	end_cons	а	devocal _ а
робИв	робИф	робИ	рабИв	рабИф	
вОрог	вОрох	вОро	вОраг	вОрах	

Для исследований произношения и для обучающей выборки при распознавании использовались записи заседаний Верховной Рады Украины продолжительностью 99 тыс. секунд, в которых встретилось более 210 тыс. слов. Всего было записано 208 дикторов. Дикторов с продолжительностью записи более 300 секунд оказалось 87 (запись длительностью менее 300 секунд является слишком короткой для объективных выводов в процессе распознавания) [8].

Для проверки правильности транскрипций и для контрольной выборки при распознавании использовались записи продолжительностью 30 тыс. секунд, в которых встретилось более 68 тыс. слов. Эти записи были сделаны в отличные от обучающей выборки дни. Всего использовались записи 118 дикторов. Дикторов с продолжительностью записи более 300 секунд оказалось 37.

Словарь для генерирования транскрипций был создан из текстов стенограмм заседаний Верховной Рады Украины. С официального сайта Верховной Рады были загружены стенограммы заседаний, начиная с 1991 г. Все тексты стенограмм (более 15 млн. слов) были модифицированы для того, чтобы устраниТЬ лишнюю информацию (например, сообщения об аплодисментах или возгласы), записать числа словами, а также отделить украиноязычный текст от русскоязычного. В результате был получен словарь примерно из 150 тыс. слов, для которого вычислена частота употребления каждого слова. Был составлен словарь на 15 тыс. элементов из наиболее частотных слов (50 и более словоупотреблений), который и стал основой для анализа.

Чтобы выяснить, насколько влияет индивидуализация транскрипции на надёжность распознавания, был проведён ряд экспериментов. Для этого сделана оценка параметров акустической и лингвистической моделей [8] с использованием инструментария HTK [2]. Результаты экспериментов распознавания для некоторых транскрипций (улучшение / ухудшение результатов распознавания для конкретного диктора по сравнению с результатами распознавания для этого же диктора с использованием литературного транскрипции) приведены в таблице 3.



Таблица 3

Образцы результатов распознавания речи дикторов с использованием
индивидуализированных транскрипций

Словарь Диктор \	end_vow_rus	end_vow	end_cons_2	end_cons_1	а	duples
lit	0,16	0,14	-0,01	-0,01	-2,6	0,26
kir	1,37	1,82	0	-1,06	-3,3	0,91
dan	0,71	1,16	-0,09	-0,44	-5,4	0,18

После обработки результатов экспериментов каждому из дикторов приписываются те правила транскрибирования речи, которые повышают надёжность распознавания. Так формируется набор правил для генерирования индивидуализированных транскрипций того или иного диктора.

Как оказалось, часть правил транскрибирования может быть использована для большинства говорящих, например, редукция окончаний вследствие быстрого темпа и эмоциональности речи. Распространённое явление — оглушение, вызванное влиянием русского языка. Значительное количество дикторов избегает произношения удлинённых и удвоенных звуков. Часто встречается слабое произношение ударных гласных (ударение исчезает не только в односложных, двусложных словах, но и в трёхсложных). Другие правила могут быть использованы только для речи одного-двух дикторов.

Генерирование индивидуализированных транскрипций для распознавания речи позволило улучшить надёжность. В дальнейшем планируется увеличить количество индивидуализированных транскрипций за счёт детального анализа речи дикторов, а также автоматически относить дикторов к той или иной группе, учитывая все особенности их произношения.

Общий словарь транскрипций для спонтанной речи был создан для распознавания речи дикторов акустического корпуса украинского эфирного вещания (АКУЕМ) [12]. Он сочетает в себе как правила литературного произношения, так и индивидуальные черты произношения отдельных дикторов, а также очерчивает специфику спонтанной украинской речи (например, задаёт ударные и безударные формы для всех односложных слов и т.д.).

Для исследований произношения и для обучающей выборки при распознавании использовались аудиозаписи спонтанной речи из АКУЕМ продолжительностью около 190 тыс. секунд (более 2000 дикторов). Особое внимание обращалось на речь дикторов с продолжительностью записи более 600 секунд (около 150 дикторов).

Словарь для генерирования транскрипций был создан из текстов стенограмм АКУЕМ (около 50 тыс. слов).

Для проверки правильности транскрипций (и для контрольной выборки при распознавании) использовались записи продолжительностью 13,5 тыс. секунд, в которых встретилось 29 500 слов. Всего использовались записи 34 дикторов со средним и быстрым темпом речи.

Возможность генерировать сразу несколько вариантов транскрипции одного и того же слова позволяет продемонстрировать в словаре вариативность произношения наиболее частотных украинских слов, редуцирование и растяжение слов во время быстрого темпа речи, нечёткое произношение и подобные явления наряду с литературным вариантом произношения. Также система транскрибирования позволяет генерировать транскрипции для та-

ких специфических подсловарей, как словарь суржика, социальных и территориальных диалектов, аббревиатур и др.

Введение нескольких способов произнесения слов в словаре в целом улучшает надёжность распознавания спонтанной речи [11].

В дальнейшем планируется сбалансировать набор правил, которые порождают варианты произношения слов в потоке спонтанной речи, а также использовать индивидуализированные словари (для 18 групп дикторов) при создании единого словаря для распознавания спонтанной украинской речи.

Выводы

Предложенная модель позволяет довольно компактно описать закономерности преобразований между графемами и фонемами в удобном для эксперта виде с учётом неоднозначности преобразований. Разработанная реализация модели также пригодна для расшифровки чисел, символов и сокращений.

Введение в модель произвольного количества уровней транскрибирования помогает разложить правила, заданные экспертом, на более простые. Это открывает путь к применению предложенного метода для транскрибирования языков, в которых традиции написания не соответствуют произношению слов. В спонтанной украинской речи наблюдается подобное явление, но в обратном направлении: уход реальной живой речи от написания (и, соответственно, от правил канонического литературного произношения). В системе распознавания спонтанной речи нужно найти баланс между детальностью фонемной транскрипции, топологией акустических моделей фонем и количеством параметров, которые уточняют эти модели.

Результатом многозначного решения являются последовательности символов, но при этом не указывается, какие из них более вероятны. Оценка соответствующих вероятностей возможна на основе результатов распознавания, что является предметом дальнейших исследований.

Литература

- 1.** Винцюк Т.К. Анализ, распознавание и смысловая интерпретация речевых сигналов. Киев: Наукова думка, 1987.
- 2.** Young S.J. et al. HTK Book, version 3.1, Cambridge University, 2002.
- 3.** Lee, T. Kawahara and K. Shikano: Julius — an open source real-time large vocabulary recognition engine. In Proc. European Conference on Speech Communication and Technology (EUROSPEECH), 2001. P. 1691–1694.
- 4.** Vintsiuk T., Sazhok M. Multi-Level Multi-Decision Models for ASR // Proceedings of the 10th Int. Conference on Speech and Computer — SpeCom'2005, Patras, 2005. P. 69–76.
- 5.** Bisani M., Ney H. Joint-sequence models for grapheme-to-phoneme conversion // Journal Speech Communication, 50: 434–451, Elsevier, 2008.
- 6.** Вінцюк Т., Людовик Т., Сажок М., Селюх Р. Автоматичний озвучувач українських текстів на основі фонемно-трифонної моделі з використанням природного мовного сигналу // Праці 6-ї Всеукраїнської міжнародної конференції «Оброблення сигналів і зображень та розпізнавання образів» — УкрОбраз'2002, Київ, 2002.
- 7.** Сучасна українська літературна мова. Фонетика: Навч. посібник для студентів-філологів. К.: Видавничо-поліграфічний центр „Київський університет”, 2002. С. 60.
- 8.** Пилипенко В.В., Робейко В.В. Автоматизированный стенограф украинской речи // Искусственный интеллект. № 4. 2008. С. 768–775.
- 9.** Робейко В.В. Генерування індивідуалізованих транскрипцій для злитого мовлення // Мовні і концептуальні картини світу. Вип. 26. Ч. 3. Київ, 2009. С. 38–42.



- 10.** Робейко В.В. Графічна омонімія як специфічна проблема синтезу мовлення за текстом. // Українське мовознавство. Вип. 39/1. Київ, 2009. С. 429–433.
- 11.** Людовик Т.В., Пилипенко В.В., Робейко В.В. Автоматическое распознавание спонтанной украинской речи (на материале корпуса украинской эфирной речи). // Компьютерная лингвистика и интеллектуальные технологии: По материалам Международной конференции «Диалог». Вып. 10 (17). М.: РГГУ, 2011. С. 478–488.
- 12.** Васильєва Н.Б., В.В. Пилипенко, О.М. Радуцький, В.В. Робейко, М.М. Сажок. Створення акустичного корпусу українського ефірного мовлення. // Обробка сигналів і зображень та розпізнавання образів: Десята Всеукраїнська міжнародна конференція. Київ, 2010. С. 55–58.

Сведения об авторах

Робейко Валентина Васильевна —

научный сотрудник отдела распознавания и синтеза звуковых образов Международного научно-учебного центра информационных технологий и систем, Киев, Украина. valya.robeiko@gmail.com

Сажок Николай Николаевич —

кандидат технических наук, старший научный сотрудник отдела распознавания и синтеза звуковых образов Международного научно-учебного центра информационных технологий и систем, Киев, Украина. sazhok@gmail.com

Система распределённого автоматизированного документирования речевых сигналов

Крак Ю.В., доктор физико-математических наук, профессор

Загваздин А.С., научный сотрудник

В статье рассматривается система автоматизированного компьютерного стенографирования, которая включает в себя механизмы предварительной обработки сигнала и сегментацию сигнала, основанную на определении голосовой активности в сигнале. Предварительная обработка сигнала включает в себя уменьшение уровня шума и возможность изменения скорости воспроизведения сигнала без изменения его акустических характеристик. Система также предоставляет возможность сегментации голосового сигнала на основании позиций изменения диктора. Система позволяет распространять сегменты голосового сигнала между операторами-стенографистами, что позволяет повысить продуктивность стенографирования в сравнении со стандартными способами стенографирования.

- *стенографирование* • *распределённое документирование* • *обнаружение изменения диктора* • *определение голосовой активности* • *речевая сегментация сигнала* • *шумоподавление*.

A system for distributed automated transcription is described. The system features automated speech signal pre-processing and segmentation based on voice activity detection. Signal pre-processing includes the noise reduction and the ability to change the speech rate of the signal without affecting its acoustical characteristics. The system also allows to segment the signal based on the speaker change. The system allows to automatically distribute the signal segments among the members of the group of transcriptionists that allows to increase the transcription performance compared to the standard transcription methods.

- *distributed transcription* • *speaker change detection* • *voice activity detection* • *speech signal segmentation* • *noise reduction*.

Текстовая стенограмма заседания — необходимая составляющая в работе многих организаций. Как правило, процесс создания и расшифровки стенограмм достаточно продолжителен и попытки его ускорения путём расширения персонала, вовлечённого в процесс, представляются неэффективными. Для автоматизации процесса создания стенограмм заседаний предлагается система распределённого компьютерного документирования.

Поскольку стенографирование заседаний — задача, которая может быть достаточно просто распределена между многими исполнителями, система распределённого стенографирования должна поддерживать однопользовательский и многопользовательский режимы работы. На сегодняшний день в мире существует несколько систем распределённого документирования. В качестве примеров можно привести модуль стенографирования, входящий в состав системы поддержки принятия решений «Рада-3» [1], которая



используется для обеспечения законодательной деятельности Верховной Рады Украины и некоторых местных законодательных органов; системы стенографирования фирмы SRS (г. Киев) [2], которые широко используются для обеспечения протоколирования судебных заседаний; систему документирования «Нестор» фирмы «Центр речевых технологий» (г. Санкт-Петербург) [3] и некоторые другие.

Однако практически всем существующим в настоящее время системам свойственен ряд недостатков. Основной недостаток существующих систем стенографирования заседаний заключается в том, что они предъявляют высокие требования к аппаратному обеспечению, на котором работает серверная часть системы, и требует существенных затрат на внедрение системы в целом. Как правило, такие системы документирования устанавливаются стационарно в залах, где проходят заседания, и имеют жёсткую привязку к звукозаписывающему оборудованию зала, а также требуют установки отдельных серверов для обеспечения распределённого документирования, сервера звукозаписи и т.п. Также, как правило, в составе подобных систем документирования присутствует ещё и выделенный сервер баз данных. Очевидно, что системы, реализованные в соответствии с подобной архитектурой, требуют значительных усилий с точки зрения системного администрирования. Такие ограничения делают внедрение подобных систем стенографирования практически невозможным в небольших организациях, организациях с ограниченным ИТ бюджетом, а также для индивидуальных пользователей. Наличие серверной части в таких системах также делает практически невозможной мобильную работу с системой и требует постоянного подключения к локальной сети.

Ещё одним общим недостатком существующих систем является недостаток внимания, уделяемого разработчиками средствам повышения продуктивности работы операторов-стенографистов. Сегментация входящего звукового сигнала, как правило, осуществляется на фрагменты фиксированной длины. Фрагмент может обрываться на середине слова или фразы, возможности по предварительной цифровой обработке сигнала (шумоочистке, изменению скорости воспроизведения сигнала, и т.д.) также зачастую ограничены. В общем, основные недостатки существующих систем распределённого документирования могут быть сформулированы следующим образом:

- высокая стоимость внедрения и сопровождения, высокие требования к аппаратным ресурсам;
- недостаточная интеллектуальность при сегментации и распределении фрагментов входящего речевого сигнала;
- недостаточные возможности по предварительной цифровой обработке сигнала;
- отсутствие либо недостаточная развитость средств повышения продуктивности работы операторов-стенографистов.

Ниже предлагается к рассмотрению система распределённого компьютерного документирования речевых фонограмм, в основе программной архитектуры которой лежит концепция, не требующая реализации основных функций системы на выделенных серверах, где среди рабочих станций операторов, стенографистов одна может быть выделена как главная. На главной рабочей станции будет проводиться предварительная обработка звука, его сегментация и распределение между операторами. Таким образом, предложенная система сможет быть развёрнута на нескольких персональных компьютерах. В предложенной системе также поддерживается однопользовательский режим работы, при котором все функции системы будут реализованы на одном компьютере.

Требования к системе распределённого документирования

Рассмотрим основные требования, которые выдвигаются к системе распределённого компьютерного документирования:

- Система должна иметь возможность работать с входящим речевым сигналом из различных источников звуковой и видеоинформации. В упрощённом варианте система должна поддерживать различные форматы звуковых и видео файлов в качестве источников речевого сигнала (wav, mp3, wma, avi, mpeg и другие). Речевой сигнал также может сопровождаться видеосигналом, и при воспроизведении звуковой и видеосигнал должны быть синхронизированы.
- Система должна иметь возможность сегментации входящего звукового сигнала на равнозначные сегменты, при этом они не должны начинаться и заканчиваться на середине слов. Длина сегментов не должна быть меньшей определённого значения, но при этом сегменты должны быть достаточно короткими для удобного запоминания, чтобы в процессе стенографирования сводилась к минимуму необходимость повторного воспроизведения сегмента. Алгоритм сегментации должен надёжно работать в условиях нестационарного шума, который может присутствовать во входящем сигнале.
- При сегментации сигнала должна учитываться информация об изменении диктора, момент изменения диктора в сигнале должен совпадать с началом нового сегмента.
- В системе должны быть реализованы следующие функции предварительной цифровой обработки сигнала: уменьшение уровня шума, усиление уровня сигнала и изменение скорости воспроизведение сигнала, причём последнее не должно вызывать изменение тембра голоса говорящего.
- Интерфейс рабочего места оператора-стенографиста должен быть простым и интуитивно понятным и не требовать высокого уровня компьютерной грамотности от операторов системы. Это позволит сократить сроки обучения пользованию системой и получить более высокую продуктивность работы.
- Система должна иметь возможность работы как в однопользовательском, так и в многопользовательском режимах, при этом она не должна требовать комплексного внедрения и администрирования. Все компоненты системы должны иметь возможность быть установленными на распространённом компьютерном оборудовании. При этом скорость обработки сигнала и выполнения основных операций системы должны быть достаточно высокими для обеспечения комфортной работы с системой.
- В многопользовательском режиме фрагменты входящего речевого сигнала должны автоматически распределяться между операторами-стенографистами. Система должна иметь возможность работы с любым количеством операторов, которые могут подключаться к системе через локальную сеть или по сети Интернет.
- Исходя из особенностей восприятия человеком информации и кратковременной памяти человека [4], длина фрагментов, на которые разбивается входящий звуковой сигнал, должна составлять в среднем 5–9 слов. Аналогично количество основных элементов управления программы рабочего места оператора не должно превышать 9.
- Для повышения продуктивности работы операторов в системе должны быть реализованы функции автоматической проверки орфографии, автозамены и автоподстановки, автоматические всплывающие подсказки для наиболее часто встречающихся слов.

Ниже будут рассмотрены особенности реализации системы распределённого компьютерного документирования, удовлетворяющей перечисленным выше требованиям, которая была создана в Институте кибернетики НАН Украины им. В.М. Глушкова.

Сегментация входящего речевого сигнала

Одной из важнейших задач системы распределённого компьютерного документирования является сегментация сигнала: разделение входящего речевого сигнала на равнозначные сегменты. Для сегментации сигнала в рассматриваемой системе используется информ



мация о паузах, присутствующих в сигнале, а также о позициях в звуковом сигнале, где происходит смена диктора.

Для качественного определения пауз в сигнале предлагается алгоритм определения пауз, устойчивый к нестационарному уровню шума в сигнале. Поиск фрагментов сигнала, где присутствуют паузы, проводится путём сравнения энергии в анализируемом фрейме с пороговым значением. Для этого по сигналу проходим прямоугольным окном продолжительностью 50 мс таким образом, чтобы начало каждого последующего окна приходилось на середину предыдущего. Такая относительно большая продолжительность окна обусловлена тем, что для поиска пауз в речи нецелесообразно использовать паузы короче 50 мс, а большая продолжительность окна позволяет сократить общее количество итераций алгоритма. Таким образом, на каждом шаге алгоритма сигнал задаётся как

$$s_k[i] = \begin{cases} s[i], & 0.05 * D * k \leq i < 0.15 * D * k \\ 0, & \text{else} \end{cases} \quad (1)$$

Здесь D — частота дискретизации сигнала, а k — номер шага алгоритма.

Полагаем, что на участке сигнала, длиной 10 с должна быть по крайней мере одна пауза. Следовательно, для адаптации алгоритма к текущему соотношению «сигнал/шум» будем использовать информацию о предыдущих 10 с звучания сигнала. Также, будем считать, что уровень энергии сигнала в участках, соответствующих паузам, ниже, чем в участках, где присутствует голосовая активность. Энергию сигнала определим как дисперсию амплитуды сигнала в заданном окне:

$$E_{s_k} = \log_{10} \left(\frac{1}{N} \sum_{i=1}^N s_k[i]^2 - \left(\frac{1}{N} \sum_{i=1}^N s_k[i] \right)^2 \right) \quad (2)$$

На рис. 1 и 2 приведены графики амплитудно-временного представления зашумленного сигнала и соответствующий ему график уровня энергии в сигнале. Уровень энергии в участке, который соответствует пазе, существенно ниже.



Рис. 1. АЧП зашумленного сигнала с паузой



Рис. 2. Энергия сигнала, рассчитанная как дисперсия амплитуды

Для устранения негативного влияния случайных возмущений на измерения к полученным уровням энергии применяется метод медианного сглаживания 5-го порядка в соответствии с формулой:

$$s[i] = \text{med}\{s[i-2], s[i-1], s[i+1], s[i+2]\}. \quad (3)$$

Для принятия решения о том, соответствует ли анализируемый фрейм пазе, значение энергии в этом фрейме сравнивается с порогом. Поскольку окружающие условия и уровень шума в сигнале могут меняться со временем, существует необходимость динамического расчёта порога в процессе обработки сигнала. Предложен следующий алгоритм адаптивного вычисления порога энергии для пауз.

На участке сигнала протяжённостью 10 с, предшествующем анализируемому фрейму, находятся минимальный и максимальный уровни энергии для данного участка: E_{\min} и E_{\max} . Далее, уровень энергии в текущем фрейме сравнивается с полученными минимальным и максимальным значениями. Решение о том, что текущий фрейм принадлежит пазе, принимается, если выполняется следующее условие:

$$E < E_{\min} \vee \frac{E - E_{\min}}{E_{\max} - E_{\min}} < 0.2 \quad (4)$$

Минимальное и максимальное значения уточняются на каждом шаге алгоритма с учётом предыдущих 10с звучания. Фреймы с низким уровнем энергии, расположенные последовательно один за другим, объединяются в одну паузу. Паузы, длина которых меньше некоторой заданной длины, исключаются из рассмотрения, так как они наиболее вероятно соответствуют участкам с низкой энергией в середине слова (например, шипящим согласным).

Графическая работа алгоритма адаптивного вычисления порога проиллюстрирована на рис. 3.

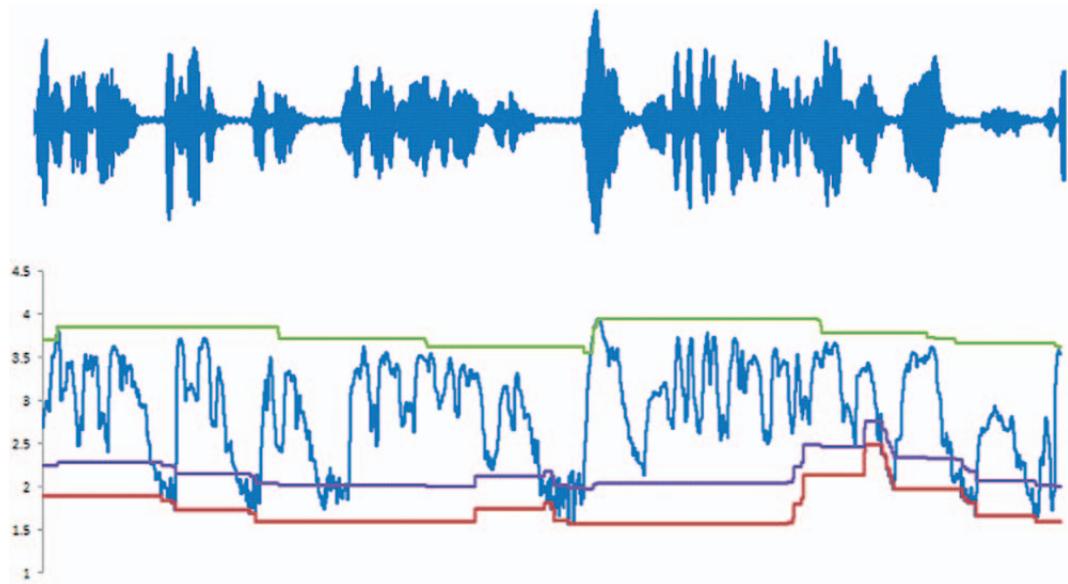


Рис. 3. Адаптивный порог при поиске пауз в сигнале

Для повышения интеллектуальности сегментации сигнала целесообразно учитывать позиции в сигнале, где происходит изменение диктора. Для нахождения таких позиций предлагается подход, рассмотренный ниже.

Полагаем, что изменение диктора в сигнале происходит в районе паузы. Иными словами, после того, как заканчивает говорить первый диктор и начинает говорить второй, есть



пауза. На практике это не всегда так, и дикторы могут перебивать друг друга, однако, такие ситуации сложно учитывать при сегментации сигнала, и в данной работе они не рассматриваются.

Положим, что $X = \{x_1, x_2, \dots, x_n\}$ — множество характеристических векторов, которые соответствуют участку сигнала до паузы, а $Y = \{y_1, y_2, \dots, y_n\}$ — множество характеристических векторов, которые соответствуют участку сигнала между текущей паузой и последующей. N_x и N_y — соответственно количество точек в первом и втором множествах. Характеристические векторы в данном случае представляют собой 13 мел-кепстральных коэффициентов, рассчитанных на участке сигнала продолжительностью 30 мс, и частота основного тона в качестве 14-го элемента вектора.

Пусть $Z = X \cup Y$ — объединение множеств характеристических векторов с количеством точек $N = N_x + N_y$. Множества X и Y сравниваются при помощи некоторой меры различия, и если они отличаются значительно, принимается положение о том, что в анализируемом участке сигнала присутствует изменение диктора.

Задачу определения наличия изменения диктора можно сформулировать в виде задачи проверки гипотезы. Пусть H_0 — гипотеза о том, что изменение диктора отсутствует, а H_1 — гипотеза о том, что происходит изменение диктора. Положим также, что векторы, из которых состоят множества X и Y являются независимыми и имеющими одинаковое распределение случайными величинами. Пусть Θ_Z — параметры распределения для множества Z , рассчитанные при помощи метода максимального правдоподобия. В таком случае логарифмическое соотношение правдоподобия для множества наблюдений Z при условии выполнения гипотезы H_0 запишется как:

$$L_0 = \sum_{i=1}^{N_z} \log p(x_i | \Theta_Z) + \sum_{i=1}^{N_y} \log p(y_i | \Theta_Z) \quad (5)$$

Здесь $p(x | \Theta)$ — вероятность того, что x выполняется при условии Θ . Функция плотности распределения находится при помощи метода Gaussian Mixture Models (GMM).

Для проверки гипотезы H_1 рассчитываются параметры индивидуальных распределений для наборов наблюдений X и Y , которые соответственно обозначаются как Θ_X и Θ_Y . При этом логарифмическое отношение правдоподобия для гипотезы записывается как:

$$L_1 = \sum_{i=1}^{N_x} \log p(x_i | \Theta_X) + \sum_{i=1}^{N_y} \log p(y_i | \Theta_Y) \quad (6)$$

Меру различия для множеств X и Y , в таком случае, можно задать как байесовский информационный критерий:

$$d_1 = L_1 - L_0 - \frac{\lambda}{2} \Delta K \log N \quad (7)$$

Здесь $\Delta Z = N_x - N_y$, а λ — параметр, который подбирается экспериментально. Решение о наличии изменения диктора в анализируемом участке сигнала принимается, если заданная таким образом мера различия превышает некоторый порог, который задаётся экспериментально.

Далее в процессе сегментации сигнала учитывается информация о найденных паузах и позициях изменения диктора. Процесс сегментации показан на рис. 4.



Рис. 4. Процесс сегментации сигнала

Предварительная цифровая обработка сигнала

В рассматриваемой системе в качестве метода предварительной цифровой обработки сигнала реализованы методы уменьшения уровня шума и изменения скорости воспроизведения звукового сигнала без изменения тембра голоса говорящего.

Шумы, которые, как правило, присутствуют в сигналах, подаваемых на вход системы распределённого документирования, могут считаться аддитивными в спектральной области. Следовательно, для фильтрации таких шумов могут быть применены распространённые методы спектрального вычитания или виннеровской фильтрации. В случае виннеровской фильтрации фильтр задаётся следующим образом [5]:

$$H(\omega) = \frac{S_x(\omega)}{S_x(\omega) + S_N(\omega)} \quad (8)$$

Здесь $S_x(\omega)$ — спектр сигнала, а $S_N(\omega)$ — спектр шума.

Функция фильтрации для фрейма m может быть задана как:

$$f(Y_m) = \frac{S_x(m)}{S_x(m) + S_N(m)} Y_m \quad (9)$$

Если спектр шума известен, то спектр очищенного от шума сигнала можно рассчитать как

$$S_x(m) = \begin{cases} |Y(m)|^2 - S_N(m), & |Y(m)|^2 > S_N(m) \\ 0 & \text{else} \end{cases} \quad (10)$$

Отсюда функция фильтрации шума может быть записана следующим образом:

$$f(Y(m)) = \begin{cases} \frac{|Y(m)|^2 - S_N(m)}{|Y(m)|^2} Y(m), & |Y(m)|^2 > S_N(m) \\ 0, & \text{else} \end{cases} \quad (11)$$

Для работы описанного метода фильтрации необходимо знать спектр шума. Предположим, что участки сигнала, которые соответствуют паузам, содержат только шум. Следовательно, для аппроксимации шума можно использовать паузы, которые были найдены при помощи алгоритма, описанного в «Сегментации речевого сигнала».

Для изменения скорости воспроизведения сигнала с сохранением тембра голоса диктора необходимо убедиться, что продолжительность сигнала изменяется, но частота основного тона говорящего сохраняется. Обеспечить это возможно при помощи использования ал-



горитмов типа PSOLA (pitch-synchronous overlap and add), которые широко применяются в системах искусственного синтеза речи. Для реализации таких алгоритмов сперва решается задача обнаружения периодов псевдопериодичности в звуковом сигнале (питч-периодов). Для этого исходный звуковой сигнал пропускается через низкочастотный и высокочастотный фильтры с конечными импульсными характеристиками. На рис. 5 и 6 приведен пример слога «ма» до и после фильтрации соответственно.

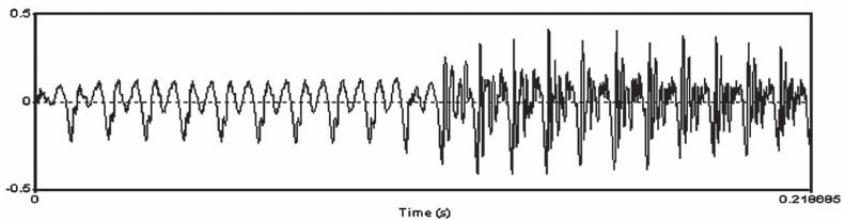


Рис. 5. Слог «ма» до фильтрации

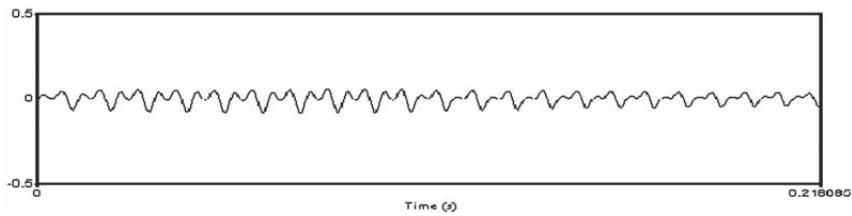


Рис. 6. Слог «ма» после фильтрации

Далее для сглаживания сигнала каждый элемент вектора исходного сигнала заменяется на взвешенное среднее четырёх окружающих его элементов по формуле:

$$d[i] = \frac{3x[i-2] + x[i-1] - x[i+1] - 3x[i+2]}{10}. \quad (12)$$

К полученному сигналу применяется медианное сглаживание порядка $y = 199$ (каждый элемент вектора заменяется на медиану вектора, состоящего из n элементов, окружающих текущий элемент). Вид сигнала после сглаживания представлен на рис. 7.

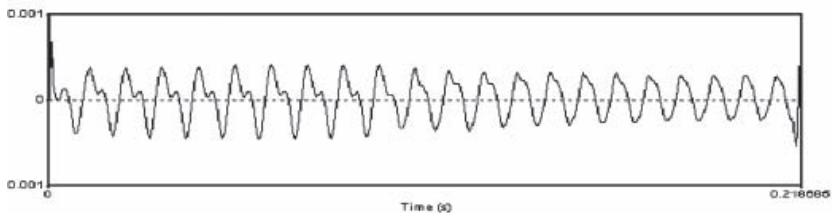


Рис. 7. Слог «ма» после фильтрации и медианного сглаживания

После этого в полученном сигнале обнаруживаются точки, где последовательность, состоящая из элементов вектора сигнала, изменяет знак с «-» на «+» и такие точки обозначаются как границы питч-периодов. Среди определённых таким образом границ обнаружаются и исключаются точки, расположенные слишком близко друг к другу, а для участков сигнала, где нет явной псевдопериодичности назначаются условные границы с некоторым постоянным интервалом.

После определения границ периодов псевдопериодичности можно изменять акустические характеристики сигнала. Исходный сигнал можно представить в виде функции периодов основного тона $x_i[n]$:

$$x[n] = \sum_{i=-\infty}^{\infty} x_i[n - t_a[i]], \quad (13)$$

где $t_a[i]$ — границы периодов псевдопериодичности сигнала, т.е. разница между двумя соседними границами $P_a[i] = t_a[i] - t_a[i-1]$ равняется периоду основного тона в момент времени $t_a[i]$. Питч-период определим через исходный сигнал помноженный на оконную функцию:

$$x_i[n] = w_i[n]x[n], \quad (14)$$

где окна w_i удовлетворяют условию:

$$\sum_{i=-\infty}^{\infty} w_i[n - t_a[i]] = 1, \quad (15)$$

что достигается использованием оконных функций типа Хэннинга или трапециевидным окном длиной в два периода основного тона.

В результате работы алгоритма необходимо получить сигнал $y[n]$, который имеет одинаковые с $x[n]$ спектральные характеристики, но отличается от него основным тоном и/или продолжительностью. Чтобы достичь этого, заменяем аналитические границы питч-периодов $t_a[i]$ границами $t_b[i]$, а аналитические периоды основного тона $x_i[n]$ — периодами $y_i[n]$ согласно

$$y[n] = \sum_{j=-\infty}^{\infty} y_j[n - t_b[j]]. \quad (16)$$

Таким образом, достаточно лишь задать границы $t_b[i]$, соответствующие продолжительности и основному тону, которые необходимо получить. Результатирующий период основного тона $y_i[n]$ получаем подстановкой ближайшего соответствующего аналитического периода $x_i[n]$. Графически работа алгоритма представлена на рис. 8.

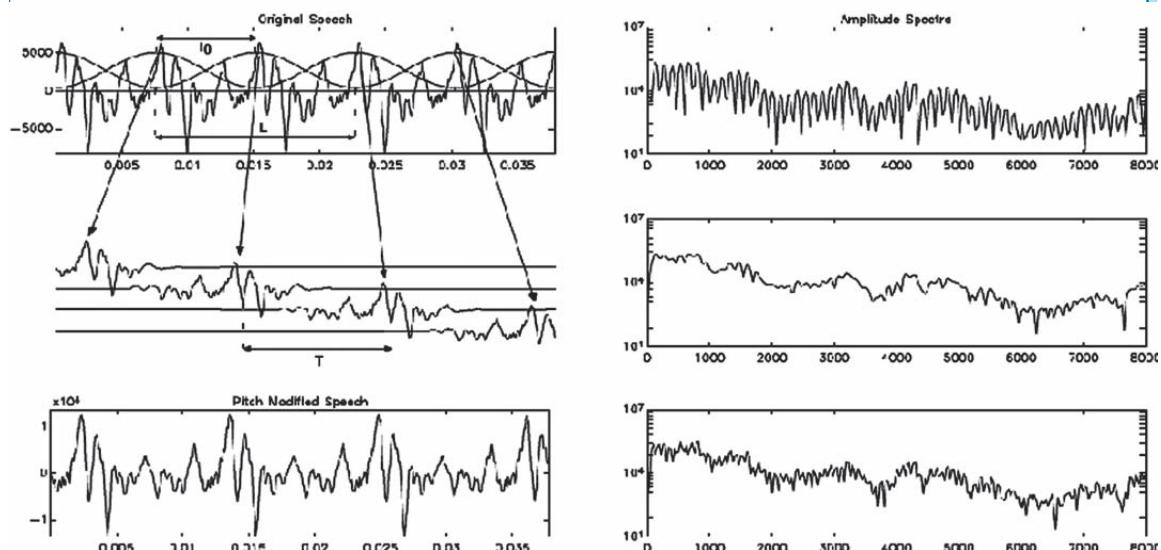


Рис. 8. Схема работы алгоритма для модификации продолжительности и основного тона



Программная реализация системы

Программная архитектура рассматриваемой системы распределённого компьютерного документирования построена без использования централизованного сервера. Такой подход обеспечивает высокую мобильность системы и небольшие требования к аппаратному обеспечению. Система состоит из двух типов модулей: главного модуля и модулей, которые являются рабочими местами операторов-стенографистов. На главном модуле осуществляется получение звукового сигнала, его предварительная обработка и сегментация. Также главный модуль управляет распределением сегментов между операторами. На нём осуществляется компоновка полученных от операторов результатов обработки сегментов в единый результирующий документ-стенограмму. В процессе обработки фонограммы оператор главного модуля может также отслеживать текущие статусы сегментов. Кроме всех перечисленных функций, в главном модуле системы доступны все функции, которые доступны прочим операторам-стенографистам: прослушивание звукового сигнала, редактирование текстового представления сегмента, изменение скорости воспроизведения и т.д.

На рабочем месте оператора-стенографиста доступны функции: получения по сети следующего доступного сегмента от главного модуля, прослушивание сегмента, редактирование текстового представления сегмента, отправка результата на сервер. Для повышения производительности работы операторам предоставляется возможность автоматической проверки орфографии, автозамены и автоподстановки, а также автоматических всплывающих подсказок для длинных слов.

Как главный модуль, так и рабочие места операторов могут быть установлены на обычных компьютерах под управлением ОС Microsoft Windows XP, Vista или Windows 7.

Распределение сегментов между операторами-стенографистами осуществляется следующим образом: при подключении к главному модулю системы оператор получает следующий необработанный сегмент из очереди. При этом на главном модуле системы такой сегмент помечается как «находящийся в обработке», и он становится заблокированным и недоступным для других пользователей системы. После завершения обработки фрагмент может принять один из статусов «завершён успешно» или «при обработке возникла проблема». Сегмент также может стать доступным автоматически для других операторов системы в случае отключения оператора, который его заблокировал.

Выводы

После реализации прототипа системы был проведён эксперимент, целью которого было сравнение эффективности работы одного стенографиста и группы стенографистов при использовании предложенной системы и без такого, т.е. при использовании лишь традиционных средств (таких как Windows Media Player для воспроизведения звукового сигнала и Microsoft Office Word для набора текста стенограммы). В качестве входящего речевого сигнала для эксперимента была выбрана запись защиты докторской диссертации продолжительностью около 2 часов. В результате эксперимента получены следующие результаты:

- При работе одного оператора-стенографиста на стенографирование записи с использованием предложенной системы было потрачено около 4 часов. Для стенографирования этой же записи при использовании стандартных средств оператор тратит в среднем 12–16 часов.

- Группе стенографистов из 5 человек для обработки записи понадобилось около 40 минут, после чего полученный текст стенограммы был направлен на обработку корректору. Вместе с коррекцией в общей сложности расшифровка стенограммы заняла около 1 часа.

Проведённый эксперимент демонстрирует эффективность предложенной системы в сравнении с традиционными средствами. Вместе с остальными преимуществами, среди которых отсутствие необходимости администрирования и внедрения, простота в использовании и качественная предварительная обработка сигнала, предложенная система является достаточно эффективным средством для автоматизации процесса создания и расшифровки стенограмм заседаний для небольших и крупных организаций, а также для индивидуальных пользователей.

Литература

1. Морозов А.О. «Рада-3» — система підтримки прийняття рішень для законотворчого процесу Верховної Ради України та рад інших рівнів. / А.О. Морозов, Л.Б. Баран, В.В. Косячиков, В.Л. Косолапов // Математичні машини і системи. 2008. № 1. С. 3–22.
2. Система стенографирования SRS Report. Сайт компании SRS. Электронный ресурс. Режим доступа: http://srs.kiev.ua/index.php?option=com_content&view=article&id=43%3Asrs-report&catid=3%3A2009-07-15-06-10-55&Itemid=13&lang=ru
3. Система распределённого компьютерного документирования устной речи Нестор. Общее описание системы. Электронный ресурс. Режим доступа: <http://www.speechpro.ru/sites/default/files/product/docs/description.pdf>
4. Miller G. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. // The Psychological Review. 1956. Vol. 63. P. 81–97.
5. Rabiner L. A comparative performance study of several pitch detection algorithms. / L.R. Rabiner, M.J. Cheng, A.E. Reseonberg, C.A. McGonegal // IEEE Tran. on. Acoustics, Speech and Signal Processing. № 24(5), 1976. P. 399–418.

Сведения об авторе

Юрий Васильевич Крак —

доктор физико-математических наук, профессор Киевского национального университета имени Тараса Шевченко, старший научный сотрудник института кибернетики им. В.М. Глушкова НАН Украины. Специалист в области искусственного интеллекта, анализа и синтеза голосовой и жестовой коммуникационной информации. Автор около 300 научных работ.

Александр Сергеевич (Семёнович) Загваздин —

закончил факультет кибернетики в 2007 году. Научный сотрудник института кибернетики им. В.М. Глушкова НАН Украины. Круг научных интересов включает цифровую обработку речевых сигналов, синтез и распознавание речи.



Система сегментно-слогового распознавания изолированных слов из больших словарей

Савенкова О.А.

В статье рассмотрены основные модули системы сегментно-слогового распознавания изолированных слов из больших словарей, математические модели и алгоритмы сегментно-слогового синтеза эталонной траектории, положенные в их основу.

- система распознавания речи • слог • сегментно-слоговое распознавание • траектория параметров • алгоритм эвристического поиска.

The basic modules of the large vocabulary segment-syllabic recognition system of isolated words, the mathematical models and the segment-syllabic synthesis algorithms of a reference parameters trajectory are considered.

- speech recognition system • syllable • segment-syllabic recognition • parameters trajectory • heuristic search.

Введение

Основные направления разработки систем распознавания речи (далее CPP): распознавание отдельно произносимых, не связанных по смыслу, слов, а также распознавание слитной речи как решение задачи декодирования языковой структуры речевого сообщения с использованием различных источников лингвистических знаний [1–10].

Решение задачи обработки и распознавания речи — задача структурной аппроксимации речевого сигнала (далее РС) как совокупности компонентов, так и иерархии категорий информации (информация о физической природе РС, о лингвистической структуре языка и речи). Существует актуальная проблема обеспечения соответствия между параметрическим, фонетическим и символным представлениями речи [1, 4, 9]. Необходим комплексный подход к построению CPP, учитывающий взаимосвязи между иерархией представления информации о РС и решающей проблему акустофонетического декодирования [1, 2, 5, 7, 9–11]. Данная проблема не имеет точного решения. Поэтому в основу её решения положим стратегию эвристического поиска с использованием специфических знаний о речевой информации на разных уровнях иерархии. Для этого согласно [12], исходную задачу сведём, во-первых, к выбору объектов, необходимых для решения задачи (алфавит объектов распознавания), а во-вторых, к разработке стратегий эффективного поиска потенциальных решений, которые могут быть сгенерированы с помощью выбранных объектов.

Однозначного ответа на вопрос о том, какие элементарные образы (фонема, слог, слово) и в каком количестве используются в процессе распознавания, пока не получено. Однако в ряде исследований показано, что обработка речи требует перехода от распознавания слов как целостных звуковых

образов к распознаванию звуковых единиц, которые меньше слова [3]. Сравнение моделей языка по количеству используемых разных языковых единиц в словаре и процентом непокрытых слов в текстовом материале показало, что чем меньше размер элементов, тем более полно они покрывают пространство слов языка [13].

Результаты исследований в области психоакустики и экспериментальной фонетики показывают, что любую речевую реализацию можно рассматривать как последовательность открытых слогов, которые представляют собой единство слогообразующего гласного (Γ) и одного или более согласных (C). Фрагмент речи между паузами является цепочкой связанных между собой слогов. Причем единого деления на слоги внутри такого фрагмента не происходит, и в этом смысле слог не отличается от отдельного звука речи [10, 11, 14–18]. Таким образом, с одной стороны, слог — основа образования разнообразных звуков, а с другой — большого числа разных ритмических структур слов [11, 18–20]. Слоги являются оптимальными с точки зрения артикуляционной организации и надежности распознавания. Их основное преимущество — учёт просодических свойств речи. О чём свидетельствуют существующие примеры систем анализа/синтеза речи, приведённые в [19, 21–24].

Элементы алфавита объектов распознавания для CPP из больших словарей должны иметь такую длину и быть подобраны в таком количестве, чтобы из них можно было построить любые слова или фразы и предложения [4, 25, 26]. Этим требованиям удовлетворяют слова-слоги длиной 2 и 3 символа-фонемы, причём в [25, 26] выявлено, что наиболее употребляемыми являются слоги, которые содержат 2 ($C + \Gamma$) и 3 ($C + C + \Gamma$) фонемы.

С учётом вышеизложенных фактов в данной работе рассмотрим модели и алгоритмы для решения задачи фонемно-слогового распознавания речи, в общем виде представленной в [4].

Пусть задан алфавит слогов $SL = \{SL_1, \dots, SL_k, \dots, SL_Z\}$, $k = 1, \dots, Z$. Каждый слог SL_k содержит K символов-фонем $SL_k = (P_1, P_2, \dots, P_K)$, $\forall P_i \in P$ (P — множество всех символов-фонем). Для каждого слога SL_k задана последовательность параметров $Y_k = (y_{k1}, y_{k2}, \dots, y_{ki}, \dots, y_{kN_k})$, $i = 1, \dots, N_k$ (траектория параметров (далее ТП) в пространстве признаков), определены границы сегментов-фонем $S_k = (s_{k1}, s_{k2}, \dots, s_{kj}, \dots, s_{kL_k})$, $j = 1, \dots, L_k$. Пусть также задана последовательность параметров $X = (x_1, x_2, \dots, x_N)$ реализации РС, предъявленного для распознавания. Для X определены границы сегментов-фонем $S = (s_1, s_2, \dots, s_L)$. Сегменты-фонемы SG_i ($SG_i = (x_{sbeg}, \dots, x_{send})$, $sbeg = s_i$, $send = s_{i+1}$, $i = 1, \dots, L$), составляющие последовательность параметров $X = (SG_1, SG_2, \dots, SG_i, \dots, SG_L)$, могут быть объединены некоторым образом в M групп-слогов XSL_p по m_p сегментов-фонем каждая, причем $\sum_p m_p = L$, $p = 1, \dots, M$. Символьная последовательность

$W^* = (P_1^*, P_2^*, \dots)$, которая соответствует последовательности параметров X предъявленной реализации РС, неизвестна. Необходимо траектории параметров X наилучшим образом сопоставить траектории параметров слогов $\{Y_k\}$, вычисляя

$$dist = \sum_p \min_k (XSL_p \# Y_k), p = 1, \dots, M, k = 1, \dots, Z, \quad (1)$$

где $\#$ — операция сопоставления. Таким образом, необходимо построить такую последовательность параметров X^* , которая будет близкой по своим параметрам к параметрам предъявленной реализации речевого сигнала X . Такую траекторию параметров X^* будем называть *эталонной траекторией параметров* (далее ЭТП). Некоторую траекторию



параметров X^* , построенную из X элементов алфавита SL , которая по количеству сегментов соответствует предъявленной траектории параметров X , будем называть *решением-кандидатом*.

Для построения решений задачи (1) рассмотрим подход, основанный на применении методов поиска в пространстве состояний с учётом особенностей исследуемой задачи [12]: выбран алгоритм эвристического поиска в пространстве состояний, который в отличие от базовых стратегий поиска, использует определенного вида оценочную функцию (далее ОФ) $f(n)$, сокращающую объем перебора. В общем случае ОФ $f(n)$ для узла n имеет вид

$$f(n) = g(n) + h(n), \quad (2)$$

где $g(n)$ — длина пути от начального узла к узлу n , $h(n)$ — эвристическая оценка расстояния из узла n к целевому узлу, для определения которой используют любую эвристическую информацию о решаемой задаче [12].

Структура системы распознавания речи «SPeach»

Структура системы «SPeach» (рис. 1) состоит из модулей, которые могут работать в режимах обучения и распознавания: (а) модуль обработки РС; (б) модуль формирования алфавита слогов; (в) модуль распознавания. В режиме обучения работают (а), (б), а в режиме распознавания — (а), (в).

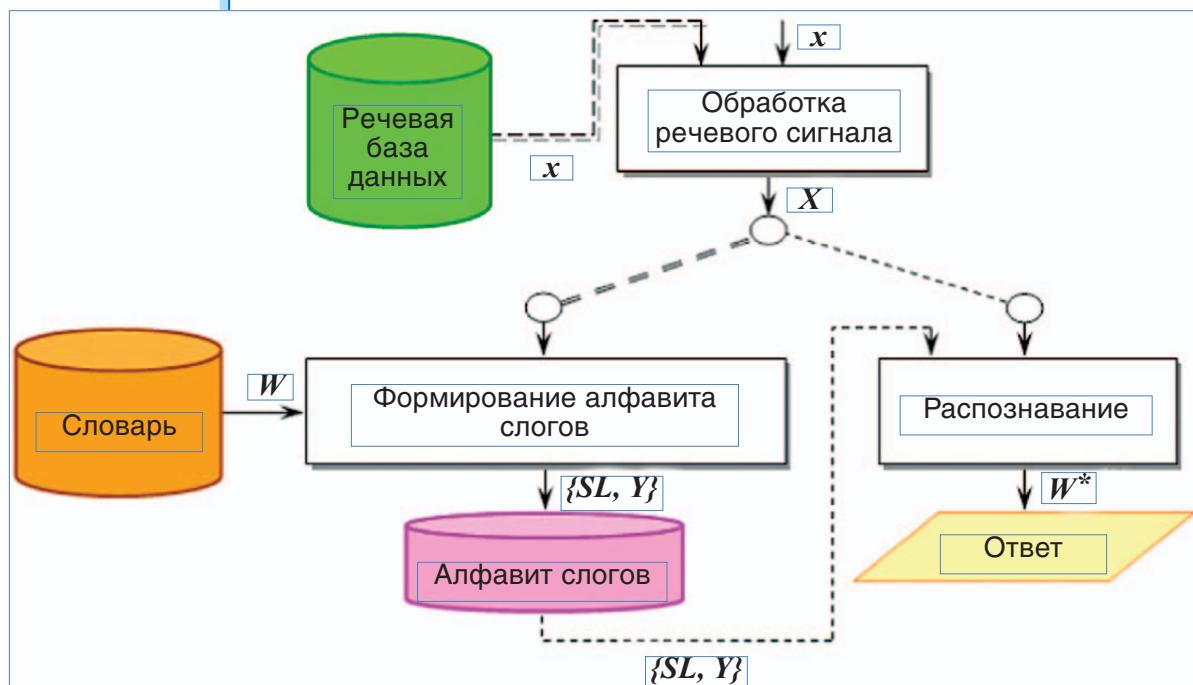


Рис. 1. Структура системы распознавания речи

Модель обучения

Последовательность этапов обработки информации о РС, которые составляют модель обучения, представлена на рис. 2.

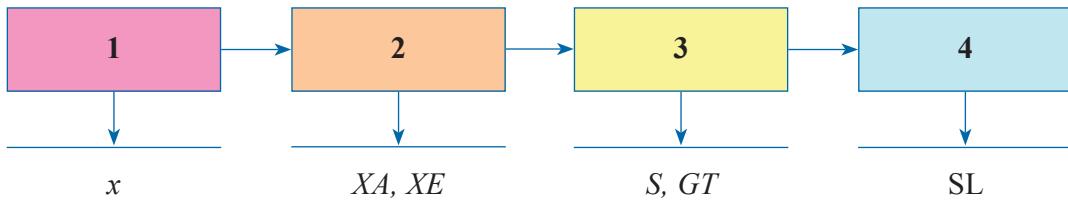


Рис. 2. Модель обучения

Шаг 1. Ввод речевого сигнала с микрофона или открытие wav-файла из речевой базы данных (далее РБД).

Частота дискретизации речевых сигналов составляет $F_S = 22050$ Гц.

Шаг 2. Первичная обработка РС, формирование траектории параметров.

Интервал анализа РС, выбранный для исследований, $\Delta T = 11,6$ мс. В данной работе траектории параметров $X = (x_1, x_2, \dots, x_i, \dots, x_T)$, $x_i = (x_{i,1}, x_{i,2}, \dots, x_{i,j}, \dots, x_{i,m})$, $i = 1, \dots, T$ (T — количество интервалов анализа РС), $j = 1, \dots, m$: спектрально-временное представление (далее СВП) $XA(\omega, t)$; спектрально-полосное представление (далее СПП) $XE(l, t)$, $l = 1, \dots, 9$ [4, 27].

Для СВП XA и СПП XE разработаны модели аналитического описания в классе функций $f \in C(\Omega)$, которые являются гладкими непрерывными функциями в соответствующих частотно-временных областях определения $\Omega: [\omega_0, \omega_1] \times [t_0, t_N]$. Согласно им можно восстановить исходную траекторию параметров РС с минимальной погрешностью.

1. Построение аналитического описания СВП слогов

СВП XA РС содержит произвольное количество всплесков/пиков спектральной энергии, которые произвольно расположены на определенной частотно-временной области D_A : $[\omega_0, \omega_M] \times [t_0, t_N]$. Естественно найти набор функций в виде пиков, с помощью которых можно описать частотные и временные свойства РС [28].

Для построения аналитического описания ТП слогов, используется колоколообразная функция (далее КФ) модифицированный локон Аньези [4, 29 — 32].

Пусть в частотно-временной области D_A таблично задана спектрально-временная функция $YA(\omega_k, t_l)$ для некоторого слога SL_j , где ω_k, t_l — дискретно заданные частота и время соответственно, $k = 0, \dots, M$, $l = 0, \dots, N$, $j = 1, \dots, Z$. Для исходной траектории параметров YA_j слога SL_j построим описание YLA_j в виде суперпозиции произведений КФ:

$$Zt_{(i)}(t_l) = \frac{a_{(i)}^3}{c_{(i)}^2 + (t_l - T_{(i)})^2}; \quad Z\omega_{(i)}(\omega_k) = \frac{b_{(i)}^3}{d_{(i)}^2 + (\omega_k - \Omega_{(i)})^2}, \quad (3)$$

где функции $Zt_{(i)}(t_l)$, $Zt \in [t_0, t_{N_j}]$, $i = 1, \dots, L$ описывают временные свойства РС; функции $Z\omega_{(i)}(\omega_k)$, $Z\omega \in [\omega_0, \omega_M]$, $i = 1, \dots, L$ — частотные свойства РС.

Произведение $Zt_{(i)}(T) \cdot Z\omega_{(i)}(\Omega)$ описывает всплеск $i = 1, \dots, L$ спектрально-временной



функции $Y\mathcal{A}$, который находится на частоте Ω , в момент времени T . Областью определения произведений функций $Zt_{(i)}(T) \cdot Z\omega_{(i)}(\Omega)$ является область $D_A : [\omega_0, \omega_M] \times [t_0, t_N]$. Тогда описание $YLA(\omega_k, t_l)$ в виде суперпозиции L произведений КФ (3) в некоторой точке (ω_k, t_l) СВП имеет вид

$$YLA(\omega_k, t_l) = \sum_{i=1}^L Z\omega_{(i)}(\omega_k) \cdot Zt_{(i)}(t_l), \quad k = 0, \dots, M, \quad l = 0, \dots, N. \quad (4)$$

Определение неизвестных параметров КФ (3) $a_{(i)}, b_{(i)}, c_{(i)}, d_{(i)}, T_{(i)}, \Omega_{(i)}$, $i = 1, \dots, L$ выполняется по алгоритму, предложенному в [29, 30].

Пример аналитического описания СВП для некоторых слогов приведён на рис. 3–6.

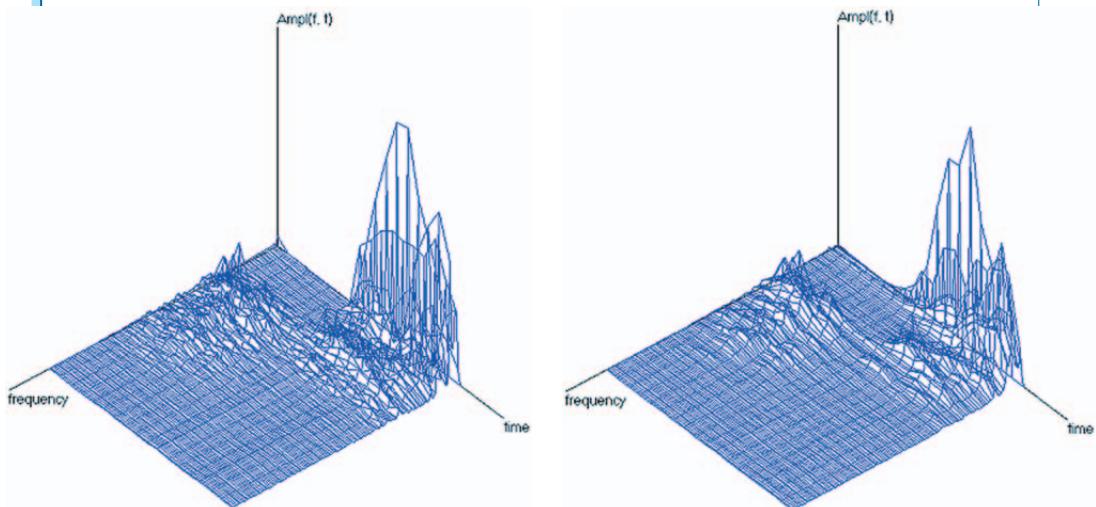


Рис. 3. Исходное СВП слога «че»

Рис. 4. Описание СВП слога «че»

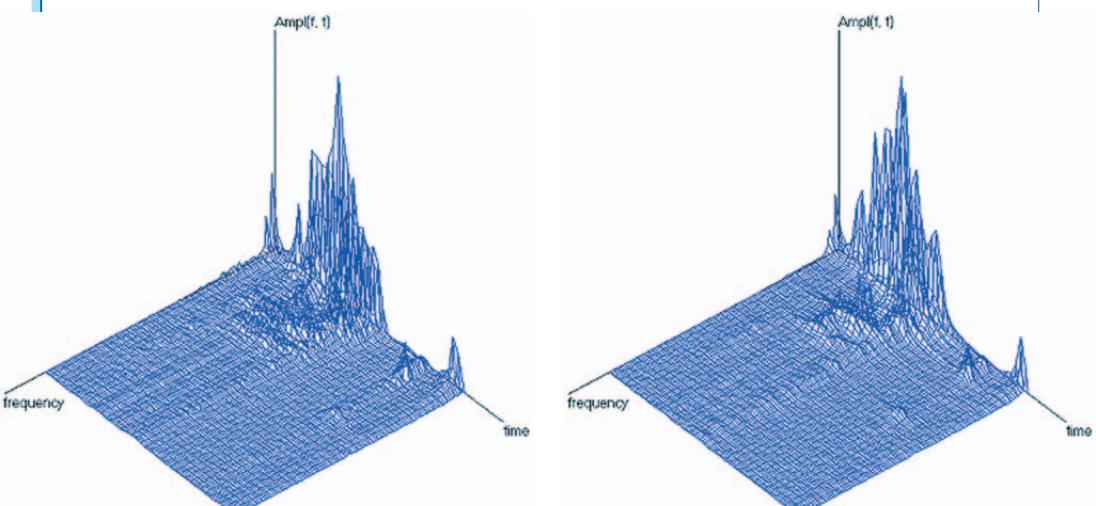


Рис. 5. Исходное СВП слога «век»

Рис. 6. Описание СВП слога «век»

2. Построение сплайн-описания СПП слогов

Анализируя вид СПП на интервалах продолжительностью сегмента-фонемы для разных звуков речи, можно сделать вывод: на каждом временном отрезке, который соответствует сегменту-фонеме, последовательности параметров СПП в каждой частотной полосе $l=1,\dots,9$ имеют вид, который можно описать полиномами низких порядков $P_n(x)$, $n \leq 3$. Таким образом, для аналитического описания СПП элементов алфавита можно применить математический аппарат сплайн-функций [4, 33].

Пусть для слога SL_k сформирована траектория параметров YE_k (СПП), найдены граничные сегментов-фонем SG_{kn} $S = (s_1, \dots, s_n, \dots, s_{L_k})$, $k=1,\dots,Z$, $n=1,\dots,L_k$, $L_k = 2; 3; 4$. Тогда модель сплайн-описания траектории параметров слога SL_k , которая аппроксимирует исходную YE_k в частотной полосе, имеет вид

$$YSE_k = \begin{cases} YSE_{k,1,i}, & s_1 \leq i \leq s_2, \\ \dots \\ YSE_{k,n,i}, & s_{n-1} \leq i \leq s_n, \\ \dots \\ YSE_{k,L_k,i}, & s_{L_k-1} \leq i \leq s_{L_k}, \end{cases} \quad (5)$$

$$YSE_{k,n,i} = a_{k,n} + b_{k,n} \cdot (t_{k,i} - s_{n-1}) + c_{k,n} \cdot (t_{k,i} - s_{n-1})^2 + d_{k,n} \cdot (t_{k,i} - s_{n-1})^3, \quad (6)$$

где $a_{k,n}$, $b_{k,n}$, $c_{k,n}$, $d_{k,n}$ — коэффициенты $P_3(x)$, который описывает n -й сегмент-фонему k -го слога $n=1,\dots,L_k$.

Для нахождения параметров модели сплайн-описания траектории параметров YE_k в каждой частотной полосе l решается задача минимизации среднеквадратического приближения с условиями в точках сегментации (на основе эмпирических данных в качестве узлов сплайн-функции выбраны точки сегментации РС), которые обеспечивают требуемую гладкость и непрерывность:

$$\sigma_l^2 = \sum_{j=1}^{L_k} \sum_{i=s_{j-1}}^{s_j} (YSE_{l,k,i} - YE_{l,k,i})^2 \rightarrow \min \quad [32].$$

Шаг 3. Сегментация речевого сигнала методом верификации [4].

Шаг 4. Формирование алфавита слогов.

1. Выбор множества слов $W = \{W_j\}$, $j = 1, \dots, NW$ и соответствующей РБД

В качестве словаря, который используется для формирования алфавита слогов, выберем список частотных слов, построенный на основе представительного корпуса современного языка. Статистические данные о зависимости процента покрытия текстового корпуса (около 16 млн слов) частотными словами ($Perc$) от количества частотных слов Nw представлены в [34].

2. Выбор структуры данных для хранения и эффективного использования информации о слогах в системе распознавания речи

Для элемента $SL_{M,k}$ $k=1,\dots,Z$, $M=2; 3$ из множества M -символьных слогов предложена структура данных, которая является совокупностью следующих категорий информации:

- лингвистическая информация о k -том слоге (символьное представление $SL_{M,k}$ транскрипция $SLT_{M,k}$, идентификатор слова, в состав которого входит и др.);



- б) вспомогательная информация об SL_k (идентификатор слога, количество временных отсчетов $NT_{M,k}$, количество сегментов L_k , границы сегментов $S_k = (s_1, \dots, s_{L_k})$, ТШП-транскрипция GT_k);
- в) параметрическое представление (СВП, СПП, параметры модели описания СВП в классе колоколообразных функций, параметры модели сплайн-описания СПП).

3. Декомпозиция слов W_j на слоги SL

В модуле формирования алфавита слогов реализованы алгоритмы для следующей задачи [35].

Пусть задано некоторое множество слов $W = \{W_j\}, j = 1, \dots, NW$, которому соответствует множество реализаций РС. Каждое слово W_j содержит LW_j символов, т.е. $W_j = (w_{j,1}, w_{j,2}, \dots, w_{j,LW_j})$. Для каждого слова W_j найдена транскрипция $WT_j = (wt_{j,1}, wt_{j,2}, \dots, wt_{j,LWT_j})$ длиной LWT_j символов, которая соответствует сегментам-фонемам траектории параметров РС.

Процесс деления слов $W = \{W_j\}, j = 1, \dots, NW$, с учетом транскрипции WT_j в цепочку M -символьных слогов будем называть декомпозицией

$$Decomp^M : W_j \xrightarrow{WT_j} Decomp^M(W_j) = SL_1^M, \dots, SL_{L_{M,j}}^M, \quad (7)$$

где $L_{M,j}$ — количество M -символьных слогов в слове W_j ($M = 2, 3$).

Таким образом, в результате декомпозиции (7) для множества слов W сформирован алфавит SL , состоящий из M -символьных слогов $SL_{M,k}$ ($SL_{2,k} = \{w_{j,m}, w_{j,m+1}\}$, $SL_{3,k} = \{w_{j,m}, w_{j,m+1}, w_{j,m+2}\}$, $m = 1, \dots, LW_j - M$), общее количество которых составляет:

$$F_{SL}(W) = \sum_{j=1}^{NW} \sum_{M=2}^3 L_{M,j} \quad (M = 2, 3). \quad \text{Очевидно, что такой набор объектов распознавания является избыточным, поэтому для покрытия множества слов } W \text{ формируется такой алфавит } SL^*, \text{ что}$$

$$F_{SL^*}(W) \rightarrow \min. \quad (8)$$

Проанализируем списки из 1000 ($W1000$), 5000 ($W5000$) и 9000 ($W9000$) наиболее частотных слов русского языка, для которых значения $Perc$: 64,07%, 82,06% и 87,82% соответственно [34].

Для выбранных списков частотных слов выполнено сравнение количества 2- и 3-символьных сочетаний, покрывающих указанные множества слов (результаты приведены на рис. 7). Выяснено, что множество таких 2- и 3-символьных сочетаний, полученное с помощью (7) для $W5000$, покрывает почти 90% сочетаний, полученных для $W9000$.

Множество слов $W5000$ покрывает 82% текстового корпуса, который составляет 13 млн слов, что достаточно для создания СПП с большим словарём. Кроме того, начиная с $Nw = 5000$ частотных слов, значение прироста процента покрытия текстового корпуса этими словами увеличивается с достаточно малым шагом.

Поэтому считаем, что наиболее оптимальным для создания алфавита слогов является использование множества из 5000 наиболее частотных слов языка.

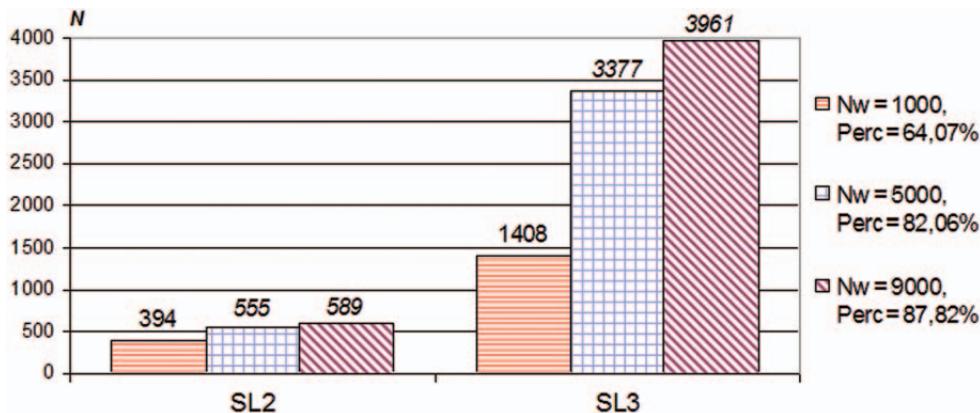


Рис. 7. Количественный анализ 1000, 5000, 9000 наиболее частотных слов

При формировании алфавита слов необходимо также учитывать информацию о фонетическом представлении слов (обобщённая звуковая транскрипция, основанная на действующих в данном языке стандартных правилах чтения), которую можно получить с помощью правил транскрибирования (например, для украинского языка [22], для русского языка [19]) [1]. Так, например, для множества слов $W5000$ определено, что 92,8% 2- и 3-символьных сочетаний соответствуют 2-, 3- и 4-символьным сочетаниям в транскрипции, а 7,2% — соответствуют более длинным сочетаниям, которые можно получить объединением 2-, 3-, 4-символьных.

Известно, что ГС-сочетания сегментов С + Г, Г + С и С + С (Γ — гласный, С — согласный) составляют три типа интеграции артикуляторных работ и составляют структуру артикуляторного жеста в таких произносимых единицах, как слог и фонетическое слово [26]. Для множества слов $W5000$ был выполнен анализ количественного состава 2- и 3-символьных ГС-сочетаний разных типов и определен порог $N_g \geq 5\%$ для включения их в алфавит (рис. 8).

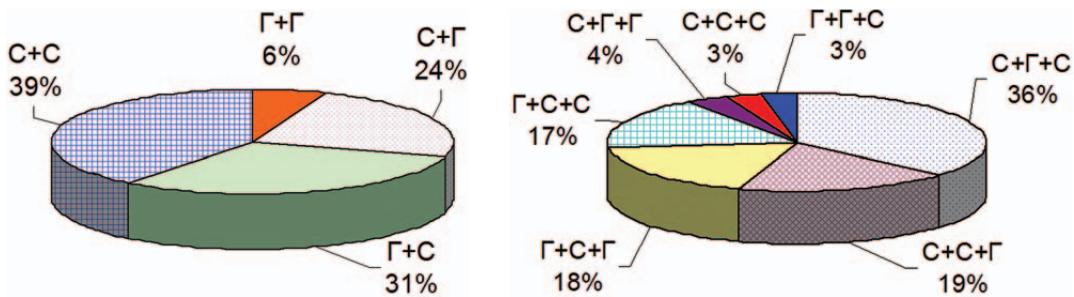


Рис. 8. Структура множеств 2- и 3-символьных сочетаний для 5000 наиболее частотных слов

Сформулируем правило выбора элементов в алфавит на основе декомпозиции наиболее частотных слов языка: выбирать 2- и 3-фонемные слоги, составляющие все типы интеграции артикуляторных работ, и которым отвечают 2, 3 или 4 символа в транскрипции и соответствующему количеству сегментов в последовательности параметров РС.

4. Группировка слов

Организация алфавита образов, которые используются в алгоритме распознавания, должна обеспечить максимально высокую скорость доступа к каждому элементу в процессе поиска. Для этого предложено применение следующих уровней группирования элементов алфавита (рис. 9).

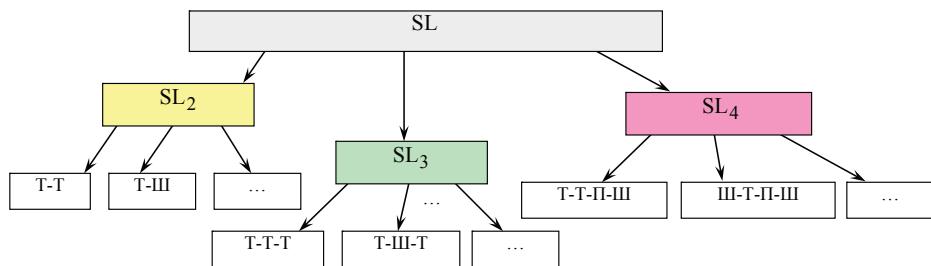


Рис. 9. Схема группирования слов

Группирование по количеству символов $\alpha = M$ и количеству сегментов $\beta = L$:

$$SL_{2,2}, SL_{3,2}, SL_{2,3}, SL_{3,3}, SL_{2,4}, SL_{3,4} \subset SL,$$

$$SL_2 \supset (SL_{2,2} \cup SL_{3,2}), SL_3 \supset (SL_{2,3} \cup SL_{3,3}),$$

$SL_4 \supset (SL_{2,4} \cup SL_{3,4})$, где $SL_{\alpha,\beta}$, $\alpha = 2, 3, \beta = 2, 3, 4$. Коэффициент сокращения рассматриваемых элементов алфавита в каждом $\alpha - \beta$ -подмножестве: $K_{\alpha,\beta} = Z/NSL_{\alpha,\beta}$, где $NSL_{\alpha,\beta}$ — количество слов в $\alpha - \beta$ -подмножестве, Z — общее количество слов.

Пусть для каждого сегмента траектории параметров РС по некоторым правилам можно определить тип сегмента (признаки (Ш) шумного, (Т) тонального или (Π) паузы). Тогда для слова можно сформировать ТШП-транскрипцию в виде символьной последовательности из обозначений типов сегментов и выполнить группирование элементов в $\alpha - \beta$ -подмножествах по типу ТШП-транскрипции ($\alpha - \beta - \gamma$ -подмножества).

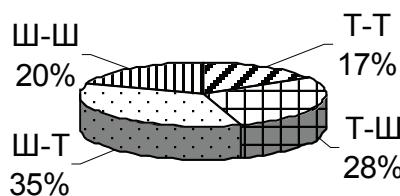
Информация о количественных данных алфавита слов приведена в таблице 1 и рис. 10 (а, б, в).

Таблица 1

Количественные данные об алфавите слов

Группа $SL_{\alpha,\beta}$	$\beta = 2$	$\beta = 3$	$\beta = 4$
$\alpha = 2$	438	121	10
$\alpha = 3$	0	1499	1579
Количество $NSL_{\alpha,\beta}$	438	1620	1589
Количество Z		3647	
Коэффициент $K_{\alpha,\beta}$	8,33	2,25	2,3

Схема группирования слов на первом уровне позволяет сократить множество рассматриваемых элементов алфавита в наилучшем случае в 8,33 раза (12,01% от общего количества слов в алфавите), а в наиболее худшем — в 2,25 раза (44,42% от общего количества слов в алфавите).



а) 2-сегментные слоги

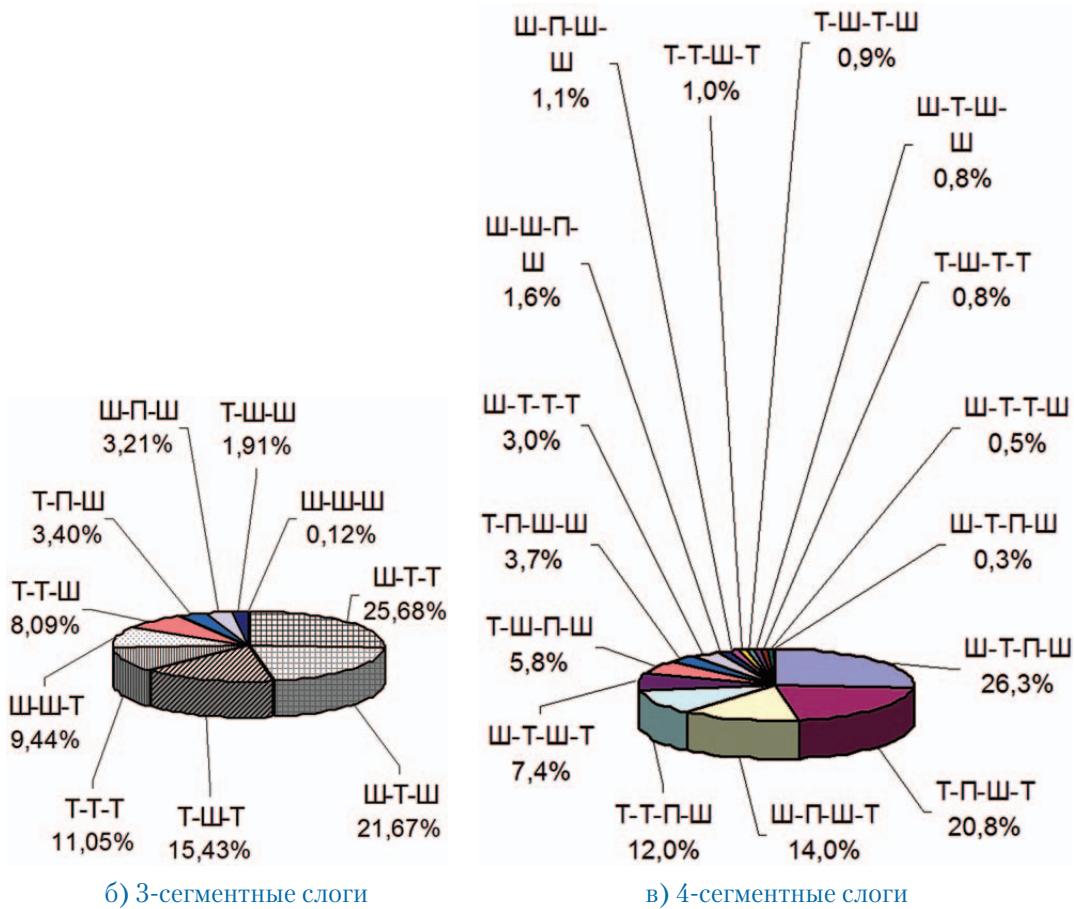


Рис. 10. Анализ алфавита слогов

5. Вычисление функции эвристической оценки состояний

Для выбора вида функции эвристической оценки состояний $h(n)$ в оценочной функции $f(n)$ (2) рассмотрены следующие характеристики сегментно-слогового представления информации о РС: вложенность слогов $SL_{2,k} \in SL_{3,k}$ на уровне слогов и траекторий параметров; тип сегментов; значение среднего расстояния между ТП слогов, которые принадлежат к разным способам сочетания признаков сегментов. Перечисленные характеристики можно использовать независимо для вычисления $h(n)$. С учётом этого, выбраны следующие слагаемые оценки $h(n)$ в $f(n)$ (2):

- 1) для узла n вычисляется значение оценки $h_v(n)$, на основе анализа вложенности слогов $SL_{2,k} \in SL_{3,k}$;
- 2) оценка $h_g(n)$ вычисляется на основе анализа вариантов сочетания признаков сегментов (Т-Т, Ш-Т и т.п.);
- 3) оценка расстояния $h_d(n)$ для всех возможных переходов из узла n для разных вариантов сочетания признаков сегментов;
- 4) оценка количества нераскрытых узлов $h_o(n)$, которые остаются нерассмотренными в процессе поиска в направлении целевого узла.

Для каждого слагаемого 1)-4) экспертом задан соответствующий коэффициент (v, g, d, o), который имеет смысл веса соответствующей эвристической оценки, и правило нормирования весовых коэффициентов: $1/v + 1/g + 1/d + 1/o = 1$.

Таким образом, выражение для вычисления значения эвристической оценки $h(n)$ из узла n к целевому узлу имеет вид

$$\begin{aligned} h(n) &= \hat{h}_v(n) + \hat{h}_g(n) + \hat{h}_d(n) + \hat{h}_o(n) = \\ &= \frac{1}{\nu} h_v(n) + \frac{1}{g} h_g(n) + \frac{1}{d} h_d(n) + \frac{1}{o} h_o(n). \quad (9) \end{aligned}$$

Для вычисления функции эвристической оценки состояний (9) был проанализирован словарь из 5000 наиболее частотных слов русского языка и соответствующие им записи из РБД [34]. Алгоритмы вычисления слагаемых функции эвристической оценки состояний и полученные результаты изложены в [36].

Модель распознавания

Последовательность этапов обработки информации о РС, которые составляют модель распознавания, представлена на рис. 11.

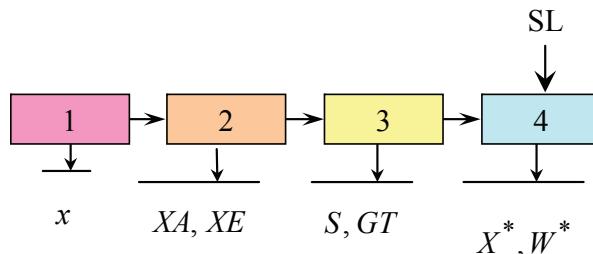


Рис. 11. Модель распознавания

Шаги 1–3 — аналогичные шагам 1–3 модели обучения, представленной на рис. 2.

В результате выполнения шагов 1–3 для предъявленного РС, сформирована траектория параметров в виде СВП XA , СПП XE , определены границы $S = \{s_1, \dots, s_L\}$ и тип GT сегментов.

Шаг 4. Поиск эвристических решений задачи сегментно-слогового распознавания.

На основе экспериментальных исследований сделан следующий вывод относительно использования представления ТП (СВП, СПП) при распознавании: СПП для предварительного распознавания и выделения списка из N -list решений-кандидатов для ЭТП; СВП — для выбора одного или нескольких решений-кандидатов из списка N -list по критерию максимальной близости в смысле (1) к ТП предъявленного РС и принятие решения о распознавании (или установление факта отказа от распознавания).

1. Поиск эвристических решений первого уровня

Для нахождения решений-кандидатов задачи сегментно-слогового распознавания (1) с помощью методов поиска в пространстве состояний (ПС) выполним формализацию задачи в терминах ПС:

1. ПС представим в виде корневого графа синтеза ЭТП X^* (рис. 12): количество узлов пространства состояний определяется в зависимости от количества сегментов L ТП XS предъявленного РС; нумерация узлов на графике отвечает номерам сегментов SG_i , $i = 1, \dots, L$ в ТП предъявленного РС; коэффици-

ент разветвления $Br = 3$; максимальная длина пути в ПС от начального узла к целевому $Len = \left\lceil \frac{L}{2} \right\rceil$; глубина самого поверхностного целевого узла $Dep = \left\lceil \frac{L}{4} \right\rceil$.

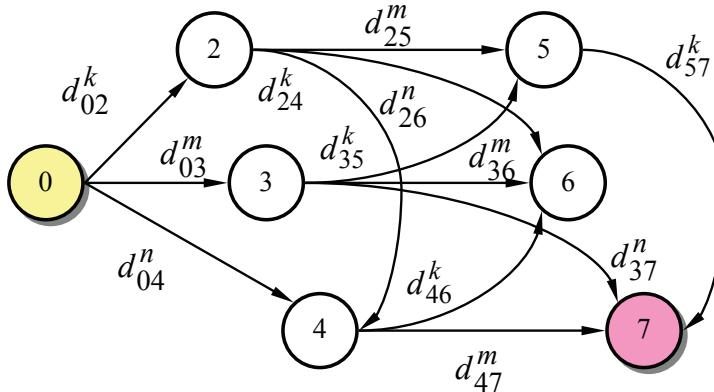


Рис. 12. Граф синтеза эталонной траектории параметров для $L = 7$

2. Состояние B_i : ТП $X_M^* = (Y_1, Y_2, \dots, Y_i, \dots, Y_M)$ на основе конкатенации ТП Y_i слов SL_i , $i = 1, \dots, M$; ТП Y_i слов SL_i содержат $L_i = 2, 3, 4$ сегмента-фонемы.
3. Начальное состояние S_0 : ЭТП X^* не содержит ТП слов, т.е. $X^* = \emptyset$.
4. Допустимые переходы на графе \mathbf{F} : дуги $l_i \rightarrow l_{i+N}$, $N = 2, 3, 4$, которые соединяют узлы на графе синтеза. Если не определены критерии для выбора дуги перехода из текущего состояния, то существует $Br!$ вариантов для перехода. Для выбранного перехода вычисляется значение стоимости дуги (евклидово расстояние) $d_{i, i+N}$, которое характеризует близость группы сегментов YSL_i предъявленного РС с траекторией параметров Y_k некоторого слова SL_k .
5. Сумма стоимостей дуг, принадлежащих некоторому пути из начального в конечный узел, определяет интегральное сходство X с X_M^*

$$D = \sum_{i=1}^M d_{i, i+N}. \quad (10)$$

6. Целевое состояние S_g : комбинация ТП Y_i с минимальным значением (10) и соответствующая комбинация слов SL_i , на пути от начального к конечному сегменту предъявленного РС.

В результате анализа элементарных шагов сопоставления N -сегментных ($N = 2, 3, 4$) траекторий параметров слов с ТП предъявленного РС, которая сегментирована на L сегментов-фонем, определена математическая модель генерирования решений-кандидатов задачи (1) без учета порядка следования слов внутри ЭТП:

$$f(\eta, \lambda, \mu) = 2 \cdot \eta + 3 \cdot \lambda + 4 \cdot \mu, \quad (11)$$

где η, λ, μ — количество слов, которые имеют 2-, 3- и 4-сегментные ТП соответственно. С помощью (11) можно оценить количество N var возможных решений-кандидатов для ЭТП: N var вычисляется исходя из количества наборов целочисленных значений $\{\eta, \lambda, \mu\}$, при которых выполняются условия $2 \cdot \eta + 3 \cdot \lambda + 4 \cdot \mu = L$, $\eta, \lambda, \mu < L$, $\eta, \lambda, \mu \geq 0$, т.е.

$$N \text{ var} = \sum_{\eta, \lambda, \mu} f(\eta, \lambda, \mu) \cdot \frac{(\eta + \lambda + \mu)!}{\eta! \cdot \lambda! \cdot \mu!}$$

$$\text{где } f(\eta, \lambda, \mu) = \begin{cases} 1, & 2 \cdot \eta + 3 \cdot \lambda + 4 \cdot \mu = L, \quad \eta, \lambda, \mu \geq 0, \\ 0, & \text{в другом случае.} \end{cases}$$

Обозначим некоторое решение-кандидат таким образом $Sol_j = \{path_j, X_j^*, dist_j\}, j = 1, \dots, N_{var}$, (путь решения $path_j = 0 \rightarrow \dots \rightarrow L$; в S_0 : $path = \{\emptyset\}$), тогда множество решений $Sol = \{Sol_1, Sol_2, \dots, Sol_{N_{var}}\}$. Множество решений для примера графа синтеза (рис. 12) содержит $N_{var} = 5$ возможных путей (рис. 13): $path_1: 0 \rightarrow 2 \rightarrow 4 \rightarrow 7$; $path_2: 0 \rightarrow 2 \rightarrow 5 \rightarrow 7$; $path_3: 0 \rightarrow 3 \rightarrow 5 \rightarrow 7$; $path_4: 0 \rightarrow 3 \rightarrow 7$; $path_5: 0 \rightarrow 4 \rightarrow 5$.

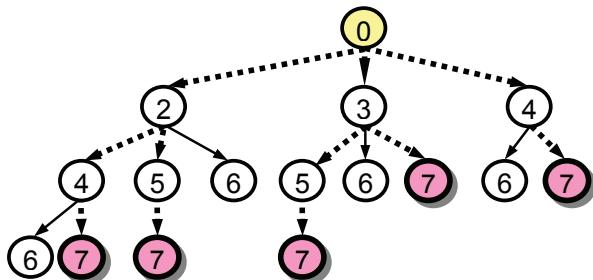


Рис. 13. Множество решений для случая $L = 7$

Зависимость N_{var} от количества сегментов-фонем L в ТП предъявленного РС имеет экспоненциальный вид. Стратегия полного перебора вариантов практически не применима для поиска решений данной задачи вследствие большого объема вычислений при сопоставлении.

Один из способов поиска решений — вычисление стоимостей $d_{i, i+N}$ для всех допустимых переходов $l_i \rightarrow l_{i+N}$ на графике синтеза и нахождение оптимального пути из начальной вершины в конечную с помощью известных алгоритмов поиска кратчайшего пути, например, Дейкстры. Очевидно, что в зависимости от L такой поиск требует вычисления значений стоимостей в количестве

$$Cnitedist(L) = \begin{cases} 1, & L = 2; \\ 2, & L = 3; \\ 4, & L = 4; \\ 3 \times (L - 3), & L \geq 5. \end{cases}$$

Если количество слов в алфавите Z , то в общем случае для нахождения оптимального пути на графике синтеза необходимо осуществить, по меньшей мере, $N_1 = Z \times Cnitedist(L)$ операций сравнения, что в свою очередь увеличивает время распознавания за счет лишних вычислений.

Другой способ использует предположение, что значения стоимостей всех дуг на графике синтеза неизвестны, и для нахождения ЭТП применяют стратегии направленного поиска, позволяющие сократить общее количество вычислений. Очевидно, что в наихудшем случае, используя второй способ поиска решений, необходимо осуществить $N_2 = Z \times Len$ операций сравнения ($Len < Cnitedist(L)$ для $L \geq 5$).

Эвристический алгоритм поиска решений задачи сегментно-слогового распознавания на основе алгоритма эвристического поиска в пространстве состояний, который использует оценочную функцию вида (2) с учетом (9)

и схему группирования (рис. 9), состоит из последовательности шагов А.1-А.7. Введены следующие обозначения: n_s , n_g — начальный и целевой узлы соответственно; nSpisok, pSpisok — списки неразвёрнутых и развёрнутых узлов; CntPall — счётчик общего количества развёрнутых узлов; CntPg — счётчик количества узлов на пути от начального до целевого узла; WsortSpisok — упорядоченный список построенных решений-кандидатов, для которых выполняется условие $D_{gr\ min} \leq D \leq D_{gr\ max}$ ($D_{gr\ min}$, $D_{gr\ max}$, определённые экспериментально границы оптимальности решения); CntW — счётчик в списке WsortSpisok.

A.1. Внести начальный узел $n_s = 0$ в pSpisok, для каждого дочернего узла которого $n = n_s + N (N=2, 3, 4)$ вычислить значение ОФ $f(n)$ (2) и внести эти узлы в nSpisok, $CntPall = 1$; $CntPg = 0$, $CntW = 0$. Особенность вычисления ОФ $f(n)$ на этом шаге — для эвристической оценки состояний $h(n)$ (9) не используется оценка h_v ($h_v(n_s) = 0$, поскольку для начального узла не существует родительских узлов, что делает невозможным её вычисление).

A.2. Если nSpisok пустой — окончание алгоритма, иначе — перейти к шагу А.3.

A.3. Выбрать из nSpisok узел $n = n_f\ min$ с минимальным значением ОФ $f(n)$:

$$f_{min} = \min([f(n_s+2), f(n_s+3), f(n_s+4), \dots], \dots, f(n+2), f(n+3), f(n+4)).$$

Изменить решение-кандидат для X^* , добавив ТП слога из соответствующего α -, β -, γ -подмножества алфавита SL в i -ту позицию (i определяется номером текущего сегмента ТП предъявленного РС X). Вычислить значение стоимостей $D_{i, i+N}$. D (10).

A.4. Если узел $n_{f\ min}$ целевой, т.е. $n_{f\ min} = n_g$, то выполняется операция композиция символьной последовательности-решения задачи распознавания W^* , которая отвечает найденной X^* .

A.5. Если $D < D_{g\ min}$, то завершается работа алгоритма поиска и на выход поступает найденный ответ W^* , в противоположном случае выполняются следующие операции: узел $n = n_{f\ min}$ вносится в pSpisok ($CntPall = CntPall + 1$) и удаляется из nSpisok; если $D_{g\ min} \leq D \leq D_{g\ max}$, то найденное решение вносится в WsortSpisok для следующего этапа принятия решения ($CntW = CntW + 1$); выполняется переход к шагу А.3.

A.6. Если $n_{f\ min} \neq n_g$, развернуть узел $n_{f\ min}$ построив все его дочерние узлы $n = n_{f\ min} + N (N = 2, 3, 4)$. Внести узел $n_{f\ min}$ в pSpisok ($CntPall = CntPall + 1$), удалив из nSpisok. Если для узла $n_{f\ min}$ отсутствуют дочерние узлы, то перейти к шагу А.2, в противоположном случае — к шагу А.7.

A.7. Для каждого дочернего узла $n = n_{f\ min} + N (N = 2, 3, 4)$ вычислить значение ОФ $f(n)$ и внести все узлы в nSpisok. Перейти к шагу А.2.

После завершения эвристического поиска на следующем этапе принятия решения выполняется анализ списка WsortSpisok и формирование списка «потенциальных» $Nlist$ ответов распознавания W^* или устанавливается факт отказа от распознавания. Нахождение списка из $Nlist$ решений-кандидатов для ЭТП X^* из условия наилучшей близости к траектории параметров XE предъявленного РС в смысле (1) осуществляется методом динамического программирования [1].

Построение траекторий параметров решений-кандидатов для ЭТП выполняется для случая СПП XE^* с помощью сегментно-слогового сплайн-синтеза на основе следующей модели.

Пусть известно, что ЭТП для предъявленного РС состоит из ТП R слогов. Модель сегментно-слогового сплайн-синтеза ЭТП в частотной полосе (для СПП) имеет вид:

$$Y(t; \vec{\theta}) = \begin{cases} Y_{1,1}(t; \vec{\theta}_{1,1}), & K_0 \leq t \leq K_1, \\ \dots \\ Y_{R,1}(t; \vec{\theta}_{R,1}), & K_{S-1} \leq t \leq K_S, \end{cases}$$

где

$$Y_{j,k}(t; \vec{\theta}_{j,k}) = a_{j,k} + b_{j,k} \cdot (t - K_s) + c_{j,k} \cdot (t - K_s)^2 + d_{j,k} \cdot (t - K_s)^3,$$

$$\vec{\theta}_{j,k} = \{a_{j,k}, b_{j,k}, c_{j,k}, d_{j,k}, K_s\}, j=1, \dots, R, k=1, \dots, L_j,$$

$$L_j = 2; 3; 4, s=1, \dots, S.$$

В каждой частотной полосе l параметры сплайн-моделей

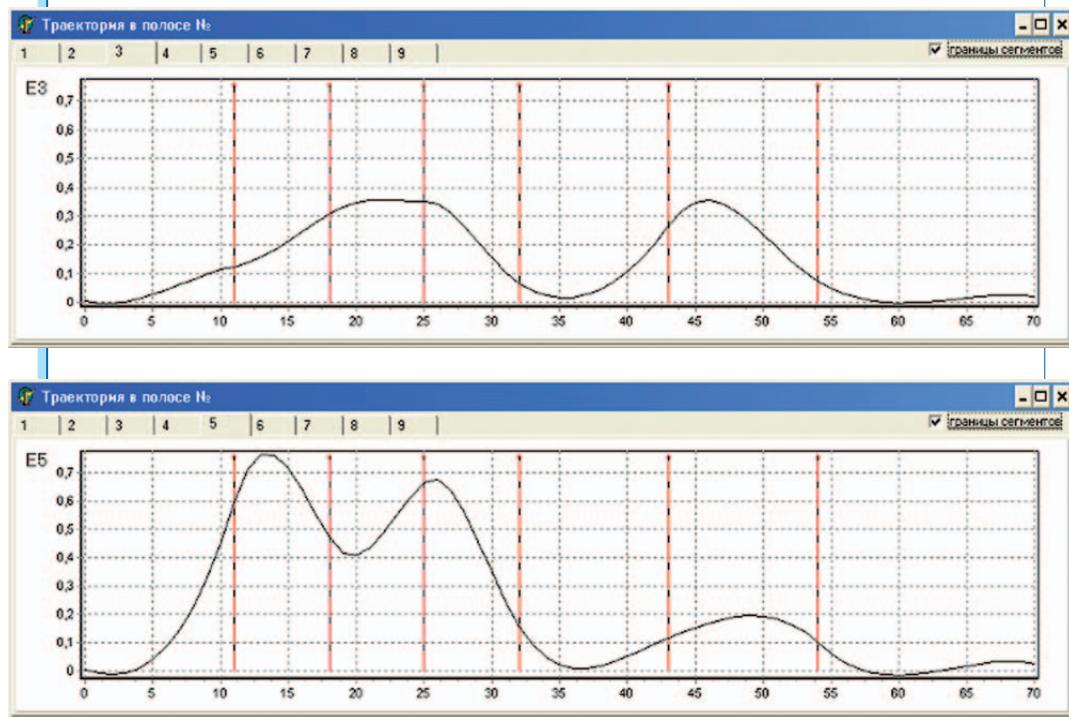
$$\vec{\theta}_{j,k} = \{a_{j,k}, b_{j,k}, c_{j,k}, d_{j,k}, K_s\}, j=1, \dots, R, k=1, \dots, L_j,$$

$L_j = 2; 3; 4$ для траекторий параметров слогов алфавита, составляющих ЭТП, вычисляются из условия минимизации ошибки аппроксимации

$$\sigma_l^2 = \sum_j \sum_k (X_{l,j,k} - Y_{l,j,k})^2 \rightarrow \min \text{ и условий, обеспечивающих гладкость}$$

и непрерывность в узлах сплайн-функций $K = (K_1, K_2, \dots, K_s, \dots, K_S)$, в качестве которых выбраны точки сегментации ТП слогов и точки конката-нации ТП слогов внутри ЭТП.

На рис. 14 приведен пример (для 3, 5, 9 полос СПП) сегментно-слогового сплайн-синтеза ЭТП для предъявленного к распознаванию слова «человек».



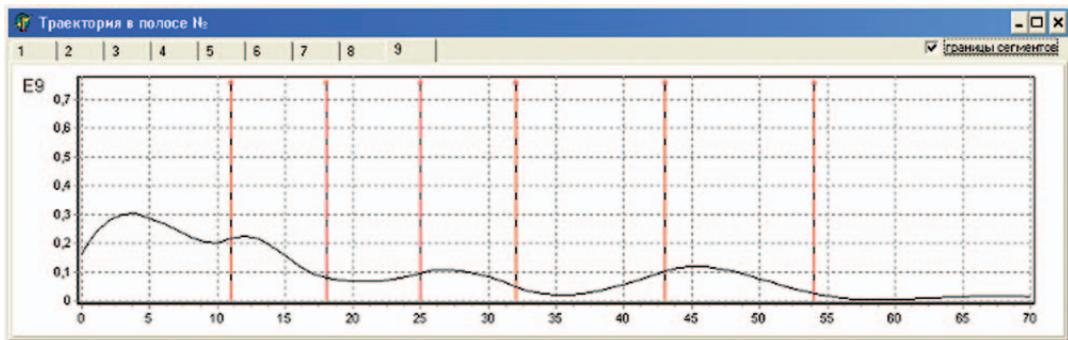


Рис. 14. Пример сплайн-синтеза ЭТП в полосе

2. Поиск решений второго уровня

Для выбранных $Nlist$ решений-кандидатов выполняется сегментно-слоговой синтез ТП XLA^* в классе КФ из ТП YLA слогов алфавита для случая СВП и сопоставление с ТП XA предъявленного РС методом динамического программирования [1]. В результате сопоставления находится такая XLA^* , которая наилучшим образом соответствует XA в смысле (1).

Модель сегментно-слогового синтеза ЭТП, состоящая из R слогов, в классе КФ имеет вид:

$$XLA^* = YLA_{1i} + YLA_{2i} + \dots + YLA_{mi} + \dots + YLA_{Ri}, N'_0 \leq i \leq N'_R \quad (12)$$

где $m = 1, \dots, R$; N_m — количество временных отсчетов ТП YLA_{mi} (4), $t_1 \in [1, N_1], \dots, t_m \in [1, N_m], \dots, t_R \in [1, N_R]$; $N'_0 = N_1 + N_2 + \dots + N_R$.

Алгоритм сегментно-слогового синтеза ЭТП для СВП состоит из шагов Б.1-Б.3 [31].

Б.1. Нахождение области определения ЭТП $D: [\omega_0, \omega_M] \times [t_0, t_N]$.

Временной диапазон ЭТП $[t_0, t_N]$, где $t_N = t_{N_1} + t_{N_2} + \dots + t_{N_R}$. Диапазон частот ЭТП $[\omega_0, \omega_M]$ совпадает с диапазоном частот YLA_{mi} .

Б.2. Восстановление ТП YLA_m выбранных R слогов из алфавита SL согласно (4) на всей области определения ЭТП XLA^* $D: [\omega_0, \omega_M] \times [t_0, t_N]$.

ТП j -го слога определена в прямоугольной области $D_A^j: [\omega_0, \omega_M] \times [t_0, t_{N_j}]$ и представлена параметрами $a_{(i)}, b_{(i)}, c_{(i)}, d_{(i)}, T_{(i)}, Q_{(i)}$ функций $Zt_{(i)}(t_l), Z\omega_{(i)}, (\omega_k)$ (3), $i = 1, \dots, L_j, k = 1, \dots, M, l = 1, \dots, N_j$.

Б.3. Синтез ЭТП XLA^* согласно модели (12).

Описание СВП XLA^* в некоторой точке (ω_k, t_l) представлено в виде суперпозиции R функций YLA_m ($m = 1, \dots, R$), которые являются аналитическим описанием СВП соответствующих m слогов из алфавита SL , таким образом

$$XLA^*(\omega_k, t_l) = \sum_{m=1}^R YLA_m(\omega_k, t_l) = \sum_{m=1}^R \left(\sum_{i=1}^{L_m} Z\omega_{(i)}(\omega_k) \cdot Zt_{(i)}(t_l) \right), \quad (13)$$

где L_m — количество параметров КФ $Zt_i(t_l), Z\omega_{(i)}(\omega_k)$ (3), $i = 1, \dots, L_m$ для соответствующего m -го слога ЭТП.



На рис. 15 — 18 приведен пример реализации сегментно-слогового синтеза СВП для РС «человек».

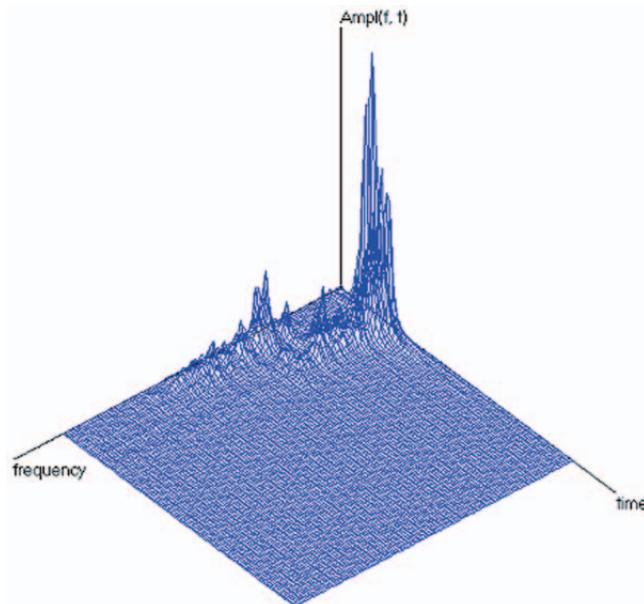


Рис. 15. ТП слова «че» из алфавита, восстановленная на области определения ЭТП

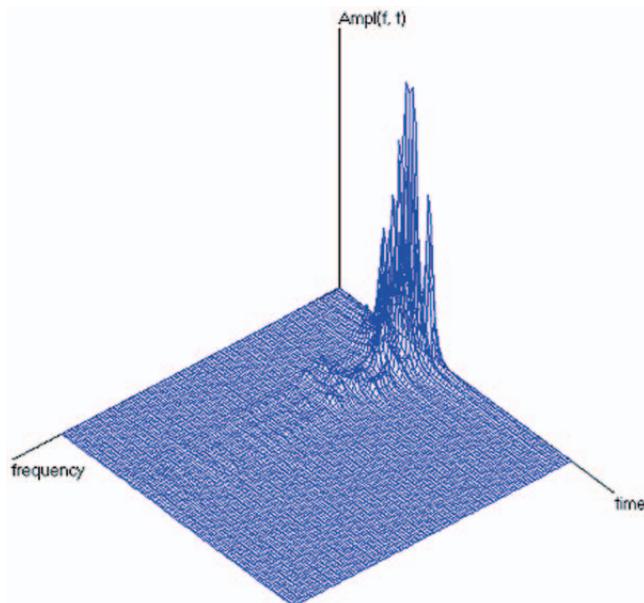


Рис. 16. ТП слова «ло» из алфавита, восстановленная на области определения ЭТП

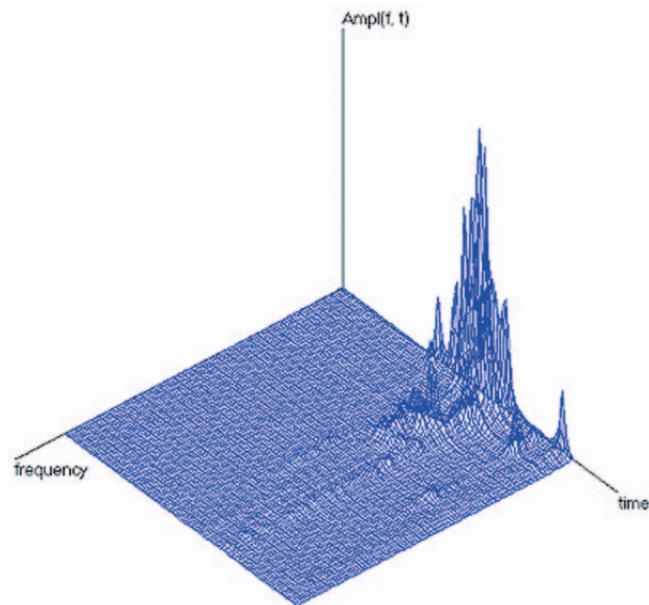


Рис. 17. ТП слова «век» из алфавита, восстановленная на области определения ЭТП

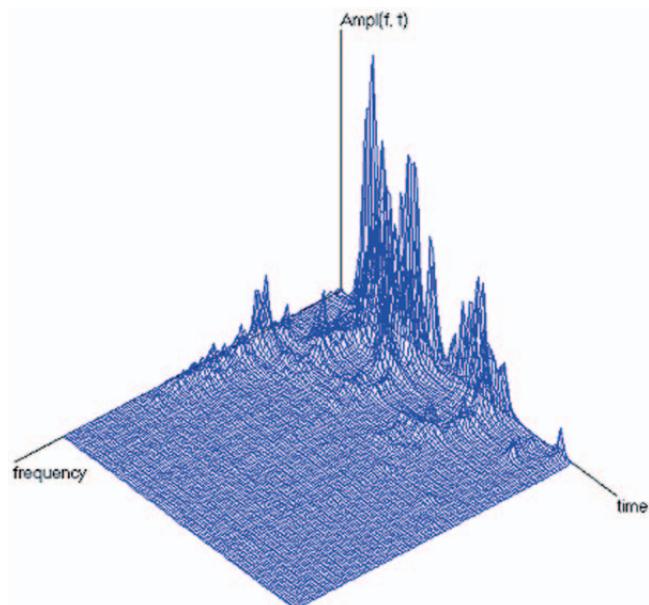


Рис. 18. ЭТП для РС «человек»

3. Композиция символьного ответа распознавания W^*

Для предъявленного РС выполняется композиция символьного ответа распознавания W^* на основе конкатенации (& — операция конкатенации) символьной информации словов алфавита, которые входят в ЭТП XLA^* : $W^* = SL_1 \& SL_2 \& \dots \& SL_R$, или в фонемном виде — $W^* = (P_{1,1} \& P_{2,1} [& P_{3,1}]) \& (P_{1,2} \& P_{2,2} [& P_{3,2}]) \& \dots \& (P_{1,R} \& P_{2,R} [& P_{3,R}])$.



Эксперимент

Для исследований в «SPEach» использована РБД для списка W_{5000} , которая содержит записи РС 10 мужчин и 10 женщин. Записи РБД сгруппированы в следующие выборки: SW_1 (выборка РС для множества слов W_1 , из которых сформировано обучающее множество РЕ); SW_2 (выборка с другими реализациями РС для множества слов W_1); SW_3 (выборка РС для множества слов W_2 , причём W_1, W_2 такие, что $W_1 \cap W_2 = \emptyset$). Средняя длина реализации РС в сегментах без учета продолжительности сегментов: $\bar{L} = 7$. Средняя длительность реализации РС в РБД составляет $\bar{t} = 1,20$ с. Примем это значение сопоставимым с реальным временем.

Проведены исследования адекватности моделей сегментно-слогового синтеза ЭТП, надёжности и быстродействия распознавания РС из РБД. Оценка адекватности моделей сегментно-слогового синтеза ЭТП выполнена по критерию надёжности распознавания для каждой выборки РС. Также вычислена оценка разборчивости синтезированного ответа системы распознавания, которая характеризует качество синтезированной ЭТП (оценка разборчивости измеряется процентом правильно распознанных слов аудиторами [37]). В экспериментах принимали участие 20 человек, которым было предложено записать услышанные слова (объём тестового словаря составил 100 слов). Количество верно распознанных слов аудиторами — 94%, что является приемлемым. Это свидетельствует о том, что эвристический алгоритм поиска решений задачи сегментно-слогового распознавания адекватно выбирает слоги из алфавита, а предложенная модель сегментно-слогового синтеза СВП в классе КФ позволяет с достаточной точностью восстановить ЭТП для распознаваемого РС из ТП слогов алфавита. Результаты исследований каждого этапа поиска эвристических решений приведены в таблице 2 (Tcp — среднее время распознавания одной реализации РС; Err — ошибка распознавания).

Таблица 2

Анализ алгоритмов сегментно-слогового синтеза

Критерии	СВП			СПП		
	SW_1	SW_2	SW_3	SW_1	SW_2	SW_3
Tcp , сек	0,5	0,55	0,45	0,5	0,35	0,40
Err , %	2,70	3,00	5,00	7,00	9,00	15,00
$Nlist$, %	7,00	10,00	10,00	10,00	12,00	15,00

Для предъявленной реализации РС время поиска на каждом этапе в среднем составляет 0,50 сек. Согласно результатам экспериментов, задержка между окончанием ввода РС с микрофона и ответом системы распознавания составляет в среднем 1,00 сек., что является допустимым для современных диалоговых систем. Использование эвристической оценочной функции позволяет уменьшить ошибку распознавания до 5%. Проведен анализ ошибок распознавания РС, в ходе которого выявлено, что большая часть ошибок распознавания связана с ошибками при определении границ сегментов, например, между безударной гласной и сонорными согласными. Поэтому одним из дальнейших направлений исследований является усовершенствование алгоритмов сегментации.

Литература

1. Винцюк Т.К. Анализ, распознавание и интерпретация речевых сигналов. К.: Наукова думка, 1987.
2. Винцюк Т.К. Образный компьютер: Концепции, методология, подходы. // Системы технического зрения и искусственного интеллекта с обработкой и распознаванием изображений. К., 2001. С. 125–138.
3. Кодзасов С.В., Кривнова О.Ф. Общая фонетика. М.: РГГУ, 2001.
4. Карпов О.Н. Технология построения устройств распознавания речи. Д.: Изд-во Днепропетр. ун-та, 2001.
5. Волошин В.Г. Компьютерная лингвистика. С.: ВТД, 2004.
6. Жожикашвили В.А., Петухова Н.В., Фархадов М.П. Компьютерные системы массового обслуживания и речевые технологии. // Проблемы управления. 2006. № 2. С. 3–7.
7. Ronzhin A.L., Yusupov R.M., Li I.V., Leontieva A.B. Survey of Russian speech recognition systems // SPECOM'2006. St. Peterburg, 2006. P. 54–60.
8. Pylypenko V. Information retrieval based algorithm for extra large vocabulary speech recognition // SPECOM'2006. St. Peterburg, 2006. P. 67–69.
9. Кушнир Д.А., Харламов А.А. Распознавание речи в базисе многомерного сигнального пространства // Информационные технологии. 2005. № 5. С. 30–36.
10. Загоруйко Н.Г. Методы распознавания и их применение. М.: Сов. Радио, 1972.
11. Распознавание слуховых образов / Под ред. Н.Г. Загоруйко, Г.Я. Волошина. Н.: Наука, 1970.
12. Рассел С., Норвиг П. Искусственный интеллект. М.: Вильямс, 2006.
13. Карпов А.А. Модели и программная реализация распознавания русской речи на основе морфемного анализа: Автореф. дисс. канд. техн. наук: 05.13.11. Санкт-Петербург, 2007.
14. Егоров А.И., Дубровский В.В. Об анализе слуховых образов речевого сигнала. http://www.kcn.ru/tat_en/science/fccl/ar1.htm
15. Теория слога. www.erudition.ru
16. Kopeček I. Speech Recognition and Syllable Segments. // Workshop on Text, Speech and Dialogue TSD'99. Lectures Notes in Artificial Intelligence. Springer-Verlag, 1999. P. 203–208.
17. Shastri L., Chang S., Greenberg S. Syllable detection and segmentation using temporal flow neural networks // Int. Congress of Phonetic sciences. San Francisco, 1999. P. 138–146.
18. Белянский В.М., Светозарова Н.Д. Слоговая фонетика и три фонетики Л. В. Щербы. <http://www.auditech.ru/doc/cherba.htm>
19. Орлов И.А. Слоговой компиляционный синтез русской речи // Речевая информатика. М.: Наука, 1989. С. 119–139.
20. Лингвистический энциклопедический словарь. М.: Сов. энциклопедия, 1990.
21. Vasylyeva N., Sazhok M. Text selection for training procedures under phoneme units variety // SPECOM'2005, St. Peterburg, June 25 — 29, 2005. St. Peterburg, 2005. P. 629–631.
22. Крак Ю.В., Горбань В.В. Один из подходов к разработке системы автоматического озвучивания текстов на украинском языке // Искусственный интеллект. 2004. № 1. С. 196–203.
23. Togawa F., Hakaridani M., Iwahashi H. Voice activated word processor with automatic learning for dynamic optimization of syllable templates // ICASSP'86. Int. Conf. Acoust., Speech and Signal Process. New York, 1986. Vol. 2. P. 1121–1124.
24. Tsuboi T., Tomihisa A., Sugamura N. Japanese linguistic processing for continuous speech recognition // ICASSP'87. Int. Conf. Acoust., Speech and Signal Process. New York, 1987. Vol. 2. P. 805–808.



Савенкова О.А.

Система сегментно-слогового распознавания изолированных слов из больших словарей

25. Бондарко Л.В., Зиндер Л.Р., Штерн А.С. Некоторые статистические характеристики русской речи // Слух и речь в норме и патологии. Л., 1977. Вып. 2. С. 3–16.
26. Дудник З.В., Затайдух О.В., Майдиков П.В. Использование Perl и Excel для создания базы данных и статистической оценки текстовых массивов в русском и украинском языках. <http://www.philol.msu.ru/~rlc2004/ru/participants/psearch.php?pid=19229>
27. Савенкова О.А., Карпов О.Н. Некоторые эксперименты по повышению надежности распознавания слов заданного словаря // Системные технологии. 2004. Вып. 35. С. 60–66.
28. Романенко С.В., Стромберг А.Г. Классификация математических моделей аналитических сигналов в форме пиков // Журнал аналитической химии. 2000. Т. 55. № 11. С. 1144–1148.
29. Карпов О.Н. Вычислительные схемы представления функций многих переменных в классах функций меньшего числа переменных. Д.: Изд-во Днепропетр. ун-та, 2003.
30. Карпов О.Н., Габович А.Г., Марченко Б.Г. Компьютерные технологии распознавания речевых сигналов. К.: Полиграф-Консалтинг, 2005.
31. Савенкова О.О., Карпов О.Н. Применение колоколообразных функций в алгоритме сегментно-слогового синтеза // Математическое моделирование. 2008. Т. 1(18). С. 5–9.
32. Савенкова О.А., Карпов О.Н. Технология построения интеллектуальной системы распознавания речи // Искусственный интеллект. 2008. № 4. С. 785–795.
33. Де Бор К. Практическое руководство по сплайнам. М.: Радио и связь, 1985.
34. Частотный словарь. <http://www.artint.ru/projects/frqlist.asp>
35. Савенкова О.А. Разработка нейросетевого алгоритма поиска решений задачи распознавания речи // Вестник академии таможенной службы Украины. 2010. № 43. С. 137–144.
36. Савенкова О.А. Вычисление слагаемых функции эвристической оценки состояний в пространстве состояний задачи сегментно-слогового распознавания речи // Труды конф. УкрОбраз'2010. Киев, 2010. С. 69–72.
37. Людовик Т.В., Сажок Н.Н. Использование речевых баз данных большого объема при синтезе речи в системах искусственного интеллекта. // Проблемы управления и информатики. 2003. № 6. С. 82–87.

Сведения об авторе

Савенкова Ольга Александровна —

Область интересов: обработка и распознавание речевых сигналов, нейронные сети. E-mail: 2sol@ukr.net

Метод выделения главных членов предложения в виде предикативных структур, использующий минимальные структурные схемы

Харламов А.А., доктор технических наук

Ермоленко Т.В., кандидат технических наук

Дорохина Г.В., младший научный сотрудник

Гнитъко Д.С., магистрант

В статье дан краткий обзор подходов, используемых при синтаксическом анализе предложений естественного языка, приведено обоснование выбора синтаксического представления предложения в виде предикатной структуры. Для формального описания базовой структуры простого предложения в работе используется предикатная конструкция, реализованная на атрибутивном уровне описания своих составляющих, включающая актанты, объединённые с предикатом системой отношений. Выявление предикативно связанных грамматических субъекта и предиката в простом предложении осуществляется с помощью минимальных структурных схем предложений. На основе минимальных структурных схем предложений строятся соответствующие им шаблоны и далее проводится их поиск в предложении.

- синтаксический анализ • семантический анализ • грамматический предикат • грамматический субъект • актант • валентность предиката • атрибутивный уровень описания • минимальная структурная схема предложения • морфологическая информация • копула.

The paper presents a short overview of approaches, used in syntactic analysis of sentences in the natural language. It also presents an argumentation of the choice of a syntactic representation of a sentence in the form of Subject-Predicate structure. To describe the basic structure of the simple sentence on a formal level the authors use a predicate structure described as a network of its constituents, including actants, being part of the predicate system of relations. Identification of grammatical subject and predicate in the simple sentence is performed using the minimal structural schemes of the sentence. On this basis corresponding templates are built, and then their automatic extraction is carried out.

- prosodic speech features • melodic contour • syntagma • intonational structure • kernel of intonational structure • pitch • microprosody • description of intonational accentual units.

Введение

В последнее время активно используются различного рода интеллектуальные информационные системы, выполняющие обработку текстов на естественном языке (далее ЕЯ). Один из ключевых элементов таких систем — лингвистический процессор. Классиче-



ская структура лингвистического процессора содержит три последовательных блока морфологического, синтаксического и семантического анализа ЕЯ-текста [1].

Морфологический анализ текста на ЕЯ не представляет серьёзных трудностей для программной реализации. Сложность создания механизмов синтаксического и семантического анализа обусловлена в значительной степени отсутствием единой теории языкового общения, охватывающей все аспекты взаимодействия коммуникантов: грамматика ЕЯ принципиально недетерминирована и неоднозначна, синтаксис ЕЯ весьма разнообразен, сложен и произволен. Поскольку полной и строгой формальной модели ни для одного естественного языка пока не создано, при разработке средств общения конечных пользователей используется ограниченный ЕЯ.

Для автоматической обработки трудны такие вполне допустимые в ЕЯ явления, как эллипсис (пропуск обязательных фрагментов предложения в силу возможности их восстановления из предыдущего контекста) и анафора (отношение между словами или словосочетаниями, при котором в смысл одного выражения входит отсылка к другому, ранее упомянутому, языковому выражению). Кроме того, при синтаксическом анализе текста на ЕЯ одна из основных проблем — разрешение неоднозначностей [1, 2]. При разработке синтаксического анализатора существуют два подхода: формально-грамматический и вероятностно-статистический [3].

Методы первого подхода направлены на создание сложных систем правил, которые позволяли бы в каждом конкретном случае принимать решение в пользу той или иной синтаксической структуры. Правила представляются в виде грамматик, задающих синтаксис языка [4, 5]. Хотя такой подход может обеспечить высокую точность анализа, возникают сложности в связи с сильной зависимостью от конкретной грамматики языка. Создание анализатора структурного типа — весьма сложный процесс. Наиболее трудоёмкую часть работы (создание системы правил) выполняет лингвист высокой квалификации.

Главная особенность методов вероятностного типа — отсутствие жёсткой системы синтаксических правил, для создания которой, собственно, и требовалось участие лингвиста. Вместо системы синтаксических правил используется обширный набор примеров предложений, разобранных человеком вручную, для получения статистики встречаемости различных структур в похожем контексте. Этот набор примеров используется для «обучения» статистического распознавателя, опирающегося на известный метод дерева принятия решений [6]. Затраты на разработку вероятностных анализаторов могут быть существенно ниже, чем на создание исчерпывающих структурно-грамматических моделей естественного языка. Однако для функционирования вероятностно-статистических методов необходим представительный банк синтаксических структур, полученный в результате «ручного» синтаксического разбора. Для достижения приемлемой точности анализа их могут потребоваться тысячи. Разновидностью статистических систем синтаксического анализа являются анализаторы, которые используют описание языка в виде моделей управления. Они настроены на работу в заданной предметной области и получены в результате предварительного анализа корпуса текстов этой предметной области. Каждой модели управления приписывается частотность, характеризующая вероятность использования этой модели управления для новых текстов данной области [7].

Таким образом, разработка методов синтаксического анализа ЕЯ-текстов без привязки к конкретному языку и легко адаптируемых под нужды конкретной

предметной области представляет собой важную научную задачу и имеет существенное практическое значение.

Особое и обязательное свойство предложения — предикативность — соотнесённость сообщаемого с действительностью. Предикатная структура простого предложения обуславливается общими принципами воссоздания действительности и не зависит от конкретного языка. Этот вывод имеет далеко идущие последствия как для структурирования речевого материала в общем (общий структурный синтаксис), так и для вопросов автоматической обработки текста (структура базы знаний, формирование лингвистического процессора и т.п.).

Данная статья посвящена решению задачи синтаксического анализа, которая заключается в получении синтаксической структуры входного предложения в виде предикатных структур на основе использования морфологической информации о словоформах, полученной на этапе морфологического анализа.

Предикатная структура как первооснова предложения

Представление о предикатной структуре как первооснове предложения возникло ещё в античный период, когда предложение и суждение строго не разграничивались, их компоненты зачастую отождествлялись. И в предложении, и выражаемой им мысли друг другу противопоставлялись субъект и предикат, поэтому в традиционной грамматикеочно утверждалось представление о двусоставности как важнейшем признаке предложения. Субъектно-предикатная структура предложения играет огромную роль в языке.

В контексте языкознания предикат обозначает то, что высказывается (утверждается или отрицается) о субъекте. Предикат находится в предикативном отношении к субъекту, способном принимать отрицание и различные модальные значения. К понятию предиката предъявляются определённые семантические требования, а именно, предикат — не всякая информация о субъекте, но указание на признак субъекта, его состояние и отношение к другим предметам.

В ряде современных направлений логики понятие «предиката» заменено понятием «пропозициональная функция», аргументы которой представлены актантами (термами) — субъектом и объектами [8].

Грамматический субъект (подлежащее) — ещё одна конститутивная знаковая единица в составе предложения [9, 10]. Его означаемым является, прежде всего, один из семантических актантов с присущей ему ролевой нагрузкой. Наряду с этим, его означаемым часто оказывается логический субъект как представление об исходном предмете мысли. Субъект обеспечивает идентификацию носителя признака. Наложение на функцию одного из нескольких семантических актантов (если их в пропозиции более одного) функции логического субъекта придаёт суждению (и выражающему его предложению) свойство ориентированности и по отношению к предикату, и по отношению к объекту/дополнению (или объектам/дополнениям). Тем самым маркируется выдвижение одного из актантов на роль первого, главенствующего в логическом плане среди равных. Субъект задаёт грамматико-смысловую перспективу предложения.

Формальными признаками грамматического субъекта могут быть его начальная позиция в линейной структуре предложения, а в языках с развитой системой словоизменения — падежные флексии. Так, в языках номинативного строя подлежащее, в основном, представляется именной частью речи в именительном падеже, реже — инфинитивом, который является формальным субъектом. В языках эргативного строя выбор падежа для подлежащего зависит от переходности или непереходности глагола. В языках активного строя для подлежащего при сказуемом со значением действия используется активный падеж, а при глаголах со значением состояния — инактивный падеж.



Грамматический предикат (сказуемое) — вторая конститутивная знаковая единица в составе предложения и может характеризоваться определённым местом в линейной структуре предложения. В его позиции чаще всего выступает глагол [9]. В языках с развитой системой глагольного словоизменения в словоформе глагола выражается набор самых разных граммем, принадлежащих к формоизменительным категориям времени, вида, наклонения, залога, отрицания, вопросительности, а также к согласовательным категориям лица, числа, иногда рода и т.д. Сказуемое может быть также представлено другими предикатными словами (прилагательное, наречие, предикатив, неличные формы глагола), а также существительным (со связкой или без неё). Возможны различные способы усложнения сказуемого. И граница между сложным глагольным сказуемым как целостным членом предложения и сочетанием сказуемого с другими компонентами часто устанавливается произвольно.

В содержательно-ориентированных теориях синтаксиса особо подчёркивается, что на сказуемое, выступающее в качестве ядра, вокруг которого организуется ближайшее окружение, или же на предикативное отношение, связывающее сказуемое с подлежащим, ложится функция актуализации предложения в модально-временном плане, отнесения его содержания к описываемой ситуации действительности, утверждения или отрицания существования этой ситуации. И сама связь между сказуемым и предикатом, и отнесённость предложения в целом к действительности (независимо от наличия или отсутствия в нём подлежащего) характеризуются в терминах «предикация» и «предикативность». Только совокупность средств выражения предикации и референции (пространственно-временной локализации) обеспечивают привязку предложения к действительности, его актуализацию [8].

В связи с вышеизложенным, наличие предикативно связанных грамматических субъекта и предиката многие исследователи считают обязательным свойством предложения. Эти члены предложения квалифицируются как главные, поскольку они формируют предикативную основу предложения, его конструктивный минимум. В конструкциях с безобъектными, непереходными (в широком смысле) глаголами позиция дополнения представлена нулём, т.е. отсутствует. Точно так же может отсутствовать и позиция подлежащего, когда предложение развертывается на основе бессубъектного глагола (*темнеет, морозит*). Субъектная позиция здесь также представлена нулём. В ряде языков появляется нечто вроде формального подлежащего. Бессубъектными следует признать предложения, ядром которых являются событийные имена и имена состояний (*война, пожар, мороз, морозно*).

В рамках данной работы используется модель языка, в которой на синтаксическом уровне предикат — ядерная структура, включающая в свой состав *п* актантов. В общем случае, само ядро — глагольная конструкция, актанты объединяются с ядром системой отношений [10]. Узлами в этой конструкции являются имена (существительное, местоимение, числительное) в их атрибутивной форме. Актанты могут быть представлены или в виде отдельных объектов, или в форме конкретных характеристик предикатора, представленных наречиями (*вчера, сегодня, там, здесь* и т.п.). Изложим описание предикатной структуры предложения более подробно.

Формальное описание базовой структуры простого предложения

Предикатная структура реализуется на объектном уровне, где каждую её составляющую (объект — Obj, субъект — Subj, действие — Pred) человек всегда воспринимает как некоторую целостность, которая всегда реализуется через совокупность своих признаков. Язык имеет средства для представления

этих признаков, с помощью которых разделяются объекты одного класса. Обозначим подобные языковые средства как Attr(Obj). Совокупность средств Obj и Attr(Obj) позволяет задавать полное описание объекта уже на уровне фиксации отдельных признаков.

Иначе говоря, предикатная конструкция, реализованная на атрибутивном уровне описания своих составляющих, является эталонной структурой описания отдельной ситуации окружающей действительности. Этую базовую конструкцию речевой деятельности будем считать простым предложением.

Особенность приведённой конструкции — иерархическая зависимость между лексическими составляющими, поступающими на вход логических схем формирования описания элементов внешнего мира. Здесь явно прослеживаются три уровня формирования описания:

ядро конструкции — имя Obj, Subj или Pred;
атрибутивный уровень — список атрибутов (Attr1(Obj), Attr2(Obj) и т.п.);
уровень меры признака (перечень элементов Attr(Attr)).

Совокупность этих трёх уровней полностью определяет атрибутивный уровень описания объекта или действия и полностью представляет языковую деятельность человека.

Конструкцию, являющуюся атрибутивным уровнем описания объекта или действия, в контексте данной работы будем называть группой существительного или глагола соответственно. Например, в конструкции *очень быстрая ходьба* объект Obj — ходьба, атрибут Attr(Obj) — быстрая, уровень меры признака Attr(Attr) — очень.

Простое предложение — это прежде всего двухсоставная конструкция

Subj — R0 — Pred,

где Subj — активный субъект, который инициирует использование предиката Pred; R0 — отношение «быть субъектом».

Если раскрыть все характеристики предиката (его валентности), то структура простого предложения будет иметь вид:

Subj — R0 — Pred — Ri — Obj_i, i = 1, n,

где Ri — предикативные отношения, n — количество актантов.

Последнее выражение определяет монопредикатную структуру описания отдельной ситуации. Простое предложение — это двусоставная конструкция отображения произвольной ситуации, объединяющая субъект с определённым предикатом, которые синтаксически соотносятся с главными членами предложения. Анализ сложных синтаксических конструкций и текста основывается на возвращении к принципам построения простых предложений [10]. Следовательно, для проведения эффективного синтаксического анализа, в первую очередь, необходимо разработать алгоритм выделения главных членов простого предложения, позволяющий представить их в виде двусоставной конструкции.

Выделение предикативной основы простого предложения

Приведём алгоритм поиска главных членов простого предложения. Для этого введём несколько обозначений, приведённых в таблице 1.

Таблица 1

Формы слов/групп, входящих в предикатные структуры предложений

Форма слова/группы	Обозначение
1. Показатели предикативности	
группа спрягаемой формы глагола (не инфинитив)	V(f)
спрягаемые формы связки — служебных слов быть, стать, являться, значит и т.д.	Cop(f)



Таблица 1 (окончание)

копула (тире, тире + это и т.п.)	Cop
группа инфинитива глагола, или связки	Inf
группа спрягаемой формы глагола 3-го лица единственного числа	V(sn,3)
группа спрягаемой формы глагола 3-го лица множественного числа	V(pl,3)
2. Имена и наречия	
группа имени (существительного, личного местоимения, количественного числительного, прилагательного, для которого нет согласованного с ним существительного) в i-том падеже	NI
группа предложной формы i-того падежа, способная сочетаться со связкой	Nip
группа беспредложной и предложной формы косвенного падежа, способная сочетаться со связкой	N2...p
группа именительного и творительного падежа прилагательных и страдательных причастий	Adj1 и Adj5
группа кратких форм и компаративов прилагательных и страдательных причастий	Adj(f)
наречия, способные сочетаться со связкой (предикативы)	Adv_pr

Главное слово в группе будем обозначать так: <обозначение группы>_1.

Следует обратить внимание на возможные варианты групп V(f) и Inf. В случае наличия в предложении нескольких групп Inf (крайне не хотеть заставить себя прилежно учиться) без копулы между ними, они объединяются в одну. Тогда Inf_1 — конструкция из нескольких инфинитивов (для словосочетания крайне не хотеть заставить себя прилежно учиться Inf_1=не хотеть заставить учиться).

Составное глагольное сказуемое (вспомогательный глагол + инфинитив) будем относить к группе V(f). Для получения составного глагольного сказуемого последовательно анализируется группа V(f) и Inf, $V(f)_1 = V(f)_1 + Inf_1$. Так, в предложении отец начинал сильно беспокоиться $V(f)_1=начинал беспокоиться$.

Введём следующие обозначения:

Subj — слово/группа, являющееся подлежащим.

Pred — слово/группа, являющееся сказуемым.

МИ — морфологическая информация словаформы.

Входные данные: простое предложение в виде

$Pr = ((W1, M1), (W2, M2), \dots, (Wn, Mn))$,

где Wi — написание i-го слова, входящего в предложение; Mi — МИ этого слова.

Выходные данные: ядро предикатной структуры предложения в виде несимметричных пар понятий $\langle ci, cj \rangle$, связанных отношением R0 (быть субъектом), где главное понятие ci — Pred; понятие-ассоциант cj — Subj.

Множество простых предложений русского языка задаётся перечнем минимальных структурных схем предложений (далее МСС), описывающих предикативный минимум предложения [11]. МСС — модель, отвлеченный образец, отражающий способ выражения предикативности.

Идея алгоритма заключается в поиске шаблона, соответствующего одной из МСС. МСС и соответствующие им шаблоны приведены в таблице 2, условные обозначения в шаблонах — в таблице 3. Алгоритм начинает ра-

ботать после того, как сформированы группы (атрибутивный уровень описания объекта/субъекта и действия).

Таблица 2

Минимальные структурные схемы и шаблоны, им соответствующие

№ п\п	MCC	Шаблон MCC	Примеры предложений
1	N1 V(f)	K1	Грачи прилетели
2	N1 Cop(f) Adj1 N1 Cop(f) Adj5 N1 Cop(f) Adj(f)	K2	Ночь тихая (тиха) Ночь тише дня
		KNC_L + KCAdj	Ночь была тихая (тихой, тиха) Ночь была тише дня
3	N1 Cop N1 N1 Cop(f) N1 N1 Cop(f) N5	K3	Маша — красавица
		KNC_L + KNC	Он был студент
		KNC_L + K3_6	Он был студентом
4	N1 Cop N2...p N1 Cop(f) N2...p N1 Cop(f) Adv_pr	KN1_P + K_P_Nobj	Дом — без лифта
		K_Nom_Obj	Подарок — Васе
		KN_Pred	Глаза навыкате
		KNC_L + KCP + K_P_Nobj	Дом будет без лифта
		KNC_L + KC_Pred	Глаза были навыкате
5	Inf V(f)	K5	Курить строго воспрещалось Не мешало б нам встречаться чаще
6	Inf Cop(f) N5 Inf Cop N1	KCI_Nom + K3_6	Дозвониться было проблемой
		K6	Любить иных — тяжёлый крест
7	Inf Cop(f) Adj1 Inf Cop(f) Adj5 Inf Cop(f) Adj(f)	KCI_Nom + KCAdj	Промолчать — самое разумное Промолчать было самым разумным Промолчать — разумно
8	Inf Cop N2...p Inf Cop(f) N2...p Inf Cop(f) Adv_pr	KI_P + K_P_Nobj	Промолчать — не в его правилах
		KI_Pred	Молчать некстати
		KCI_Nom + KCP + K_P_Nobj	Отвечать было в его правилах
		KCI_Nom + KC_Pred	Идти было трудно
9	Inf Cop Inf Inf Cop(f) Inf	K9	Отказаться — обидеть хозяина
		KCI_Nom + KCI	Отказаться было обидеть
10	Cop(pl) N2...pr Cop(pl) Adv_pr	KCP + K_P_Nobj	Дома были в слезах
		KC_Pred	С ним были запросто
11	Cop(f) N1 N1	KNC	Будет дождь. Была зима
		_11	Шепот. Робкое дыхание. Тишина
12	Cop(sn,3) Adj(f)	K12	Ночью будет морозно
13	Cop(pl,3) Adj(f)	K13	Результатом были довольны. Отказом были обижены
14	N(2–6)p Cop(sn,3) N2...p Cop(sn,3) Adv_pr	K_P_Nobj	На улице без осадков
		KCP + K_P_Nobj	Завтра будет без осадков
		KC_Pred	Было поздно.
15	V(sn,3)	_15	Скрипело, свистало и выло в лесу. Ему нездоровится. У него кипело на сердце
16	V(pl,3)	_16	За столом зашумели. Его обидели
17	Inf	_17	Не нагнать тебе бешеной тройки. Быть рекам чистыми



Таблица 3

Условные обозначения в МСС и шаблонах МСС

Обозначение	Описание
K+индекс	Индекс соответствует номеру МСС, указанному в таблице 2, К означает наличие координационной связи между словами в предложении
_индекс	Предложение односоставное: субстантивное или с простым сказуемым
Cop	Наличие копулы в явном виде в предложении
Pred	Наличие предикатива в предикатной структуре
Nobj	Главное слово группы N1 в объектном падеже ($I \neq 1$)
Nom	Главное слово группы N1 — номинатив
I_Nom	Инфинитив является номинативом
V(pl,3)	Форма глагола множественного числа 3-его лица
V(sn,3)	Форма глагола единственного числа 3-его лица
Обозначение конструкции	Описание конструкции
KNC_L	Главное слово группы N1 стоит слева от копулы
KCAdj	Копула + зависимое слово, которое является компаративом или краткой либо полной формой прилагательного в именительном или творительном падеже
KNC	Копула + зависимое слово, которое является главным словом группы N1
K3_6	Копула + зависимое слово, которое является главным словом группы N5 (используется в МСС 3 и 6)
KN1_P	Главное слово группы N1 управляет предлогом
K_P_Nobj	Предлог управляет главным словом группы Nobj
K_Nom_Obj	Главное слово группы N1 + главное слово группы Nobj
KN_Pred	Главное слово группы N1 + предикатив
KCP	Копула управляет предлогом
KC_Pred	Копула + предикатив
KCI_Nom	Копула + инфинитив
KI_P	Инфинитив управляет предлогом
KI_Nom_Obj	Инфинитив + группа Nobj
KI_Pred	Инфинитив + предикатив

Пример работы алгоритма

Сочинять музыку — значит поручить цапфенштетсерскому оркестру исполнить хор ангелов (Т. Манн «Доктор Фаустус»).

Объединяя в одну группу Inf неразделённые Cop инфинитивы «поручить» и «исполнить», в итоге получаем Inf_1=«поручить исполнить».

Копула в явном виде (тире), до неё и после — инфинитив, получаем шаблон: KCI_Nom + KCI, который соответствует МСС9.

Результат работы алгоритма: <«поручить исполнить», «сочинять»>

Таблица 4

Выходные данные, соответствующие найденным шаблонам МСС

№ п\п	МСС	Шаблон МСС	Выходные данные
1	N1 V(f)	K1	<V(f)_1, N1_1>
2	N1 Cop(f) Adj1 N1 Cop(f) Adj5 N1 Cop(f) Adj(f)	K2	<Cop+Adj1, N1_1> <Cop+Adj5, N1_1>
		KNC_L + KCAdj	<Cop+Adj(f), N1_1>
3	N1 Cop N1 N1 Cop(f) N1 N1 Cop(f) N5	K3	<Cop+N1_1, N1_1>
		KNC_L + KNC	<Cop+N1_1, N1_1>
		KNC_L + K3_6	<Cop+N5_1, N1_1>
4	N1 Cop(f) N2p N1 Cop N2p N1 Cop N5p N1 Cop(f) N5p N1 Cop(f) Adv_pr	KN1_P + K_P_Nobj	<Cop+предлог+Nobj_1, N1_1>
		K_Nom_Obj	<Cop+Nobj_1, N1_1>
		KN_Pred	<Cop+Pred, N1_1>
		KNC_L + KCP + K_P_Nobj	<Cop+предлог+Nobj, N1_1>
		KNC_L + KC_Pred	<Cop+Pred, N1_1>
5	Inf V(f)	K5	<V(f)_1, Inf_1>
6	Inf Cop(f) N5 Inf Cop(f) N1	KCI_Nom + K3_6	<Cop+N5_1, Inf_1>
		K6	<Cop+N1_1, Inf_1>
7	Inf Cop(f) Adj1 Inf Cop(f) Adj5 Inf Cop(f) Adj(f)	KCI_Nom + KCAdj	<Cop+Adj_1, Inf_1>
8	Inf Cop(f) N2p Inf N2p Inf Adv_pr Inf Cop(f) Adv_pr Inf Cop Adv_pr	KI_P + K_P_Nobj	<предлог+Nobj_1, Inf_1>
		KI_Nom_Obj	<предлог+N1_1, Inf_1>
		KI_Pred	<Cop+Pred, Inf_1>
		KCI_Nom + KCP + K_P_Nobj	<Cop+предлог+Nobj_1, Inf_1>
		KCI_Nom + KC_Pred	<Cop+Pred, Inf_1>
9	Inf Cop(f) Inf Inf Cop Inf	KCI_Nom + KCI	<Cop+Inf_1, Inf_1>
			<Cop+Inf_1, Inf_1>
10	Cop(pl) N2...pr Cop(pl) Adv_pr	KCP + K_P_Nobj	<Cop(pl)+N2...pr_1, 0>
		KC_Pred	<Cop(pl)+Adv_pr, 0>
11	Cop(f) N1 N1	KNC	<Cop+N1_1, 0>
		_11	<Cop+N1_1, 0>
12	Cop(sn,3) Adj(f)	K12	<Cop(sn,3)+Adj(f), 0>
13	Cop(pl,3) Adj(f)	K13	<Cop(pl,3)+Adj(f), 0>
14	N2p Cop(sn,3) N2p Cop(sn,3) Adv_pr	K_P_Nobj	<Cop(sn,3)+предлог+Nobj_1, 0>
		KCP + K_P_Nobj	<Cop(sn,3)+предлог+Nobj_1, 0>
		KC_Pred	<Cop(sn,3)+Adv_pr, 0>
15	V(sn,3)	_15	<V(sn,3)_1, 0>
16	V(pl,3)	_16	<V(pl,3)_1, 0>
17	Inf	_17	<Inf, 0>

Использование МСС в качестве формального образца позволяет получить предикативную основу (структурную схему) простого предложения, и в дальнейшем — его предикатную



структурой. Это первый и обязательный шаг для проведения первичного семантического анализа в формировании информационного портрета текста, поскольку смысловая связь между понятиями предложения (объектом/субъектом) в общем случае может быть описана предикатом, актантами которого выступают данные понятия. Установление таких синтаксико-семантических связей позволяет сформировать схему ситуации, описываемой во фразе.

Обусловленный валентностью предиката семантико-синтаксический уровень анализа конструкций, не соответствующий узкому собственно формально-синтаксическому подходу, даёт возможность даже из набора неправильных форм (посредством приведения их к начальным формам) с помощью заполнения валентных гнёзд определить схему предложения.

Семантико-синтаксический анализ предложения предусматривает создание электронного словаря валентности глаголов. При этом для каждого глагола (около 20 тысяч в русском языке) необходимо указать, какими падежами и с какими предлогами он может управлять, а также в каких семантических ролях (семантических падежах) выступают актанты глагола. Разработкой такого словаря для русского языка авторы планируют заняться в ближайшем будущем.

Литература

1. Волкова И.А. Введение в компьютерную лингвистику. Практические аспекты создания лингвистических процессоров. М.: Издательство ВМиК МГУ, 2006.
2. Ножов И.М. Морфологическая и синтаксическая обработка текста (модели программы). М.: Наука, 2003.
3. Евдокимова И.С. Естественно-языковые системы: курс лекций. Улан-Удэ: Издательство ВСГТУ, 2006.
4. Ахо А., Сети Р., Ульман Дж. Компиляторы: принципы, технологии и инструменты. М.: Вильямс, 2001.
5. Волкова И.А., Руденко Т.В. Формальные грамматики и языки. Элементы теории трансляции. М.: Изд-во МГУ, 1999.
6. Андреев А.М., Берёзкин Д.В., Брик А.В., Кантонистов Ю.А. Вероятностный синтаксический анализатор для информационно-поисковой системы [Электронный ресурс]. http://www.inteltec.ru/publish/articles/textan/1kx5_9.shtml.
7. Волкова И.А., Мальковский М.Г., Одинцов Н.В. Адаптивный Синтаксический анализатор // Диалог 2003: Труды Международного семинара. М., 2003, Т. 1. С. 401–406.
8. Сусов И.П. Введение в языкознание. М.: Восток-Запад, 2006.
9. Загнітко А.П. Теоретична граматика української мови: Синтаксис: Монографія. Донецьк: ДонНУ, 2001.
10. Загнітко А.П. Теоретична граматика української мови. Морфологія. Донецьк: ДонНУ, 1996.
11. Современный русский язык: Учебник для филологических специальностей высших учебных заведений / В.А. Белошапкова, Е.А. Брызгунова, Е.А. Земская и др.; Под ред. Белошапковой. 3-е изд., испр. и доп. М.: Азбуковник, 1997.

Сведения об авторах

Харламов Александр Александрович —

доктор технических наук, старший научный сотрудник Института высшей нервной деятельности и нейрофизиологии РАН. Область научных интересов: нейроинформатика, распознавание речи, анализ текстов, распознавание изображений, семантические представления, искусственные нейронные сети.

Ермоленко Татьяна Владимировна —

кандидат технических наук, научный сотрудник отдела распознавания речевых образов Института проблем искусственного интеллекта МОНМС и НАН Украины. Распознаванием и обработкой речевых сигналов занимается с 2002 года. К области интересов также относится автоматическая обработка ЕЯ-текстов.

Дорохина Галина Владимировна —

младший научный сотрудник Института проблем искусственного интеллекта МОНМС и НАН Украины. Область научных интересов: распознавание образов, автоматический морфологический и синтаксический анализ текстов, ассоциативная память, искусственный интеллект.

Гнитъко Дмитрий Сергеевич —

магистрант Института информатики и искусственного интеллекта Донецкого национального технического университета. Область научных интересов: автоматический синтаксический анализ текстов, искусственный интеллект, формально-грамматический метод.



Распознавание речевых образов суб-словного уровня в слитной украинской речи

Васильева Н.Б., научный сотрудник

В статье описывается пример разработки экспериментальной системы распознавания речевых образов, являющихся составляющими слов. В основу системы положена скрытая Марковская модель. Большое внимание уделяется созданию речевого корпуса как относительно компактной обучающей выборке, в которой представлено всё звуковое разнообразие языка. Проводится оценка параметров акустической модели на основе созданного речевого корпуса. Наряду со свободным порядком следования речевых образов анализируется способ ограничения следования фонем на основе статистической модели. Выбираются коэффициенты, компенсирующие несоответствие шкалы акустической и лингвистической составляющих модели распознавания. Описывается разработанный инструментарий, приводятся результаты экспериментальных исследований.

- суб-словный • слог • фонемы-трифоны • распознавание речи • слитная украинская речь • обучающая последовательность.

The paper presents advances in a multi-level multi-decision automatic speech understanding approach that is initially developed for highly inflective languages with relatively free word order. On the first level a sub-word-based grammar phoneme recognizer is applied, which output is post-processed at the second level. The research is concentrated on the phoneme level aiming to apply the lexical level to the phoneme level output. The idea to use sub-word units for recognition vocabulary appears productive since word lexicon growth leads to practically no new sub-word items, so we can say about the alphabet of sub-words. The ways to select a set of sub-word units like syllables are considered. Both ways operate with a text corpus converted to sequences of phonemes and syllables. To analyze a phoneme error rate we consider a free sub-word order grammar integrated to the HMM-based decoder. The proposed procedure to select a set of about 18000 sentences containing all phoneme-triphones allowed for creation the text for training corpus that has been read by the ordinary speaker. Three control sets were formed by different ways. Acoustic model built for the basic phoneme alphabet is complemented with grammar-based language models for two types of syllables. The recognition accuracy has been compared to free phoneme order grammar. The obtained results show the promising input for the next lexical level of the multi-level automatic speech understanding system. Experimental results, problems and future research are discussed.

- sub-word • syllable • phoneme-triphone • speech recognition • continuous Ukrainian speech • training set.

Введение

Общепринятые системы пофонемного распознавания оперируют алфавитом фонем, из которых состоят речевые образы слов. На слова уже накладываются ограничения их следования путём построения грамматик или лингвистической модели (далее ЛМ). При обогащении лексики увеличиваются объёмы рабочего словаря, существенно усложняется грамматика или ЛМ. Это приводит к уменьшению продуктивности системы распознавания.

Если использовать вместо слов речевые образы частей слова (слогов или морфем), то обогащённая лексика не приведёт к заметному возрастанию рабочих словарей и усложнению грамматик или ЛМ.

В этом случае самой большой препятствием будет переход от последовательности слогов (морфем) к последовательности слов, поскольку ошибка распознавания слогов или морфем может создать ситуацию, когда их последовательностям напрямую невозможно сопоставить слово. Также сама процедура перехода к последовательности слов неоднозначна и недостаточно исследована.

В работе [1] исследовались надёжность распознавания монофонов и двух видов слогов. Для проведения экспериментальных исследований использовался многодикторный речевой корпус отдельно произнесённых слов. Обучающая выборка (далее ОВ), сформированная на основе этого корпуса, состояла из относительно небольшого количества изолированных слов. Использовался словарь только на 4000 слов по имеющимся около 20 тыс. реализаций этих слов, произнесёнными 70 дикторами. Результаты показали перспективность исследований послогового распознавания. Вместе с тем очевидны ограничения при использовании ОВ из изолированных слов.

С учётом результатов было решено сформировать корпус слитной речи, в котором бы наблюдалось всё разнообразие звуков украинского языка. На основе этого корпуса планировалось проводить эксперименты, сравнивать результаты распознавания как отдельно произнесённых слов, так и слитой речи для различных речевых образов. Присутствие в корпусе лишь одного диктора не должно ограничивать проводимые исследования, поскольку в современных системах распознавания применима процедура адаптации акустических моделей фонем к голосу диктора.

Цель данной работы — поиск путей повышения фонемной надёжности распознавания слитой речи, что создаст предпосылки реализации эффективных алгоритмов перехода от последовательности слогов (морфем, фонем) к словам [2].

Формирование текста для акустической базы обучающей выборки

Для проведения как обучения, так и тестирования распознавания необходимо иметь широкую экспериментальную базу, в которую входят:

- однодикторные или многодикторные обучающие и контрольные выборки (далее ОВ и КВ) для исследования индивидуализированного и кооперативного распознавания;
- текстовый корпус для формирования алфавита речевых образов и текстов для записи речевого корпуса.

Формирование речевой базы данных и знаний требует больших временных затрат, в особенности детальной подготовки текста ОВ, содержащей широкое разнообразие элементов (фонем-трифонов или слогов). Для её получения использовались тексты, которые находятся в свободном доступе в Интернете, в основном художественные сочинения украинских авторов, публицистические сочинения, новости, исторические справки. Исключались стихотворные произведения: стихи читаются с особенностями, не свойственными повседневной речи (нестандартное словарное ударение, интонация, ритм). Для ОВ



с изолированными словами использовали частотный словарь украинского языка и словарь УМИФ [3].

«Слитная» ОВ

В процессе формирования текста ОВ были произведены такие действия:

- предварительная обработка текстов (709 файлов; ~ 50 МБ): удаление примечаний, номеров разделов, замена сокращений и т.д.;
- выделение предложения в отдельную строку;
- преобразование орфографического текста в фонемный [4];
- прореживание первоначального корпуса (для каждого из элементов выбираются самые короткие предложения);
- обработка полученного прореженного корпуса результата «жадным» алгоритмом (далее ЖА) [5].

В выбранные таким образом предложения попадают те, которые содержат новую фонему-трифон и являются самыми короткими из рассматриваемых. В результате получаем существенное сокращение текста ОВ, не теряя фонемного разнообразия.

Графики встречаемости фонем-трифонов в разных источниках (текстовый корпус, словарь УМИФ и частотный словарь) и полученных соответствующих ОВ приведены на рис 1. Здесь мы видим, например, что при работе ЖА количество элементов, встречающихся один раз, увеличивается в несколько раз для каждой ОВ. Также из рисунка следует, что частота фонем-трифонов соответствует распределению Ципфа — Мандельброта, как для исходных корпусов, так и после работы ЖА.

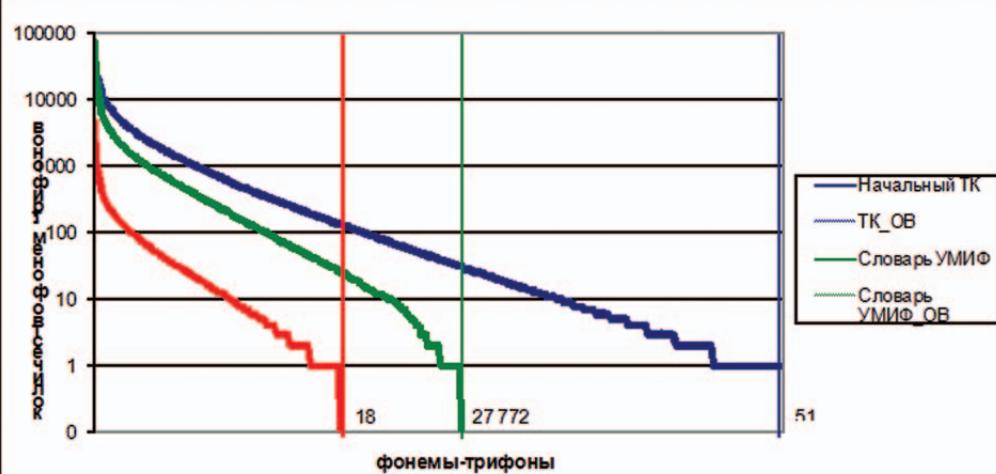


Рис. 1. Распределения фонем-трифонов по встречаемости в текстовых выборках

На первом этапе (обработка текстового корпуса и формирование ОВ) рассматривались фонемы-трифоны как речевые образы, поскольку они имеют регулярную структуру и дают возможность моделировать фонемное разнообразие, учитывая правый и левый контексты [6]. Структурно фонема-трифон имеет три символа в отличие от слогов, которые могут содержать разное количество символов из алфавита фонем: от одной до шести фонем в слогах и до пяти в открытых слогах.

В таблице 1 показана статистика по фонемам-трифонам в ОВ, оптимизация при работе ЖА.

Следующий этап создания ОВ — записывание речи по сформированному тексту обучающей выборки. Во время записи проводится апробирование полученных результатов на удобство чтения, проверка транскрипций, выявление ошибок, которые не обнаруживаются автоматически и мешают нормальному произношению диктора и т.д.

При обработке текстов не было возможности учесть позднее выявленные проблемы:

- ошибочно написанные фразы-предложения (написаны орфографически правильно, но лишённые семантики);
- опечатки;
- визуальная схожесть букв в кириллице и латыни: а, о, е, у, і, р, с, х, Е, Х, Н, В, А, О, Р, М, Т;
- буква вместо цифры (в основном, это касается римских обозначений цифр) и наоборот;
- сокращение типа 1-ї, 1-го, 1-е, 1-й, 1-м, 1-му, 1-ої, 1-у;
- написание и произношение слов, обозначенных цифрами (календарные даты, дробные числа и другое);
- изолирование буквы, например, в конце фразы, после которой стоит многоточие, или перед цифрой и т.д.

Таблица 1

Сравнение количества элементов (в тысячах) в начальном корпусе и тексте ОВ

Начальный текстовый корпус, с которого выбиралась ОВ	Общее количество предложений (слов для словарей) до работы ЖА	Общее количество предложений (слов) после работы ЖА	Общее количество реализаций фонем-трифонов до работы ЖА	Общее количество реализаций фонем-трифонов после работы ЖА	Алфавит фонем-трифонов
Текстовый корпус	816,0	18,0	41 179,8	1 020,3	51,4
Словарь УМИФ	1 874,7	13,7	23 734,3	120,1	27,7
Частотный словарь	137,6	8,2	1 488,0	71,0	18,3

«Словарные» ОВ

На основе словарей были составлены две ОВ: одна на основе словаря УМИФ, другая — на частотном словаре украинской речи.

При обработке словарей также была проведена предварительная работа перед записью: были удалены слова, содержащие одинаковые фонемы-трифоны.

Фонем-трифонов, принадлежащих обеим словарным выборкам, 15 431 элементов. При этом 12 327 фонем-трифонов принадлежат только ОВ на основе словаря УМИФ, а 2 916 — только ОВ частотного словаря.

Запись ОВ и последующих выборок проводилась с помощью модуля *Sigrs* [7] на звуковой карте *Creative Audigy2 ZS* гарнитурой *SteelSeries 5H v2*. Получено около 36 часов записи слитой речи. Объём словаря ОВ — 47 621 слов. Общее количество реализаций слов в ОВ — 184 910.

В процессе записи наблюдались такие физиологические и психолингвистические явления:

- уставление голосового тракта;
- изменение голоса в разных жизненных ситуациях (заболевание, волнение, время суток и т.д.);



- специфика произношения некоторых словосочетаний и словоформ (редуцирование и ассимиляция по глухоте и звонкости согласных звуков).

Объём словаря ОВ отдельных слов составлял 12 870 слов, около 12 часов записи.

Формирование контрольной выборки

Для проверки предложенных речевых образов, т.е. фонем, открытых слогов и слогов, полученных по правилам деления слогов, были сформированы тексты контрольной выборки (далее КВ) слитой речи и проведена её запись.

Решено было провести тестирование на трёх КВ, сформированных разными способами.

«Частотная» КВ

Первый способ выбора КВ основан на проверке распознавания часто употребляемых слов, предложений, фраз, т.е. формирование КВ по частоте фонем-трифонов.

Алгоритм получения «частотной» КВ:

- из начального текстового корпуса удаляется текст ОВ;
- оставшийся текстовый корпус подвергается тем же процедурам обработки, что и ОВ (см. выше);
- из этих предложений выбирается некоторое количество первых предложений (в нашем случае — 3000);
- удаляются повторяющиеся предложения.

Запись проводилась в тех же условиях, что и запись ОВ.

Полученная КВ имеет 3,6 часа записи. Объём словаря составляет 3225 слов. Общее количество реализаций слов — 8987.

«Случайная» КВ

Второй способ — сформировать КВ случайным образом, из тех же текстов, из которых выбирался текст ОВ, но с запрещением выбора тех предложений, которые вошли в ОВ.

Алгоритм получения «случайной» КВ:

- из начального текстового корпуса удаляется текст ОВ;
- из оставшегося текстового корпуса случайным образом берётся некоторое количество предложений (в нашем случае — 2000);
- удаляются повторяющиеся предложения.

Запись проводилась в тех же условиях, что и запись ОВ.

Полученная КВ имеет 4,3 часа записи. Объём словаря составляет 10 013 слов. Общее количество реализаций слов — 22 864.

КВ «Википедия»

Эту КВ предложено выбирать из текстов, которые не использовались ни для выбора предыдущих КВ, ни для ОВ. Для этого из сайта украиноязычной Википедии [8] случайным образом выбрано 100 МБ текстов.

Алгоритм получения КВ «Википедия»:

- с текстов сайта Википедия удалены предложения, которые встретились в ОВ и предыдущих КВ;

- из оставшихся текстов случайным образом выбирается 1000 предложений;
- удаляются предложения, которые повторяются;
- добавлено 200 последовательных предложений из одной случайно выбранной статьи.

Запись проводилась в тех же условиях, что и запись ОВ.

Полученная КВ имеет 3,0 часа записи. Объём словаря составляет 7330 слов. Общее количество реализаций слов — 16 073.

Экспериментальное распознавание и сравнение полученных результатов

Было проведено оценивание параметров акустических моделей с использованием программного инструментария *HTK* [9]. Акустические модели формировались на основе контекстно-независимых фонем. Поскольку объём их алфавита небольшой, а значит, для статистических оценок необходима меньшая база акустических сигналов, чем для слогов и фонем-трифонов, которых больше в тысячи раз и топология их акустических моделей требует дополнительных исследований. Для каждого из 57 фонем-монофонов украинской речи и двух фонем-пауз получены модели, имеющие каждая три состояния и от 4 до 36 смесей нормальных законов в зависимости от частотности.

Декодер пытается найти последовательность суб-словных элементов $\mathbf{q}_{1:L} = q_1, \dots, q_L$, которые наиболее правдоподобно генерируют последовательность наблюдаемых веторов $\mathbf{Y}_{1:T} = \mathbf{y}_1, \dots, \mathbf{y}_L$, исходя из интегральной меры схожести:

$$\hat{\mathbf{q}} = \arg \max_{\mathbf{q}} \{ \log p(\mathbf{Y} | \mathbf{q}) + (\alpha \log(P(\mathbf{q})) + \beta |\mathbf{q}|) \},$$

где α и β — коэффициенты, компенсирующие несоответствие шкалы акустической и лингвистической составляющих модели распознавания. Поэтому на первом этапе проводились эксперименты с целью эмпирически подобрать параметры α и β , рекомендуемый диапазон которых составляет 0–20 и 0 — (— 20) соответственно [9, 11].

При оценке надёжности использовались показатели пофонемной ошибки (*PER* — *Phoneme Error Rate*):

$$\%PER = 100\% - \frac{H - I}{N} 100\%$$

и пофонемной некорректности (*PIR* — *Phoneme Incorrectness Rate*):

$$\%PIR = 100\% - \frac{H}{N} 100\%,$$

где: H — количество правильно распознанных суб-словных элементов,

I — количество ошибочно вставленных суб-словных элементов,

N — общее количество произнесённых суб-словных элементов.

На рис. 2–7 проиллюстрированы показания $\%PER$ и $\%PIR$ пофонемного распознавания описанных выше трёх КВ при изменениях коэффициента β в пяти точках (0, — 5, — 10, — 15, — 20) для α , равное 0, 5 и 10.

Убывание *PER* происходит главным образом за счёт уменьшения вставленных суб-словных элементов, которых не должно быть. Рост некорректности обусловлен уменьшением правильно распознанных элементов. Из рисунков следует, что наименьшая фонемная ошибка достигается при значениях параметров $\alpha = 5$ и $\beta = -5$. Показатель корректности *PIR* дал возможность определить, что надёжность возросла вследствие сокращения числа вставок.

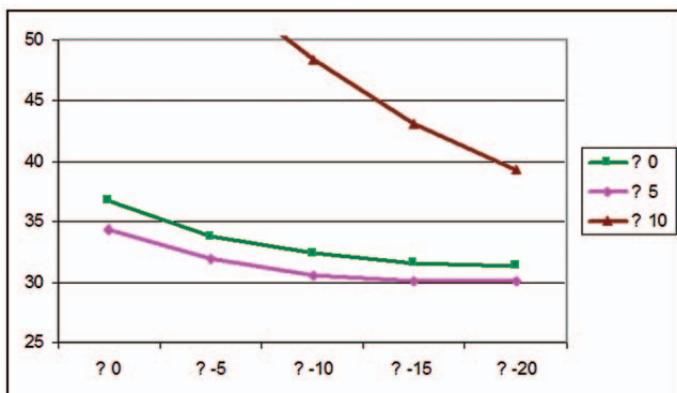


Рис. 2. Показатели PER распознавания (%) для слитной речи на «частотной» КВ

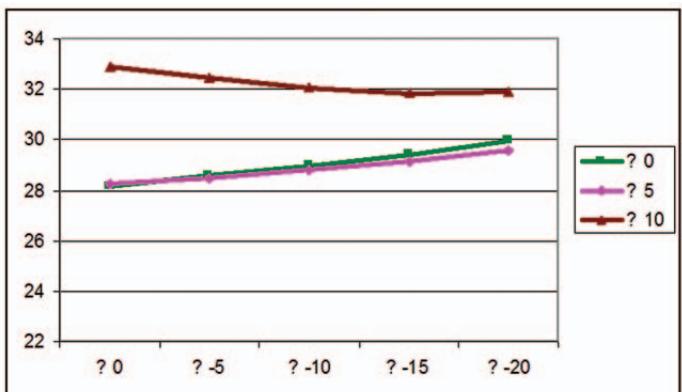


Рис. 3. Показатели PIR распознавания (%) для слитной речи на «частотной» КВ

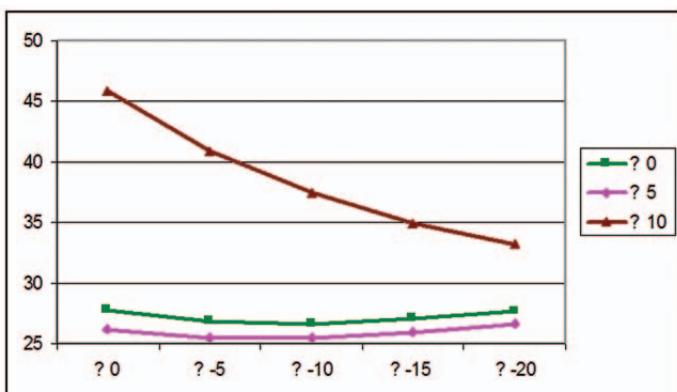


Рис. 4. Показатели PER распознавания (%) для слитной речи на «случайной» КВ

Рис. 5. Показатели PIR распознавания (%) для слитной речи на «случайной» КВ

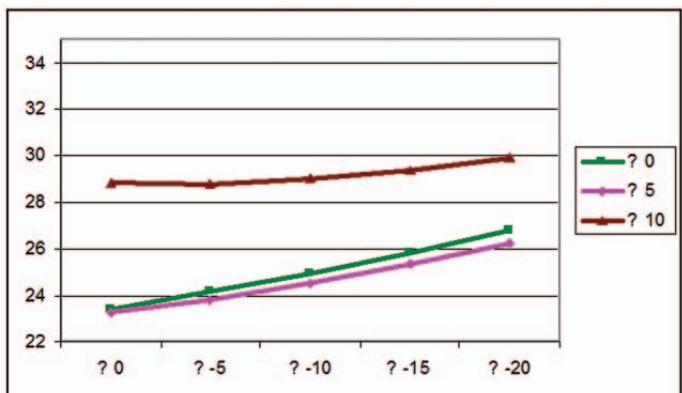


Рис. 6. Показатели PER распознавания (%) для слитной речи на КВ «Википедия»

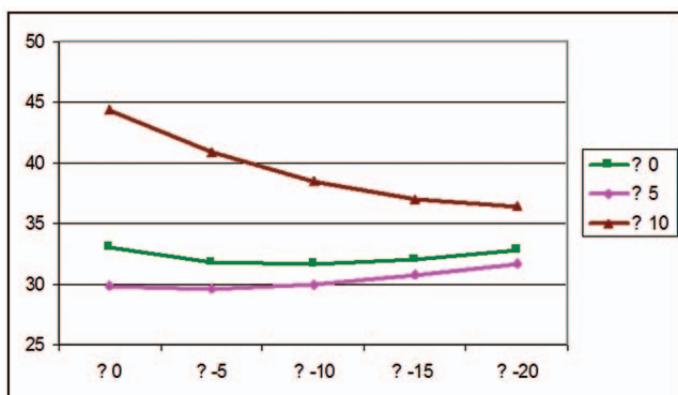
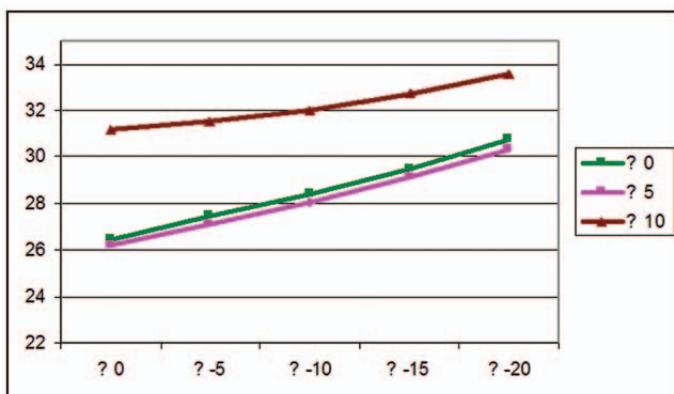


Рис. 7. Показатели PIR распознавания (%) для слитной речи на КВ «Википедия»



При распознавании допускалась свободная грамматика следования фонемных образов, как для фонем, так и для слогов. Только для открытых слогов было наложено ограничение: слоги, которые не имеют гласной, могут находиться лишь перед паузой.

Процедура распознавания проводилась с помощью декодеров *HTK* и *Julius* [9] на трёх КВ: «частотной», «случайной» и «Википедия». В качестве словарного элемента брали: фонемы (59), открытые слоги (7 270), и слоги, поделенные по правилам деления на слоги (10 200).

Ответы распознавания сводились к фонемному виду с целью дальнейшей оценки надёжности сравнительно с эталонным фонемным текстом. В таблице 3 приводится фонемная ошибка для описанных выше КВ. Заметим, что в алфавите фонем различаются ударные и безударные фонемы. Нечёткое произношение ударной гласной и безударной может привести к искажению содержания. Впрочем, на письме ударение обычно опускается. Исходя из этих соображений, в результатах распознавания также подаётся погрешность без учёта ударения, что дало значительно меньшую оценку *PER*.

Таблица 3

Показатели фонемной ошибки распознавания (%) для слитой речи на основе разных речевых образов инструментариев *HTK* и *Julius*

КВ	Фонема		Открытый слог		Слог по правилам деления слогов	
	<i>HTK</i>	<i>Julius</i>	<i>HTK</i>	<i>Julius</i>	<i>HTK</i>	<i>Julius</i>
«Случайная»	28,86	29,11	24,92	24,46	24,54	24,03
«Случайная» (без ударения)	21,39	22,28	17,68	17,47	17,29	17,01
«Частотная»	36,6	-	37,75	-	-	-



Окончание таблицы 3

«Частотная» (без ударения)	26,1	–	27,95	–	–	–
«Википедия»	31,93	35,48	28,01	30,17	28,18	31,08
«Википедия» (без ударения)	24,72	23,19	28,81	20,81	21,00	22,37

И хотя для ряда экспериментов про окончательный результат говорить ещё рано, очевидным является факт зависимости ошибки от метода формирования КВ. Так, тексты из выборки «Википедия» не входили в начальных корпус, а значит, эта выборка содержит определённое количество фонем-трифонов, отсутствующих в ОВ. «Случайная» КВ отвечает общей статистической картине, поэтому следует ориентироваться на показатели надёжности именно этой выборки. Также отметим, что длина предложения для «частотной» КВ составляет в среднем 3,2 слова, тогда как в «случайной» КВ среднее количество слов в предложении — 10,5, почти как в ОВ. Более детальные исследования ОВ изолированных слов для обучения могут объяснить эти результаты.

Выводы

По сравнению с предыдущими исследованиями [1], PER распознавания для слитой речи в отдельных случаях уменьшилась более чем наполовину. Это обусловлено усовершенствованием ОВ для оценки параметров акустических моделей и учитыванием индивидуальных особенностей произношения диктора. Однако пока из результатов чётко не прослеживается лучший вид разделения на слоги.

Предложенный способ формирования ОВ даёт возможность широко охватить фонетическое разнообразие языка, используя около 2% предложений из всех рассмотренных.

В приведённых экспериментах допускалась свободная грамматика следования частей слов. Для проводимых исследований выбраны коэффициенты α и β , компенсирующие несоответствие шкалы акустической и лингвистической составляющих модели распознавания.

Планируется применить статистические лингвистические модели для суб-словных элементов, что должно привести к уменьшению ошибки распознавания. Остаётся неисследованным влияние ряда параметров декодера на надёжность и скорость. В частности, будут разрабатываться подходы к уменьшению алфавита слогов, что должно ускорить распознавание.

Дальнейшие исследования покажут, насколько достигнутого уровня надёжности достаточно для перехода от последовательности фонем (с сопровождающей оценкой акустических параметров) до последовательности слов.

Литература

1. Vasylieva N., Sazhok M. Modeluvannya bahatorivnevoho poskladovoho roz-piznavannya movlennevoho syhnalu. Shl. Donets'k, 2008. № 3. P. 801–808.
2. Sazhok M. Generative Model for Decoding a Phoneme Recognizer Output, Proc. of the 8th International Conference «Text, Speech and Dialogue», TSD'2005, Karlovy Vary, 2005. P. 288–293.
3. Shyrokov V., Monako V. Organizatsiya resursiv natsionalnoyi slovnykovoyi bazy. Movoznavstvo. № 5. 2001.

4. Robeiko V., Sazhok M. Bahatorivneva bahatoznachna model peretvorennya orforhafichnogo tekstu na fonemnyy. Shl. Donets'k, 2011, № 4. P. 117–126.
5. Goncharov E., Kochetov Yu. Povedenie veroyatnostnykh zhadnykh algoritmov dlya mnogo-stadiynykh zadach razmeshcheniya. Diskretnyy analiz i issledovaniye operatsiy. Seriya 2, 6(1), 1999. P. 12–32.
6. Vintsiuk T., Sazhok M. Speaker Voice Passport for a Spoken Dialogue System. Proceedings of the 3rd International Workshop «Speech and Computer» — SPECOM'98, St.-Petersburg, 1998.
7. Sazhok M. Speech Modelling Virtual Laboratory. Speech Processing, Recognition and Artificial Neural Networks. Proc. of the 3rd International School on Neural Nets «Eduardo R. Caianiello», Vietri sul Mare (SA), Italy, 1998. P. 229–232.
8. <http://uk.wikipedia.org>
9. Young S.J. et al.. HTK Book, version 3.1, Cambridge University, 2002.
10. Lee, T. Kawahara and Shikano K.: Julius — an open source real-time large vocabulary recognition engine. In Proc. European Conference on Speech Communication and Technology (EUROSPEECH), 2001. P. 1691–1694.
11. Gales M., Young S. The Application of Hidden Markov Models in Speech Recognition. Foundations and Trends in Signal Processing Vol. 1, No. 3 (2007). P. 195–304.

Сведения об авторе

Васильева Нина Борисовна —

научный сотрудник отдела распознавания и синтеза звуковых образов Международного научно-учебного центра информационных технологий и систем, г. Киев, Украина.

E-mail: n.vassilieva@gmail.com; ninel@uasoiro.org.ua



Sentiment Analysis for Hotel Reviews

Kasper Walter, Mihaela Vela,
DFKI GmbH Stuhlsatzehausweg 3

User reviews and comments on hotels on the web are an important information source in travel planning. Therefore, knowing about these comments is important for quality control to the hotel management, too. We present a system that collects such comments from the web and creates classified and structured overviews of such comments and facilitates access to that information.

I. INTRODUCTION

Travel planning and booking on the web has become one of its most important commercial uses. With the rise of the Web-2.0 user-generated reviews, comments and reports about their travel experiences play an increasing role as information source. Especially for hotel booking, such user reviews are relevant since they are more actual and detailed than reviews found in traditional printed hotel guides etc., they are not biased by marketing considerations as e.g. the hotels' home pages or catalog descriptions, and they reflect actual experiences of guests.

Though nearly every internet travel agency and hotel booking service nowadays offers also ratings and/or reviews of hotels, it is not that easy for hoteliers who want to know what is published about their hotels on the web to gather the user-generated information. A standard search engine like Google will give thousands of hits for a hotel. But, though there seems to be a huge number of sites providing user reviews, often these are just the same because many sites use the same source, such as openholidayguide.com. In other cases, the links lead only to some general page from which one can access reviews besides other information and lacking transparent navigation structure. Also, the links might point to some individual review but leaving it open whether there are other reviews on the site. An additional problem is that the Web-2.0 provides a large number of publication types: besides travel agencies and hotel booking services there are numerous blogs, fora, newsgroups, social networks etc. related to traveling.

Another problem concerns the kind of information: travel agencies and hotel booking services often only publish scalar ratings, e.g. scores between 1 and 5. Such scores are not very helpful for hotel managers as the numeric value does not provide information of what guests actually considered positive or objectionable. Also, the numeric scores are not comparable: when a 3-star hotel receives a higher score than a 4-star hotel, that does not imply that the one is better than the other. For hotel managers the textual user comments would be much more significant than the numeric scores since they would be interested to know *what* the users exactly commented on and *how* they thought of it.

Another problem for hotel managers is that of following updates and new reviews. Hotel booking services and travel agencies collect and publish user reviews systematically, e.g. by asking their customers for comments or ratings. So, new reviews appear quite frequently on their pages but it would be difficult to follow these by just using general search.

For the traveling user who is accessing reviews on the web for planning his travel, many of these considerations are not relevant, as he will be content with a momentary snapshot of reviews. But for hoteliers interested in user comments on the web a service that automatically and systematically collects and summarizes the relevant information from the web would be advantageous and perhaps even more useful than the paper forms many hotels use for gathering feedback from their guests.

The BESAHOT service presented in this paper aims at providing such a service for hotel managers that collects user reviews for hotels from various sites on the web, analyzes and classifies the textual content of the review and presents the result in a concise manner.

We will give an overview of the system in Section II and discuss the major components in more detail, the data acquisition from the web (Section II-B), the statistical polarity classification (Section II-C) and the linguistic information extraction (IE) components (Section II-D). The user interface will be presented in Section III. In Section IV evaluation results for the analysis system will be presented. In Section V we will relate our work to other work in opinion mining.

II. OVERVIEW OF THE SYSTEM

The target users of the system presented here are hoteliers who want to get actual overviews and summaries of textual comments about their hotel(s) on the web. At present, only German reviews from German sites are handled.

The BESAHOT system is an interactive web application based on the GWT framework. The core system on the server-side handles *data acquisition*, *analysis* and *storage* as shown in Fig. 1. The user interface provides various types of summaries of the analyzed data, allows direct access to the information sources on the web as well as free text search.

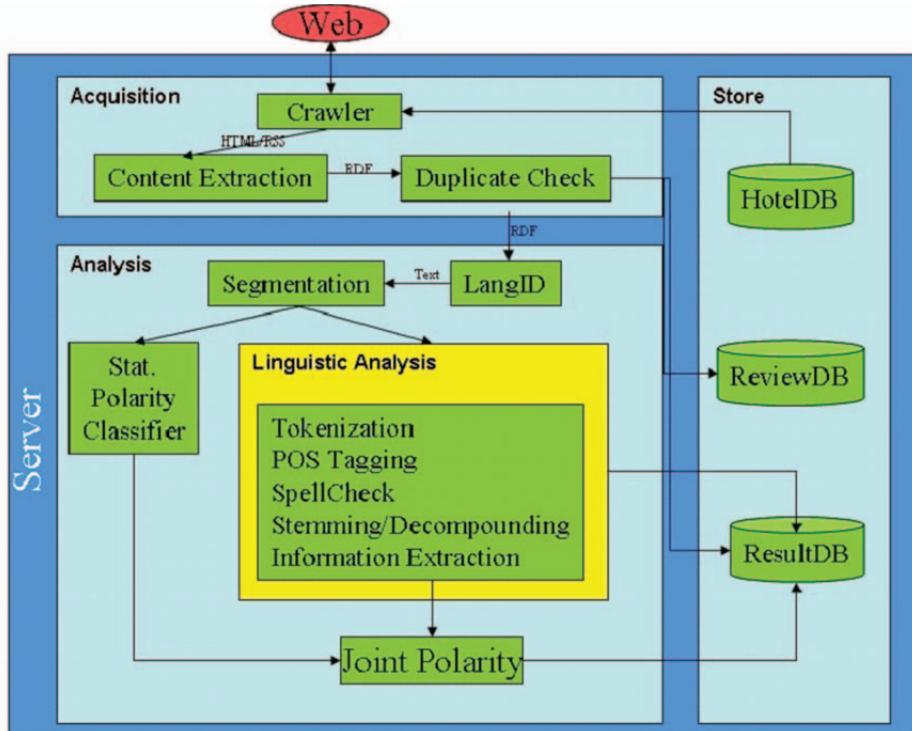


Fig. 1. BESAHOT system overview

New data retrieved from the web by the acquisition system (cf. Section II-B) are passed to the analysis system. The analysis systems first does a *language check* (*LangID*) to filter out re-



views in other languages than German because even German hotel review sites occasionally contain reviews in other languages. The review texts then get segmented into segments ("sentences")¹. These segments are then subjected to further analysis by the *statistical polarity classifier* (cf. Section II-C) and *linguistic information extraction* components (cf. Section II-D) for finer grained analysis of the polarity and the topic of the review. Polarity values are always assigned to text segments, not to reviews as a whole.

The polarity values from the statistical and the linguistic classification are then combined into a joint global polarity value that is used for presenting the segments in the user interface.

Finally, the analysis results are stored together with the review segments in a special *ResultDB* optimized to the retrieval and interaction requirements of the user interface.

II-A. Polarities

In general, we distinguish three possible polarity values for text segments: the segment can express a *positive* opinion, a *negative* one or a *neutral* one. By neutral segments, we mean purely descriptive ones that do not carry an evaluation, such as *We spent three days at the hotel*. The delimitation of neutral/descriptive and evaluative text is not always easy, not even to humans. A remark like *no minibar* on the one hand just describes a fact but on the other hand the remark is probably meant as a negative comment describing a deficiency.

Another problem for a polarity classification on text segments is that a segment might address more than one topic. For example, *clean rooms and friendly personnel* addresses the two topics *room* and *personnel* both rated as positive. But for a comment like *Room ok, but poor breakfast* it would be unclear what the overall polarity value of the comment should be, as there are actually two ratings on two different topics. Similar issues arise with respect to multiple ratings on the same topic as in *clean, but tiny room*.

The BESAHOT IE system is able to detect such multiple topics and ratings on a text segment. Nevertheless, as we have not yet found a good solution for handling these cases in the user interface, at present we prefer to disregard them in favor of a global polarity assignment, even if that sometimes might be a bit random. This will be further discussed in Section II-E and Section IV.

II-B. The Acquisition System

The acquisition of reviews from the web is handled by a web crawler. The *HotelDB* defines for each hotel a set of crawl configurations that define a start URL, URL patterns for links to follow, target URL patterns for pages containing reviews, the potential crawl depth and an indicator whether the content of a target page is mutable or not. The crawler handles HTML pages as well as RSS feeds. All the URLs usually point to dynamic web pages, that is, the content of the web pages can change between visits. Also, the web pages most times contain hundreds of links, most of them being irrelevant for retrieving reviews (e.g. advertisements, other hotels, etc). Therefore, filter patterns are used to restrict the crawler to follow only relevant links. The distinction between links to follow and target pages is required because the crawler often has to go through several intermediate pages to get at the review pages, e.g. from the hotel overview page to the review overview page to individual review pages and to more reviews.

¹ We prefer the term "segments" to "sentence" because the segments are not always sentences in a linguistic sense but just phrases.

At present, we ignore sites that present only numeric scores for hotel ratings and no textual reviews.

Also, when we found that sites use the same data source for the reviews, we chose one of the sites as a representative and do not use the alternative sites for data retrieval.

When a target page is retrieved a *content extraction module* is applied that extracts the relevant textual content of the review but also other metadata such as scores and information about the reviewer/guests. The content extraction is based on XSLT scripts for known sites (*screen scraping*). If a page contains several reviews, for each of them a separate review instance is created. Extracted content is represented as RDF instance of a *Review* ontology defined in OWL (*Ontology Web Language*, <http://www.w3.org/2004/OWL/>). Fig. 2 shows an example of the structure.

```
<bes:Review>
  <bes:about rdf:resource="urn:hotel:687_02">
  <bes:fullText>
    Parkmöglichkeiten eingeschränkt ...
  </bes:fullText>
  <bes:reviewer>
    <bes:Guest>
      <bes:travelTime>Juli 2010</bes:travelTime>
      <bes:age>45-50</bes:age>
      <bes:guestType>
        geschäftlich allein reisend
      </bes:guestType>
    </bes:Guest>
  </bes:reviewer>
  <bes:source rdf:about="http://www.hotel..."/>
  <bes:rating>
    <bes:Rating>
      <bes:ratingCategory>
        Gesamtbewertung
      </bes:ratingCategory>
      <bes:ratingScore>
        8,1 von 10
      </bes:ratingScore>
    </bes:Rating>
  </bes:rating>
</bes:Review>
```

Fig. 2. Extracted content as RDF

Since the content of the web page is dynamic the system needs to determine whether it has seen a review before or whether it is a new review. The *duplicate check* uses review fingerprints created from the textual content without any formatting. This provides reliable and efficient tests independent of text size and formatting. Reviews that survive the duplicate check are stored in the *ReviewDB* and passed to the analysis system. The review texts there first are split into text segments that become the units of further analysis.

II-C. Statistical Polarity Classification

The statistical polarity classifier assigns to each text segment a polarity value. As a basis for statistical polarity classification we used the classification engine of [1]. This engine is based on *character n-grams* instead of terms. Classification is achieved by computing for each class the probability of the text segment as that of the “best” matching n-gram sequence based on Bayes chaining rule given in (1), where N is the length of the segment.

$$P(c_1, \dots, c_N) = \prod_{i=1}^N P(c_i | c_{i-n+1}, \dots, c_{i-1}) \quad (1)$$



For our application this approach has several advantages.

- Robustness against orthographic errors that are quite frequent in the reviews, especially transposed or omitted letters.
- Robustness against unknown terms from word compounding that are very frequent in German: the terms get automatically split up into the smaller n-gram sequences.
- It diminishes the sparse data problem as no huge training corpus is required
- Applicability to short texts such as segments, not just longer documents

For getting training data for the statistical classifier we exploited the fact that on some hotel sites users themselves classify their contributions into positive and negative text items. An example is shown in Fig. 3.

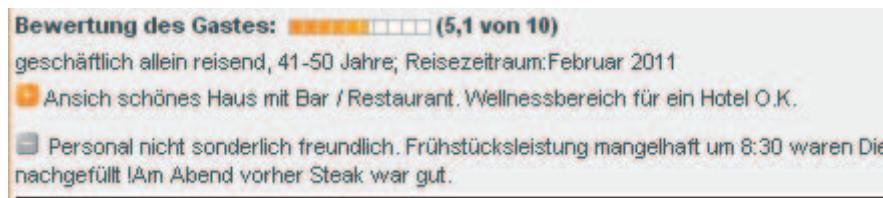


Fig. 3. Polarity classification by users

So we collected a corpus of such hotel review texts from these sites and used them for training the classifier with 2 polarity classes (positive/negative). We use 4-grams with Goodman smoothing ([2]), trained on roughly 7200 text segments for each class. Crossvalidation benchmarks demonstrated a satisfactory performance as shown in Table I.

Table I

Table CLASSIFIER BENCHMARK: 10-FOLD CROSSVALIDATION

Training	Precision	Recall	F-Measure
50%	0.90002	0.90017	0.90008
90%	0.92846	0.92855	0.92851

The benchmarks illustrate the robustness of the classifier: performance of the classifier does not increase very much when more data are used for training.

We use only two polarity values for the statistical classifier. An experiment to add a neutral category from manually classified data showed a clear performance degradation. Therefore, we preferred to leave the detection of neutral segments to the IE.

In Section IV we will further discuss the performance with respect to manually annotated data and the problem of multi-topic and neutral text segments.

II-D. Information Extraction

The main task of the linguistic analysis components in the BESAHOT system is to identify from a text segment its *topics* (what is talked about) and how these get rated within the segment. The core of that analysis is an information extraction (IE) component based on the SProUT platform (Shallow Processing with Unification and Typed Feature Structures; [3]). SProUT is a rule based IE system combining finite state technology with unification on typed feature structures for imposing type constraints on possible feature values and propagating constraints by coreferences. Fig. 4 shows an example for a rule in the SProUT system. The left-hand side of each rule consists of a regular pattern over the input sequence,

while the right-hand side specifies the output structure. The `@seek` operator allows to call other rules and use their output.

```

premod_nn :>
    (@seek(quantifiers_rule) &
     quantifier &
     [ NEGPOL #neg ])??
    (@seek(conj_adj_phrase) &
     %mods & property &
     [ NEGPOL #neg ])*
    (@seek(noun_match) &
     gazetteer &
     [ SUPERCLASS #class,
      SURFACE #surf,
      POLAR #pol ]
    ->
    object & [OBJECT #surf,
              CATEGORY #class,
              NEGPOL #neg,
              LEXPOL #pol,
              RATING %<mods>].

```

Fig. 4. A SProUT rule for NPs

The IE system is designed to supply answers to the following questions:

- *Topic* of the review segment: what is evaluated?
- *Dimension* of the evaluation: what properties are evaluated?
- *Dimension value*: what is the value on that dimension?
- *Polarity* of the evaluation: is it positive or negative or none at all (neutral)?

For the IE component we created a dictionary of *domain-specific terms* relevant for the hotel domain as well as a *sentiment dictionary* that associates basic polarity values with terms. Besides that, the dictionaries assign topic terms to a semantic category indicating what aspect of hotels this topic refers to, e.g. *Service*. Also, the *dimension* of evaluative terms are defined by the dictionary. Fig. 5 gives an impression of these categories and dimensions.



Fig. 5. Main topics and dimensions in the review ontology



The IE system distinguishes several types of possible roles for a polarity value that influence in different ways what actual polarity is expressed in a segment.

- evaluative speech act indicators, such as *regrettably*. These can override any other polarity expressed.
- negation particles, e.g. *not* that will turn polarities in their scope to the opposite.²
- polarity modifiers, e.g. the *too* in a phrase like *too small* that can override the default polarity at phrasal level.
- “missing things” indicators, such as *without*
- negative and positive polarity items as well as idiomatic polarity expressions
- a default lexical polarity, e.g. that *nice* expresses a positive rating

Fig. 6 gives an impression of the IE markup applied to stemmed text input, each line representing a text segment. Each colored sequence represents one or more semantic annotations on the text.

The screenshot shows the 'Test grammar' application window. On the left, the 'Input text' pane contains a multi-line text of stemmed German words. The right pane, 'Output text', displays the same text with various colored sequences (blue, green, red) highlighting semantic annotations. A vertical scroll bar is visible between the two panes. On the far right, the 'Active components' panel lists three checked components: 'ExtendedGazetteer', 'Morphology', and 'Tokenizer'. At the bottom, there are buttons for 'Search', 'Matches', 'Statistics', and 'Close', along with a checkbox for generating XML output.

Fig. 6. Information extraction markup

Fig. 7 depicts the semantic representation of *kein kostenloses schnelles WLAN* (*no free fast WLAN*) from IE as a feature structure. It can be read as follows: *WLAN* is the topic belonging to the *telecommunication* category. There are two properties attached that by default denote positive properties (*free, fast*), shown as values of the LEXPOL feature. But these occurrences are in the scope of a negation polarity, the NEGPOL value that is propagated down to the rating elements by a coreference and that will invert these default values.³ This is handled by an IE postprocessor. So in the end we will have two negative ratings for the *WLAN* topic as being neither free nor fast.

² Of course, this is a simplified assumption: *not bad* does not mean the same as *good*, but in this context we ignore such subtle distinctions.

³ The value *polarity* on any of the *POL features that correspond to the different roles of the polarity values just designates a neutral value, that is, neither positive nor negative.

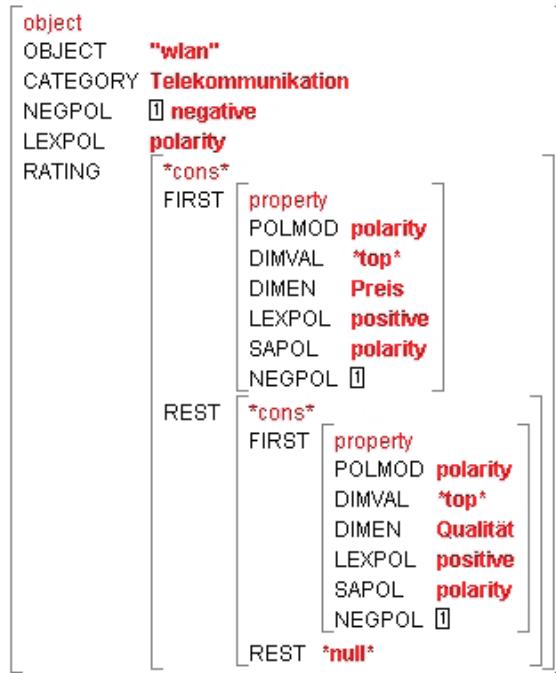


Fig. 7. Semantic representation for *kein kostenloses schnelles WLAN*

Since the review texts often are not well-formed linguistically with respect to orthography and grammar, some preprocessing and normalization steps are applied before actually submitting the text to the SProUT IE system. Part of Speech (POS) tagging is used to reduce the search space for lexically ambiguous words and word forms. Also, to improve input quality a spell-check is applied that automatically can correct frequent types of spelling errors like transposed and left-out characters. To prevent over-correction the similarity measure between word and possible replacement must be set very high.

A frequent problem in processing German is word composition by which several terms are combined into a single word. This compounding generates new words missing from the usual dictionaries and so these are difficult to process. The SProUT morphology includes a decomposition component for German compounds that allows us to handle compounds as multi-word expressions and simplifies building the semantic dictionaries⁴. The morphological stem assignment is also used to correct strange POS tag assignments from the tagger for terms for which the morphology provides a more plausible POS.

After the SProUT IE has marked up the found structures, the resulting feature structures are passed to a postprocessor that evaluates the found structures and computes the final rating values for a segment taking into account the different types of polarities and their scopes. This postprocessor would recognize that the positive lexical default polarity values of the adjectives in Fig. 7 occur in the scope of a negative polarity marker and therefore would invert them such yielding finally 2 negative ratings instead of 2 positive and some negative polarity. Also, isolated annotations that cannot be related to ratings get eliminated here.

It is obvious that for the IE system the representation of multiple topics and multiple ratings in a text segment is not a problem. Also, we treat the absence of rating annotations in a segment as evidence that the segment belongs to the neutral polarity category.

⁴ Usually, the last component of a compound is regarded as the headword as that governs the morphological properties of the compound. Semantically, we found that often the other components are more significant.



II-E. Combining Statistical and IE Polarities

For each segment the statistical polarity classifier yields a positive or negative polarity value. More fine-grained polarity values are available even for parts of the segments (sub-segments). We developed an experimental system that would use the IE to create finer phrases as subsegments of the text segments according to the recognized topic changes. Unfortunately, in many cases that resulted in text fragments that are incomprehensible without their syntactic context and so cannot be presented to users⁵. Therefore we kept to the approach to assign a global polarity value to the whole text segment, but the assignment of that global value would take into account both classification sources, the statistical value and the IE values. In that approach, the statistical value is regarded as baseline value and the ratings from IE are used to possibly correct that value. As an approach that would give the IE ratings preference to the statistical value proved unsatisfactory, we developed a method for using the IE ratings as length-normalized weights on the statistical values: for each polarity, the IE weight is defined as the number of ratings of that polarity divided by the token length of the segments. On short segments, the IE ratings thus will have larger weight than on longer segments. The global polarity values then are computed by combining the scores of the statistical classifier with these weights according to (2),

$$\begin{aligned} \text{sp}(p) \\ \text{pol} = \underset{\substack{\text{p} \in \{\text{pos}, \text{neg}\}}}{\operatorname{argmax}} \dots \dots \dots \\ (2) \end{aligned}$$

where p is a polarity, $\text{sp}(p)$ its statistical score, s/l the segment length and $\text{ie}(p)$ the number of the IE ratings with that polarity. This approach reconciles the confidence of the statistical classifier with the IE results better than a preference based approach. A side effect of the formula is that the statistical polarity value will be kept, if the IE does not yield ratings. The motivation for this is that the statistical classifier has larger coverage than the current IE. Therefore we keep the statistical polarity value and treat the absence of IE ratings as meaning "IE does not know" rather than "This is neutral polarity". This provides more flexibility for the user interface that can decide how to handle this case.

III. THE USER INTERFACE

The BESAHOT system is a tool to support hotel managers in quality control. So it should provide them with fast and comprehensive overviews and summaries of how their hotel is rated on the web and how it is commented on by guests and visitors on the web.

Fig. 8 shows the main result overview that the user will see when accessing the BESAHOT service after selecting a hotel. The top panel displays some statistics about scores from source sites, normalized to a scale between 1 and 10, and about guest types, as far as this information could be extracted from the source web pages. Also, the time range can be restricted to show only recent reviews. The *Aktualisieren* button allows to start the crawler to search for new reviews on the web for the selected hotel⁶.

The main panel provides a summary of the reviews by displaying text snippets from the reviews according to their polarity and category. A click on a segment opens

⁵ A possible solution would be the use of a text generator to generate some simplified text from the semantic structures of the IE instead of using only text pieces from the original review texts. At present, this is outside the scope of the project.

⁶ This *Actualize* button exists only in the demonstration system. In the final system the server would automatically update the databases periodically.

a popup panel that displays the full review text highlighting the displayed segment in context. This allows users to check the text in context and also makes it unnecessary to visit the source page, though this would be easy by just following the provided link to the source page. Additionally, the popup displays information about the guest that provided the rating.

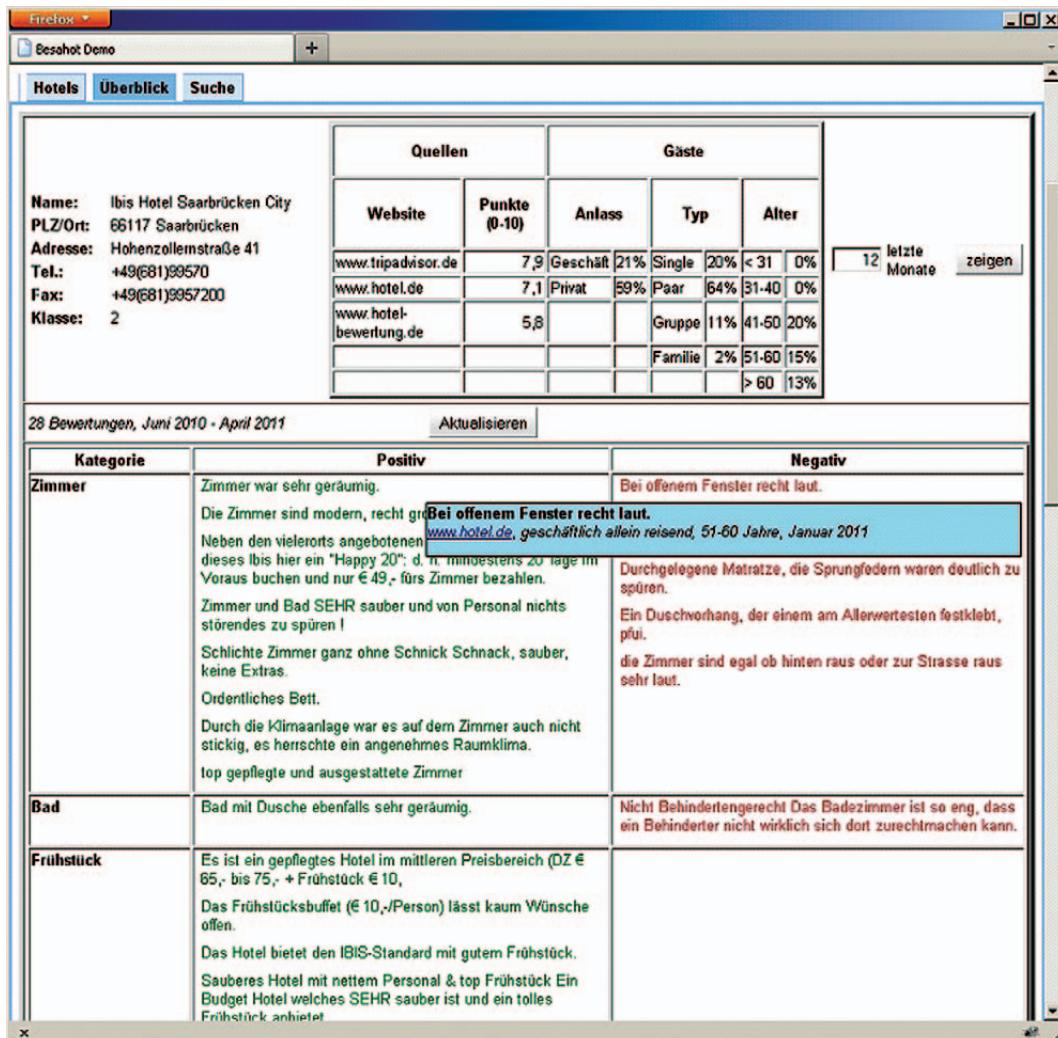


Fig. 8. Classified review summary for a given hotel

For this display we exploit the IE's capability to identify neutral text segments: text segments that do not receive an IE rating here are omitted from the view. An open issue in designing the user interface is the handling of text segments belonging to more than one category. Adding these segments to each category tends to result in rather crowded and redundant category fields, impairing the usefulness. So, presently such ambiguous segments currently are displayed only in one category, preferably a dominant one.

In addition to the overview presentation, a free text search function allows users to search the review database by freely chosen keywords, independent of the predefined categories and polarity values.

A usability test for the user interface with members of the Saarland hotel association is in preparation.



IV. EVALUATION

We evaluated the analysis system on a corpus of 1559 hotel reviews crawled from the web. These reviews contained 4792 text segments. For the evaluation, these segments were manually classified with respect to their polarity, including the neutral polarity besides positive and negative ones. Also, we annotated the segments whether they cover more than one topic. The distribution from this manual classification is shown in Table 1.

Table 1
MANUAL CORPUS CLASSIFICATION

Segments	positive	negative	neutral	multi-topic
4792	2240	1183	938	431

We evaluated the performance of the statistical classifier alone, the IE system alone and the hybrid system combining the polarity classifications from the statistical classifier and the IE system as described in Section II-E. Evaluated on all segments, the results in Table 2 were achieved.

Table 2
CLASSIFICATION ON ALL SEGMENTS

	Correct	False	Accuracy
Stat	3145	705	0,66
IE	2604	486	0,54
Stat+IE	3208	646	0,67

It shows that the IE system currently covers less data than the statistical classification, but that it slightly improves the overall classification accuracy. These data relate the results to the complete corpus not taking into account the presence of neutral and multi-topic segments. As discussed in previous sections, the assignment of only positive/negative polarities in these cases can be a bit random, or, for the cases of neutral polarity that make up about 19% of the corpus, the positive/negative assignment is rather uninteresting.

Therefore, in a second experiment, we evaluated the classification performance on only the subset of manually verified positive/negative segments and achieved considerably better results, shown in Table 3.

Table 3
CLASSIFICATION WITHOUT NEUTRALS

	Correct	False	Total	Accuracy	F-measure
Stat	3145	705	3854	0,82	0,80
IE	2604	486	3090	0,68	0,66
Stat+IE	3208	646	3854	0,83	0,81

These values demonstrate that it would be beneficial to be able to identify neutral and multi-topic/multi-polarity ratings. As mentioned in Section II-C the pure statistical classifier did not look promising in that respect. Therefore we evaluated how well the IE system would recognize the neutral and the multi-topic cases identified in our corpus. The results are shown in Table 4.

Table 4
RECOGNITION OF NEUTRAL AND MULTI-TOPIC POLARITY

	Correct	False	Total	Accuracy
Neutrals	682	256	938	0,72
Multi-topics	324	107	431	0,75

These values look promising. We expect that improving the coverage of the IE system will also improve these figures. That will also provide a strong motivation for changing the interpretation of the absence of a polarity rating from IE as “don’t know” to “classify that as *neutral*”.

V. RELATED WORK

The development of the WWW and the possibility for customers/users to express their opinion online made the online available reviews interesting for both the vendor as well as for the potential customer. Therefore, the interest on opinions and sentiments of (former or future) customers has increased tremendously. In parallel, the development boosted research in opinion mining and sentiment analysis in recent years. Good overviews on existing opinion mining techniques and methods are given by [4] and [5].

Most research in this area concentrates on opinions about products. Also, domains such as movie reviews or news found considerable interest especially in research, since large datasets and corpora are publicly available.

The goal of opinion mining can vary considerably. In many cases, one is only interested in a global overview: how many users/reviews rate a product positive or negative. For these, a global polarity classification is sufficient without having to go into details of a product. More fine-grained is an approach as that of [6] who present an opinion mining approach for news articles. They do not just global classification at document level but split up the review into phrases. Based on a predefined lexicon and contextual information they apply machine learning techniques for determining the polarity of the phrase. But different from our approach, they do not identify specific features that are evaluated.

Research in opinion mining often requires specific resources such as suitably classified corpora and sentiment dictionaries that associate terms with sentiments. For English, a large set of resources is publicly available for research. Therefore also most research is done on English data, such as ([7], [6], [8]). For opinion mining approaches that also do feature extraction for the rated product features, also domain-specific dictionaries can be needed that specify product-specific features.

For German (or other languages), there are less of such resources available, even though the situation starts to improve. A large sentiment dictionary for German has been built by [9] that we used to initialize our sentiment dictionary for the terms extracted from our hotel review corpora. The dictionary of domain-specific terms and concepts for the hotel and tourism domain we had to create ourselves.

While our IE system for feature extraction relies on manually created rules, there are a number of approaches to use machine learning techniques to achieve that, such as the work of [7] on mining opinions about products. They describe an unsupervised information extraction system which determines the relevant features and the corresponding opinion. The method uses relaxation labeling ([10]) for finding the semantic orientation of words in the context of given product features and sentences. A more linguistically inspired approach that resembles ours is described in [11].

The tourism domain is not one of the mainstream domains for opinion mining research. [8] uses a corpus of English reviews from *tripadvisor.com* in order to present a rule-based method for classifying opinions. Different from other approaches she takes also the context into account. This way she differentiates between the needs of a person on a business trip and the needs of the same person on a family trip. A larger English corpus also from *tripadvisor.com* is used in the study of [12] that uses linguistic preprocessing with the SENTIWordnet ([13]) but machine learning techniques for feature assignment. [14] describe in their work a framework for constructing Thai language resource for feature-based opinion mining for hotel reviews. Their approach for extracting features and polarity words from opinionated texts is based on syntactic pattern analysis. In general it is left unclear how the high number of misspelled and ungrammatical data, we found in our corpora, are handled in these approaches and how they affect the result.



In general, these approaches focus on research on specific technologies but there is little indication about what the results are used for in an application, who the users of the results are and how results can be used by them. In many cases the research is related to building recommendation systems so that the results are not directly used by humans but just by machines. The BESAHOT system, on the hand, targets explicitly human users, not machines.

Closely related to BESAHOT is the work on review summarization such as [15, 16]. Summarization there means extracting relevant sentences classified according to their polarity and some category, called *features* or *aspects* in these papers. They focus on adjectives as carriers of polarity and nouns/noun groups as designators for features, ignoring other word classes. Negation seems to be recognized only if adjacent to an opinion term. Irrelevance/neutral is defined by thresholds on scores. The methods of feature extraction based on nouns in the context of opinion terms tend to yield high numbers of features. [16] therefore introduce a second level of manually created static high-level aspects that resemble more the high-level categories used in BESAHOT. It is unclear whether sentences belonging to more than one category are treated in the user interface in a special way. The BESAHOT-IE approach looks more flexible as it is not restricted to few word classes and it can handle larger contexts and relevant linguistic phenomena better than these approaches. Also, resources for the IE are easy to extend and to adapt for new data and phenomena.

VI. CONCLUSION

We presented a web based opinion mining system for hotel reviews and user comments that supports the hotel management in monitoring what is published on the web about their houses. The system is capable of detecting and retrieving reviews on the web, to classify and analyze them, as well as to generate comprehensive overviews of these comments. We showed that, despite some remaining issues, the system provides good performance for the analysis and the classification tasks. Further research will be necessary especially with respect to the demarcation of evaluative and neutral text as well as to the handling of multi-topic segments, especially for the user interface.

Besides that extension of coverage to more sites is under work. One further direction is to include web search into the data acquisition to find reviews on sites that only infrequently or just by chance publish guest comments on hotels registered on the BESAHOT service. Also, we are preparing a pilot test of the BESAHOT service with members of the Saarland hotel association to improve the information value and usability of the system.

Acknowledgments

The reported work was supported by a grant from the Saarland Ministry Of Economics and by the European Commission under contract number FP7-231527 (IKS).

REFERENCES

1. J. Steffen, "N-Gram Language Modeling for Robust Multi-Lingual Document Classification," in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*. Lisboa: ELRA, 2004. P. 731–734.
2. S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," Computer Science Group, Harvard University, Cambridge (Mass.), Tech. Rep. TR-10-98, 1998. [Online]. Available: <http://research.microsoft.com/~joshuago/tr-10-98.pdf>

3. W. Drozdzynski, H.-U. Krieger, J. Piskorski, U. Schafer, and F. Xu, "Shallow processing with unification and typed feature structures — foundations and applications," *Kunstliche Intelligenz*, vol. 1. 2004. P. 17–23.
4. B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2. 2008. № 1–2. P. 1–135.
5. B. Liu, "Opinion mining and sentiment analysis," *Handbook of Natural Language Processing*, 2010.
6. T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis," *Computational Linguistics*, 2005.
7. A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, 2005.
8. S. Aciar, "Mining context information from consumer's reviews," in *Proceedings of the Context-Aware Recommender Systems (CARS) Workshop*, 2009.
9. U. Waltinger, "GermanPolarityClues: A lexical resource for german sentiment analysis," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, 2010.
10. R. A. Hummel and S. W. Zucker, "On the foundations of relaxation labelling processes," *PAMI*, 1983.
11. L.-W. Ku, T.-H. Huang, and H.-H. Chen, "Using morphological and syntactic structures for chinese opinion analysis," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, 2009.
12. S. Baccianella, A. Esuli, and F. Sebastiani, "Multi-facet rating of product reviews," in *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, ser. ECIR '09. Berlin, Heidelberg: Springer-Verlag, 2009. P. 461–472.
13. A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC-06)*, 2006.
14. C. Haruechaiyasak, A. Kongthon, P. Palingoon, and C. Sangkeettrakarn, "Constructing thai opinion mining resource: A case study on hotel reviews," in *Proceedings of the Eighth Workshop on Asian Language Resources*, 2010.
15. M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004.
16. S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. A. Reis, and J. Reynar, "Building a sentiment summarizer for local service reviews," in *NLPIX2008*, Beijing, 2008.

Сведения об авторе

Walter Kasper —

Walter Kasper is researcher in the areas of semantics and discourse. He studied German linguistics, literature and philosophy at the University of Dusseldorf. There he received a PhD in German linguistics. He worked in numerous research projects focussing on Natural Language Processing, Knowledge Management and Discourse.

1980–1984 he was teacher for German as a Foreign Language in Dusseldorf. 1985–1992 he worked as research scientist and project leader at the Institut fur Maschinelle Sprachverarbeitung (IMS) at the University of Stuttgart. Since 1992 he is senior researcher at the at the LT-Lab of DFKI for numerous collaborative projects, national projects, EC funded projects as well as industrial projects.

The current major occupation is the MESH project.

<http://www.dfg.de/~kasper/>

DFKI GmbH. Language Technology Lab.



Памяти Тараса Климовича Винценюка



29 мая 2012 г. после продолжительной болезни ушёл из жизни выдающийся учёный, основатель перспективного направления в теории распознавания образов, доктор технических наук, профессор Тарас Климович Винценюк.

Тарас Климович Винценюк родился 10 марта 1939 г. в с. Кульчин Волынской области в семье репрессированных. В 1956 г. окончил школу с золотой медалью. С отличием окончил Киевский политехнический институт. Свой научный путь прошел от инженера (1962) до заведующего отделом распознавания и синтеза звуковых образов (1988) в Институте кибернетики им. В.М. Глушкова НАН Украины и Международном научно-учебном центре информационных технологий и систем (1997).

Широкое признание получила генеративная модель распознавания образов, впервые сформулированная Т.К. Винценюком еще в 1967 г. Этот подход, известный в мире как Dynamic Time Warping (DTW), нашел свое применение не только в теории распознавания речевых и зрительных образов, но и в обработке текстов, естественных сигналов и в сфере биологии. Подобная модель, известная как Hidden Markov Models (HMM), берёт начало в 1973 г. и является наиболее цитируемой в мире. Обе модели — самые продуктивные в системах распознавания речи.

С конца 60-х гг. прошлого столетия под руководством Т.К. Винценюка разработан ряд систем распознавания речи. Был пройден долгий путь от систем на основе БЭСМ до портативных устройств с голосовым управлением.

Научные разработки Т.К. Винценюка отражены более чем в 300 работах и двух монографиях, отмечены высшими наградами ВДНХ СССР, дипломами многих выставок. В составе авторских коллективов он удостоен Государственных премий Украины 1988 и 1997 гг. в области науки и техники.

Т.К. Винценюк — создатель концепции образного компьютера, которая легла в основу Государственной научно-технической программы (2000–2010).

Тарас Климович Винценюк являлся членом ряда научных сообществ, входил в программные комитеты и редакционные коллегии многих научных конференций и издательств. Он основал и неизменно возглавлял Украинскую ассоциацию обработки информации и распознавания образов (УАсОИРО). Начиная с 1992 г., под его руководством прошло 10 международных конференций

по обработке сигналов и изображений и распознаванию образов «УкрОбраз», изданы сборники трудов конференции, а с 2004 г. проводились ежегодные летние школы, посвящённые речевым информационным технологиям.

Тарас Винцюк открыл новую эпоху распознавания речевых образов, в которой нам предстоит жить с памятью о великом учёном, энтузиасте и учителе — основателе всемирно известной научной школы.

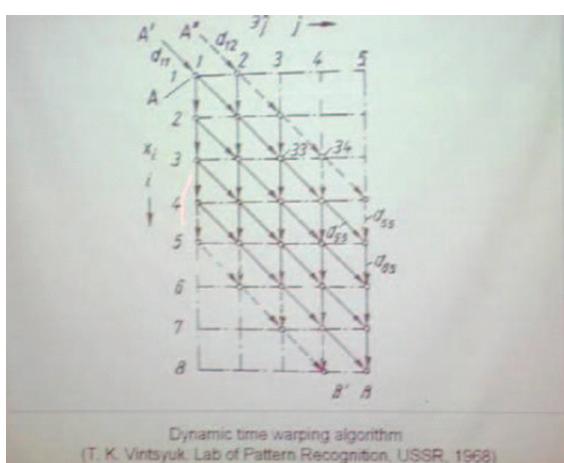


Система распознавания сплитной речи на БЭСМ-6 (1970)



Многоязычная система речевого диалога
Речь-121 (1986)

Портативные устройства (2001–2006): цифровой диктофон с голосовым управлением «Вокофон» и устный словарь-переводчик



Dynamic time warping algorithm
(T. K. Vintsyuk, Lab of Pattern Recognition, USSR, 1968)

Обзорный слайд на одной из международных конференций

Редакция:

Редактор — Елена Долматова
Выпускающий редактор — Анастасия Чипенко
Корректор — Татьяна Денисьева
Дизайн — Анна Ладанюк
Вёрстка — Александр Перевозов

Адрес редакции: 109341, Москва, ул. Люблинская, д. 157, корп. 2
Тел.: 8 (495) 979-54-27

Подписано в печать 5.10.2012. Формат 60x90/8. Бумага офсетная. Печать офсетная
Печ. л. 7,0. Тираж 1000 экз. Заказ № . Издательский дом «Народное образование»
Отпечатано в ООО «Чебоксарская типография № 1». 428019, г. Чебоксары, пр. И. Яковleva, 15

© «Народное образование»