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 $\mu = [\mu_1, \mu_2, ..., \mu_n]^T$ и ковариационная матрица Σ, размерностью $n \times n$, где n — размерность вектора первичных признаков речевого сигнала.

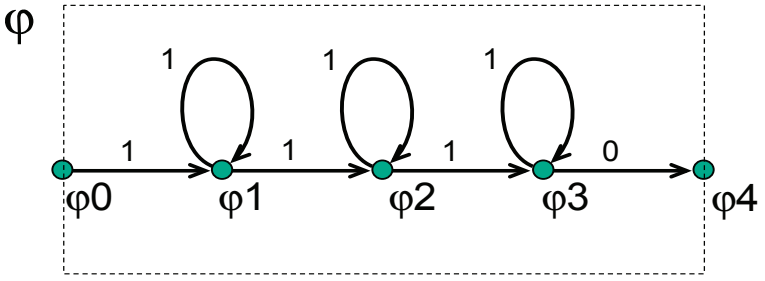

Рис. 1. Генеративная модель фонемы φ с тремя фазами-состояниями φ1, φ2, φ3

Начальное состояние φ0 и конечное φ4 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,

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

В свою очередь, матрица W представляется в виде:

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

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:

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

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 $t, 1 \le t \le T$,

$L_{m_r}(t)$ — 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:

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

После ряда преобразований получаем формулу в виде:

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

где

w_{ii} — 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:

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

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

Количество слов на адаптицию / Дикторы		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.

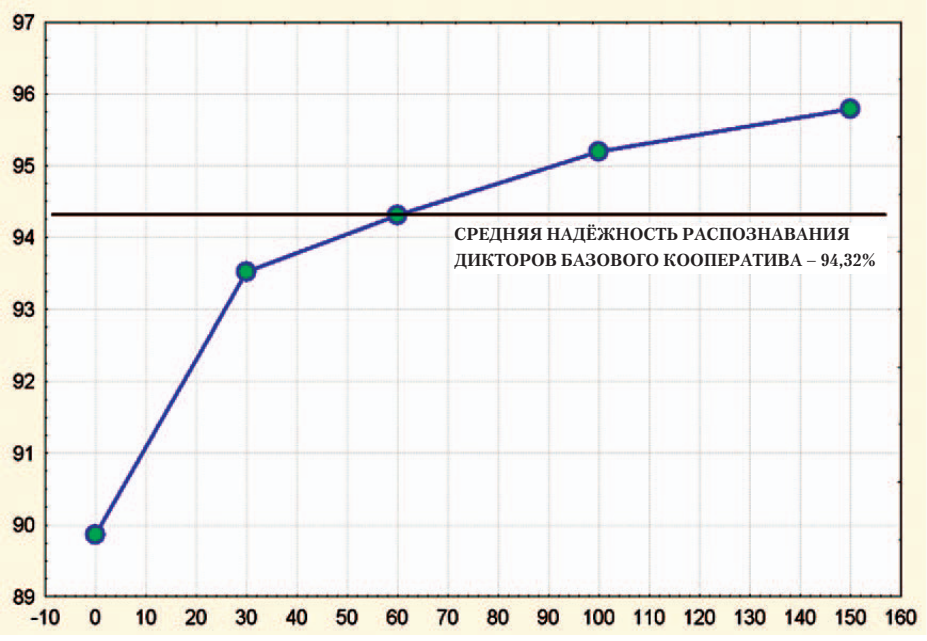

Рис. 2. Средняя надёжность распознавания дикторов контрольной группы до и после адаптации

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 нагова-ривали один и тот же набор слов (241 слово). Из таблицы видно, что по-сле адаптации на голос нового диктора надёжность распознавания в сред-нем 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

Количество слов на адаптицию / Дикторы		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

Количество слов на адаптицию / Дикторы		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

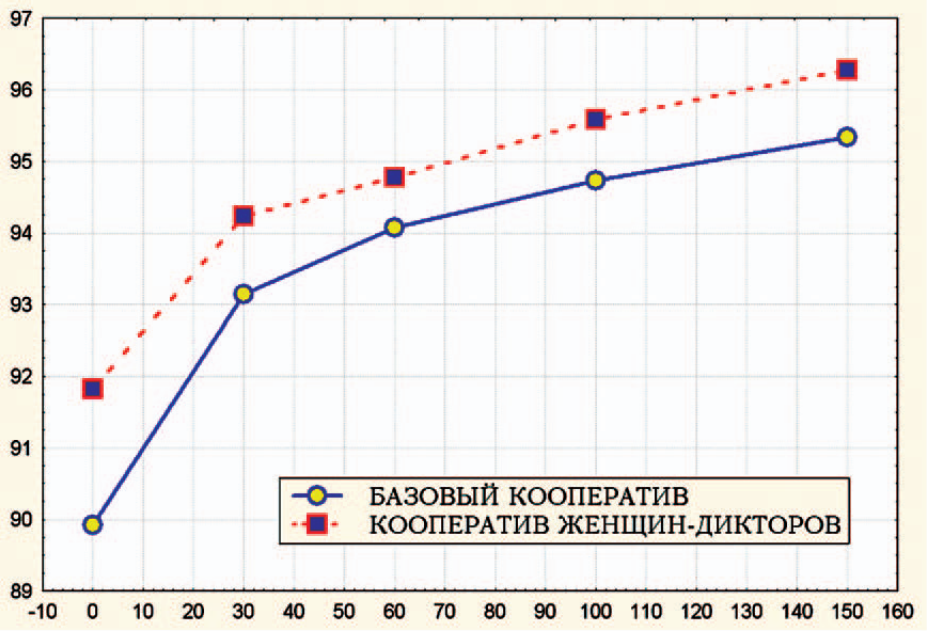

Рис. 3. Средняя надёжность распознавания дикторов женского пола

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.

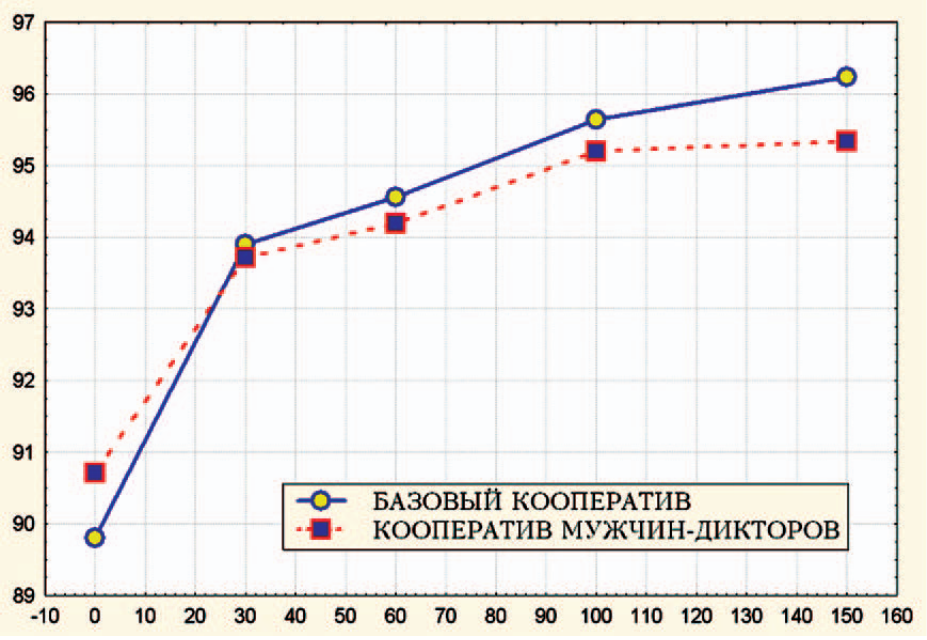

Рис. 4. Средняя надёжность распознавания дикторов мужского пола

Выводы

Результаты гендерозависимого распознавания показывают уменьшение количества ошибочно распознаных 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 наговорили свыше 12 000 слов в общей обучающей выборке. В такой ситуации преимущество адаптации очевидно.

Дальнейшая адаптация к голосу диктора на базе гендернозависимых 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.

Литература

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. *Vasylieva N*. Text Selection for Training Procedures under Phoneme Units Variety / N. Vasylieva, M. Sazhok //Proceedings of the 10th International Conference on Speech and Computer — SpeCom'2005. Patras, 2005. P. 69–76.

3. *Винцюк Т.К*. Анализ, распознавание и смысловая интерпретация речевых сигналов. Киев: 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