УДК 004.934

А.А. Юхименко

Адаптация к голосу диктора для пофонемного распознавания изолированных слов и спонтанной слитной речи украинского языка

Рассмотрены вопросы адаптации к голосу нового диктора к предварительно созданным системам пофонемного распознавания. Описан метод максимальной правдоподобности линейной регрессии. Приведены результаты экспериментальных исследований по адаптации для распознавания изолированных слов и спонтанной слитной речи. Проанализирована эффективность разных подходов в адаптации.

This paper is devoted to the problems of adaptation to a new speaker voice for speech recognition systems. The method of Maximum Likelihood Linear Regression (MLLR) is described. The results of different adaptation experiments with isolated words and continuous speech are discussed. Particularly the effectiveness of different approaches to the adaptation is analyzed.

Розглянуто адаптацію голосу нового диктора до створених попередньо систем пофонемного розпізнавання. Описано метод максимальної правдоподібності лінійної регресії. Подано результати експериментальних досліджень з адаптації для розпізнавання ізольованих слів та спонтанного злитого мовлення. Обговорюється ефективність різних підходів в адаптації.

Введение. Пофонемное распознавание речевого сигнала предусматривает формирование речевого паспорта диктора, включающего в себя акустические модели фонем (вероятностные параметры моделей) [1]. Оценка этих параметров моделей фонем проводится с использованием данных обучающей выборки (ОВ) определенного диктора, которая должна содержать все фонемное разнообразие речи. Опыт формирования таких выборок показывает, что их объем должен быть достаточно большим, так что диктору необходимо потратить не один час для записи речевых сигналов с целью создать систему распознавания с приемлемой надежностью при пофонемном распознавании изолированных слов и слитной речи [2]. Такая система распознавания дает неплохие результаты для диктора, на обучающей выборке которого происходило обучение распознаванию (оценка параметров). Этого диктора назовем опорным. Но для другого, нового диктора, эта же система распознавания дает не слишком хорошие результаты, если не сказать – плохие. Напрашивается вывод – провести точно такое же обучение для нового диктора, как и для опорного, с такой же большой обучающей выборкой. Но в данном случае вполне возможна следующая гипотетическая ситуация – либо новый диктор совершенно не имеет возможности наговаривать большую обучающую выборку (совершенно осознанно, из-за нехватки времени, например), либо не имеет ни малейшего желания этого делать (и с этим, пожалуй, тоже нужно считаться в определенном смысле), либо новый диктор совершенно не знает о том, что его речь распознается, и потому нет никакой возможности обращаться к нему с просьбой наговорить большую ОВ. Резонно возникает вопрос – а нельзя ли новому диктору произнести относительно небольшую ОВ из изолированных слов или слитной речи, а потом с помощью определенных методов приспособиться (адаптироваться) к уже существующей системе распознавания, обученной на опорного диктора, и при этом получить приемлемую надежность распознавания? Принципиально, такая возможность должна существовать. Сравнение видеоспектрограмм, полученных при анализе речи разных дикторов, показывает, что при всем разнообразии проявления индивидуальных особенностей голосов видеоспектрограммы одних и тех же слов достаточно похожи [3]. Таким образом, необходимо преобразовать речевые сигналы одного диктора в сигналы другого.

Следовательно, задача адаптации к голосу диктора предусматривает предварительное проведение обучения на голос определенного опорного диктора или базового кооператива дикторов. Затем осуществляется корректирование параметров акустических моделей фонем для нового диктора на относительно небольшой адаптационной выборке (АВ). Адаптация может применяться и к смене условий распознавания, например, при переходе на иной канал полу-

чения речевой информации (другой микрофон, телефонная линия).

Цель статьи – исследование и применение к украинской речи одного из наиболее распространенных подходов в адаптации на голос диктора при пофонемном распознавании.

Задача адаптации и методы ее решения

При пофонемном распознавании речи каждая фонема имеет свою акустическую генеративную модель, представляющую собой последовательность состояний с определенными переходами между ними. На этапе обучения распознаванию речи параметры этих моделей вычисляются на основании итерационных процедур для опорного диктора или базового кооператива дикторов [3, 4]. Для каждой фазы–состояния фонемы φ (рис.1) известны средний вектор $\mu = [\mu_1, \mu_2, \ldots, \mu_n]^T$ и ковариационная матрица Σ, размерностью $n \times n$, где n – размерность вектора первичных признаков речевого сигнала.

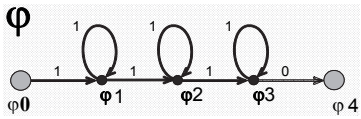

Рис. 1. Генеративная модель фонемы φ с тремя фазами–состояниями φ1, φ2, φ3

Начальное состояние φ0 и конечное φ4 служат для соединения с другими моделями фонем в словах. Цифра рядом со стрелкой обозначает количество временных отсчетов, за которое осуществляется переход по данной стрелке; μ и Σ – параметры n-мерного нормального закона распределения. Состояние модели может быть описано несколькими параметрами (парами), и тогда его описывают смесью гауссианов (нормальных распределений). Для двух систем распознавания, обученных на двух разных дикторов, эти вероятностные параметры будут различаться между собой, чем и объясняется неудовлетворительная точность распознавания какого-то диктора на чужой системе.

Имеется возможность вычислить линейные преобразования, которые переводят средние векторы и ковариационные матрицы опорного диктора или базового кооператива дикторов в средние векторы и ковариационные матрицы нового диктора. Действием этих преобразований есть 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 среднего вектора записывается в виде:

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

где $\hat{\mu}$ – средний вектор нового диктора, W – матрица преобразования размерностью $n \times (n+1)$, ξ – средний расширенный вектор опорного диктора,

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

В свою очередь матрица W представляется в виде:

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

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

Линейное преобразование ковариационных матриц представляется в виде:

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

где H– линейное преобразование размерностью $n \times n$, которое необходимо вычислить, B – разложение Холецкого для ковариационной матрицы Σ такое, что:

$$\Sigma = CC^T, \qquad (5)$$

$$B = C^{-1}. \qquad (6)$$

Матрицы линейных преобразований получают путем оптимизации значения критерия распознавания. Одним из таких оптимизационных алгоритмов есть линейная регрессия максимальной правдоподобности (*Maximum Likelihood Linear Regression – MLLR*) [4].

Для повышения гибкости процесса адаптации можно определить соответствующее множество базовых классов, которое будет зависеть от объема доступных данных адаптации [4]. Если доступны малые объемы данных адаптации, то тогда будет генерироваться общее адаптационное преобразование, применяемое к ка-

ждой компоненте гауссианов в множестве моделей. Однако, если адаптационных данных становится больше, то возможно улучшить адаптацию путем увеличения количества преобразований. В этом случае каждое преобразование становится более конкретным и применяется к определенной группе гауссианов. Например, гауссианы могут быть сгруппированы в широкие фонетические классы: пауза, гласные, назальные, фрикативные и т.д. И теперь адаптационные данные должны использоваться для вычисления более конкретных преобразований широких классов, чтобы применить такие преобразования к этим группам.

Связывание каждого преобразования через множество компонентов смеси позволяет адаптировать и те распределения, для которых вообще не было наблюдений. В таком процессе все модели могут быть адаптированы, и адаптационный процесс динамично улучшается как только появляется больше адаптационных данных.

Дерево классов регрессии построено таким образом, чтобы объединить компоненты близкие в акустическом пространстве, и, таким образом, похожие компоненты будут преобразовываться. Следует отметить, что дерево построено с использованием индивидуального диктороннезависимого множества моделей фонем, а значит – не зависит от какого-либо нового диктора. Терминальные узлы или листья дерева определяют конечные группы компонентов и называются базовыми классами (классами регрессии). Каждый гауссиан в множестве моделей фонем принадлежит к одному определенному базовому классу.

На рис. 2 приведен простой пример бинарного дерева регрессии с четырьмя базовыми классами, обозначенными как $\{C_4, C_5, C_6, C_7\}$. На диаграмме изображены сплошные стрелки и сплошные окружности, и это значит, что адаптационных данных, связанных с этим классом, достаточно для вычисления матриц преобразований. Пунктирные стрелки и окружности обозначают классы, для которых недостаточно адаптационных данных. В данном примере узлы 6 и 7 не имеют достаточно данных; но в узле 3, который есть родительским для 6 и 7, данных достаточно. Аналогично для узлов 5 и 2.

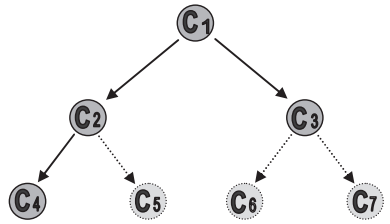

Рис. 2. Бинарное дерево регрессии

Преобразования вычисляются только для тех узлов, которые:

- имеют достаточно данных;
- являются либо терминальными узлами (то есть базовыми классами), либо имеют потомков с недостаточным количеством данных.

В этом примере преобразования вычисляются лишь для узлов 2, 3 и 4, и эти преобразования обозначим соответственно W_2, W_3, и W_4. Отсюда, когда нужно иметь преобразованное множество моделей фонем, матрицы преобразований применяются к компонентам гауссианов в каждом базовом классе следующим образом:

$$\left\{ \begin{array}{c} W_2 \to \{C_5\} \\ W_3 \to \{C_6, C_7\} \\ W_4 \to \{C_4\} \end{array} \right\}$$

Отметим, что случай общей адаптации похож на случай, когда дерево имеет только один корневой узел.

Экспериментальная база

Проведены две серии экспериментальных исследований. В первой серии было задействовано 67 дикторов (25 мужчин и 42 женщины). Учитывая тот факт, что надежность распознавания женских голосов ниже [5], количество женщин–дикторов было больше, чем мужчин. Каждый диктор достаточно четко наговаривал свою определенную ОВ, состоящую в среднем из 250 изолированных (отдельно произносимых) слов. Поскольку этих определенных ОВ было 10, то разные дикторы могли наговаривать одинаковые слова. Канал записи был для всех один, условия записи – практически студийные. Всего дикторами наговорено 2 416 разных слов.

В рамках этой серии экспериментов также проведены исследования с гендерным распознаванием.

Вторая серия экспериментов проводилась со спонтанной речью. Спонтанная речь заключалась в том, что дикторы, записи речей которых использовались в исследованиях, говорили свободно, не специально для каких-то экспериментов, порядок слов в их речи был свободным, некоторые слова они повторяли и не всегда полностью, говорили с разной степенью эмоциональности, в разном темпе, при этом речь была слитной. Распознавание также проводилось для слитной речи. Каналов записи было много, они различались между собой по характеристикам. Записи дикторов были не одинакового объема – от коротких по времени до длинных. Использовались записи с теле- и радиоэфира. Все эти записи были собраны в так называемый корпус АКУЕМ – «акустичний корпус українського ефірного мовлення» [6]. В этом корпусе словарь насчитывает 71545 словоформ, около 60 часов аудиозаписей, в которых представлена речь почти 2000 дикторов. Следует отметить, что дикторы произносили и такие слова, которых не было в словаре вообще, в отличие от первой серии экспериментов. Это усложняло ситуацию тем, что автоматически понижало точность распознавания. Большинство дикторов было представлено короткими записями, тогда как у почти 150 дикторов длина записей более 10 мин. Из этого следует, что условия распознавания здесь менее благоприятны, чем в первой серии экспериментов.

В алфавит фонем вошло 55 элементов. Все фонемы моделировались тремя состояниями Марковской цепи без пропусков.

Исследования

Первый эксперимент заключался в том, что создавались системы распознавания на базе одного опорного диктора. Затем проводилась адаптация новых дикторов к этим системам. Следует отметить, что при таком подходе получена невысокая точность распознавания после адаптации от 30 до 50 процентов. Отсюда возникла мысль о переходе от одного опорного диктора к кооперативу дикторов.

Во *втором эксперименте* был создан базовый (опорный) кооператив дикторов из 53 человек. Дикторы разного пола, возраста, из разных городов Украины. Остальные 14 дикторов вошли в контрольную группу. Дикторы из контрольной группы наговаривали один и тот же набор слов (241 слово), реализации которых не произносились дикторами базового кооператива.

Результаты второго эксперимента представлены на рис. 3. На нем графиком изображена усредненная точность распознавания до и после адаптации к базовому кооперативу дикторов контрольной группы на 30, 60, 100 и 150 слов. Видно, что после адаптации на голос нового диктора точность распознавания в среднем выросла на 3,66 процентов для адаптационной выборки объемом в 30 слов, на 4,45 – для 60 слов, на 5,33 – для 100 слов, на 5,93 – для 150 слов.

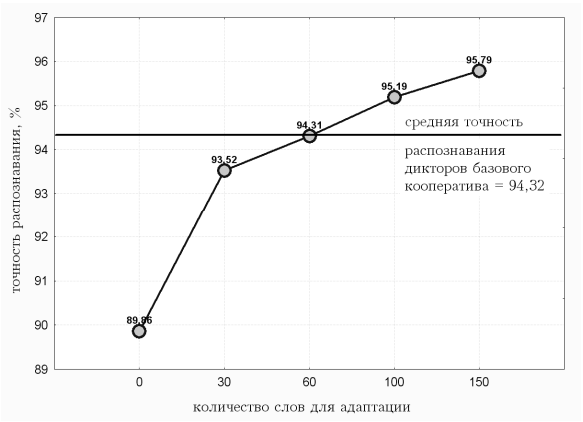

Рис. 3. Средняя точность распознавания дикторов контрольной группы до и после адаптации

При адаптации вычислялись матрицы перехода для среднего значения и дисперсии.

Третий эксперимент заключался в том, чтобы ответить на вопрос: что лучше – обучение нового диктора на выбранное количество слов (30, 60, 100 и 150) или адаптация на это количество слов? В итоге получилось, что лучшие результаты получаем при адаптации. Даже 150 слов на отдельное обучение нового диктора недостаточно, чтобы получить точность распознавания выше, чем при адаптации на этот же набор слов.

Идея *четвертого эксперимента* – базовый кооператив разделить на два по гендерному признаку. По такому же признаку контрольная

группа делилась 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 точности распознавания после адаптации.

Сравнительные графики точности распознавания в среднем по контрольной группе женщин–дикторов относительно базового кооператива и кооператива женщин–дикторов показано на рис. 4. Контрольная группа – семь женщин–дикторов, кооператив женщин–дикторов – 35 человек. Ясно видно, что гендерный подход в данном случае целесообразен, так как приводит к повышению точности.

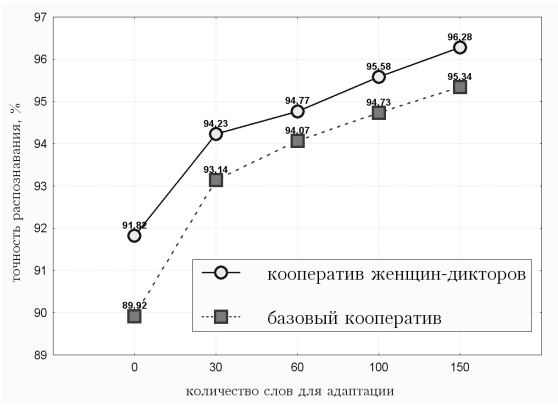

Рис. 4. Средняя точность распознавания дикторов женского пола

На рис. 5 изображены сравнительные графики точности распознавания в среднем по контрольной группе мужчин–дикторов относительно базового кооператива и кооператива мужчин–дикторов. Контрольная группа – семь мужчин–дикторов, кооператив мужчин–дикторов – 18 человек. С мужчинами–дикторами ситуация неоднозначна.

Во второй серии было проведено три эксперимента с тремя разными контрольными группами дикторов.

В *первом эксперименте* контрольная группа № 1 состояла из дикторов, принимавших участие в обучении. Это значит, что аудиозаписи разговоров этих дикторов были разделены на две части: записи из первой части полностью использовались при обучении системы распознавания (это была ОВ), записи со второй части использовались для тестирования и адаптации (это была независимая выборка (НезВ) этих дикторов). Цель данного эксперимента – выяснить, когда результаты адаптации будут лучше: когда АВ брать из ОВ или из НезВ.

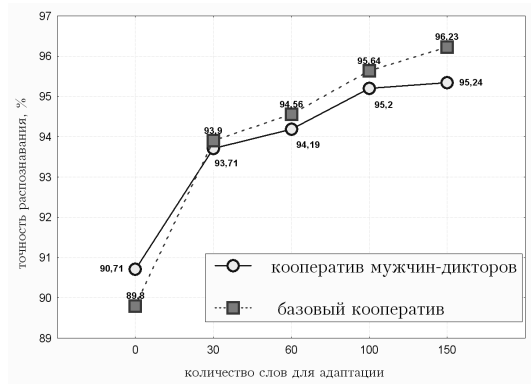

Рис. 5. Средняя точность распознавания дикторов–мужчин

Попутно необходимо было уточнить: как зависят результаты адаптации от количества линейных преобразований, которые применяются при этой самой адаптации, т.е. количество адаптационных данных не изменялось, АВ оставалась той же, а изменялся вручную порог достаточности данных в дереве классов регрессии. Чем больше порог, тем меньше будет линейных преобразований на всю систему при адаптации. Принималось четыре разных значения порога – 2000, 1000, 500 и 200. Также строились разные деревья классов регрессии – с 1, 2, 3, 4, 6, 8, 10, 13, 16, 20, 25 и 30 терминальными узлами. Для каждого дерева в зависимости от значения порога вычислялось разное количество линейных преобразований. Результаты данного эксперимента изображены на рис. 6.

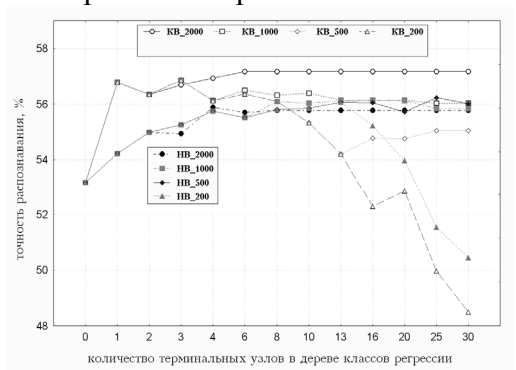

Рис. 6. Усредненная точность распознавания дикторов контрольной группы № 1 до и после адаптации

Пояснения к рисунку: КВ2000 – это значит, что АВ выбиралась из НезВ, значение порога

2000; НВ500 – АВ из ОВ, порог – 500. Когда количество терминальных узлов нуль, то это соответствует распознаванию без адаптации. Очевидно, что результаты адаптации лучше, когда АВ выбирают из НезВ (при порогах 2000 и 1000), при порогах 500 и 200 получаем весьма сомнительный результат. Получалось, что простое увеличение количества линейных преобразований вследствие понижения порога без роста объема АВ не приводит к автоматическому улучшению распознавания. Можно констатировать, что рост точности распознавания при выборе АВ из НезВ достигает почти четырех процентов (при пороге 2000), при выборе АВ из ОВ – почти трех при пороге 500. Результаты адаптации при выборе АВ из ОВ менее разбросаны. Исследования проводились при 16 гауссианах в смесях моделей фонем.

Во *втором эксперименте* контрольная группа № 2 состояла из дикторов, не принимавших участия в обучении. Записи разговоров этих дикторов не использовались при обучении системы распознавания, у них были только НезВ. Цель – выяснить, будут ли результаты адаптации для группы, не принимавшей участия в обучении, лучше чем для группы, принимавшей участие. Кроме того, как зависят результаты адаптации при увеличении количества гауссианов в смесях моделей фонем. Поскольку в предыдущем эксперименте при значении порога 200 получали неудовлетворительный результат, то этот порог тут не использовали. Результаты данного эксперимента изображены на рис. 7.

Пояснения: Г128_2000 – гауссианов 128, порог 2000. Однозначно при 128 гауссианах точность распознавания выше как до, так и после адаптации, результаты менее разбросаны. Рост точности – до 4,5 процентов (порог 2000) при 128 гауссианах, до 5,5 процентов (порог 500, 1000) при 16 гауссианах. Сравнивая с результатами первого эксперимента этой серии можно сделать вывод, что при 16 гауссианах результаты адаптации улучшились – 5,5 процентов против 4, хотя при этом говорить о значительной разнице не приходится.

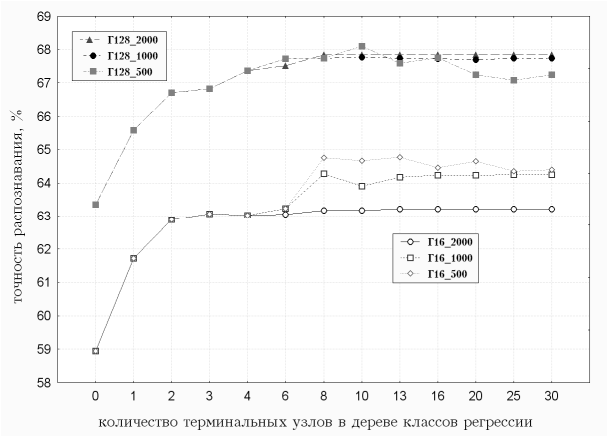

Рис. 7. Усредненная точность распознавания дикторов контрольной группы № 2 до и после адаптации при 16 и 128 гауссианах

В *третьем эксперименте* – контрольная группа № 3, дикторы которой не принимали участия в обучении, так как это – депутаты Верховной Рады Украины, т.е. они говорили со спецификой парламентских выступлений и со спецификой записей этих выступлений из парламентского зала заседаний. Цель – опять-таки экспериментально выяснить, будут ли результаты адаптации для группы, не принимавшей участия в обучении, лучше, чем для группы, принимавшей это участие. Поставлена задача: проводить адаптацию не для одной определенной АВ для каждого диктора, а для нескольких разных по объему АВ, чтобы оценить качество адаптации в зависимости от этих объемов. АВ для всех дикторов выбирались объемом в 30, 60 и 90 секунд. Деревьев классов регрессии построено немного меньше. Результаты представлены на рис. 8.

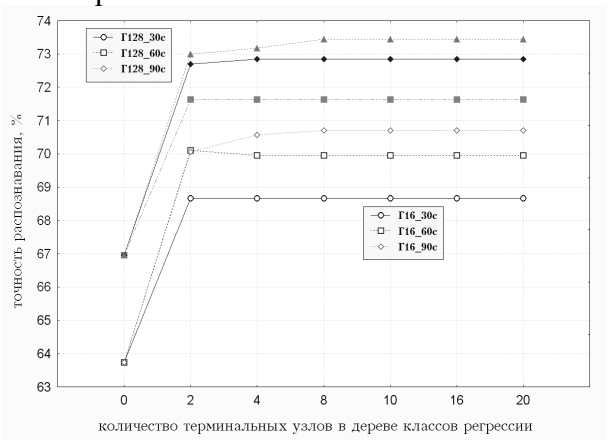

Рис. 8. Усредненная точность распознавания дикторов контрольной группы № 3 до и после адаптации при 16 и 128 гауссианах при значении порога 2000

Пояснения: Г16_60 с – гауссианов 16, объем АВ – 60 с. На графиках видно, что при увеличении объемов АВ растет точность распознавания после адаптации. Разница между результатами при АВ в 30 и 60 с больше, чем разница при АВ в 60 и 90 с. При 128 гауссианах наблюдаем рост точности от 4,5 процентов (при 30 с) до 6,5 (при 90 с), при 16 гауссианах – от 5 (при 30 с) до 7 процентов (при 90 с).

Заключение. Результаты экспериментов подтвердили целесообразность применения адаптации к голосу нового диктора.

Исследования гендерозависимого распознавания показывают уменьшение количества ошибочно распознанных слов до 10–20 процентов в сравнении с распознаванием на акустических моделях, сформированных на базовом кооперативе дикторов. Средняя точность распознавания самих дикторов из базового кооператива составляет 94,32 процента, при этом они наговорили свыше 12 тысяч слов в общей обучающей выборке. Преимущество адаптации очевидно.

Дальнейшая адаптация к голосу диктора на этой базе показала такую же динамику уменьшения ошибок для дикторов–женщин. Этот эффект не наблюдался для мужских голосов, очевидно, по причине меньшего количества дикторов–мужчин в базовом кооперативе.

Во второй серии экспериментов (со спонтанной речью) выяснено, что при увеличении количества гауссианов (16 – 128) происходит повышение точности распознавания. Однако после адаптации рост точности наблюдался при 16 гауссианах.

Для дикторов, принимавших участие в обучении, рост точности распознавания после адаптации был несколько больше, когда АВ выбиралась из НезВ. Для дикторов, не принимавших участие в обучении, рост точности распознавания после адаптации был больше в сравнении с дикторами, принимавшими участие в обучении.

Увеличение АВ улучшает результаты адаптации, по крайней мере до какого-то момента. Задача на будущее – выяснить, когда наступает этот момент, т.е. такие объемы АВ, что дальнейшее наращивание уже не приводит к росту точности распознавания.

1. *Vintsiuk T.K.*, *Sazhok N.N.* Speaker Voice Passport for a Spoken Dialogue System // Proc. of the 3rd Int. Workshop «Speech and Computer». – SpeCom'98. – St.–Petersburg, 1998. – P. 275–278.
2. *Vasylieva N.B.*, *Sazhok N.N.* Text Selection for Training Procedures under Phoneme Units Variety // Proc. of the 10th Int. Conf. on Speech and Computer – SpeCom'2005. – Patras, 2005. – P. 69–76.
3. *Винцюк Т.К.* Анализ, распознавание и смысловая интерпретация речевых сигналов. – Киев: Наук. думка, 1987. – 264 с.
4. *HTK* Book, v. 3.4 / S. Young, G. Evermann, M. Gales et al. // Cambridge University. – 2006. – 368 p.
5. *Peder Olsen.*, *Satya Dharanipragada.* An efficient integrated gender detection scheme and time mediated averaging of gender dependent acoustic models // Eurospeech'4, Sept. 1–4, 2003, Geneva Switzerland. – P. 2509–2512.
6. *Створення* акустичного корпусу українського ефірного мовлення / Н.Б. Васильєва, В.В. Пилипенко, О.М. Радуцький та ін. // Десята всеукр. міжнар. конф. «Оброблення сигналів і зображень та розпізнавання образів». – К.: УАсОІРО, 2010. – С. 55–58.

Тел. для справок: +38 044 502-6334 (Киев)