УДК 004.934

А.А. Юхименко

Подбор адаптационных выборок при адаптации к голосу нового диктора

В предыдущих исследованиях показана целесообразность применения адаптации системы пофонемного распознавания речи к голосу нового для системы диктора. Смысл адаптации заключается в улучшении распознавания нового диктора. Оговорен вопрос подбора наиболее оптимальных адаптационных выборок.

В попередніх дослідженнях показано доцільність застосування адаптації системи пофонемного розпізнавання мови до голосу нового для системи диктора. Сенс адаптації полягає в покращенні розпізнавання нового диктора. Обговорено питання підбору найбільш оптимальних адаптаційних вибірок.

Введение. Существуют задачи, когда система распознавания должна обслуживать достаточно большое количество пользователей. При этом все эти пользователи будут для системы незнакомыми, чужими дикторами. А поэтому система распознавания для этих чужих дикторов может выдавать не слишком приемлемую точность распознавания. Многочисленные эксперименты [1–4] показали целесообразность применения в этом случае адаптации имеющейся готовой системы распознавания к новому диктору вместо обучения этого диктора, так как ему придется наговаривать значительно больше речевого материала, что неприемлемо по условиям. Система распознавания оперирует определенным, может быть и небольшим, словарем. По условиям, новому диктору следует предложить наговорить *адаптационную выборку* (АВ) с небольшим количеством слов. Каким будет новый диктор – неизвестно. Какую АВ он подготовил, чтобы повышение точности распознавания было ощутимым?

Подходы в решении задачи

Итак, адаптация обеспечивает возрастание точности распознавания. Результаты экспериментов показали, что в среднем точность распознавания возрастает при увеличении объемов АВ. Но при этом наблюдается такой факт: приблизительно одинаковые по объему АВ (т.е. количество слов в них одинаковое, только количество фонем отличается несущественно) дают рост точности распознавания, которые могут отличаться достаточно существенно. Например, пригласили 67 дикторов (25 мужчин и 42 женщины) – дикторы разного возраста из пяти городов Украины. Каждый диктор достаточно четко наговаривал определенную обучающую выборку (ОВ), состоящую в среднем из 240 изолированных (отдельно произносимых) слов. Поскольку этих определенных ОВ было 10, то разные дикторы могли наговаривать одинаковые ОВ. Канал записи для всех был один, условия записи – почти студийные. Всего этими дикторами было наговорено 2416 разных слов. В алфавит фонем вошло 55 элементов. В *базовый кооператив* (БК) отобрано 53 диктора. ОВ дикторов БК использовались для обучения системы распознавания. Остальные 14 дикторов вошли в *контрольную группу* (КГ) и использовались для адаптации. Дикторы из КГ наговаривали одну и ту же ОВ (241 слово). Адаптационные выборки (АВ) выбирались из этой ОВ в 241 слово, а именно выбирались АВ по 30, 60, 100 и 150 слов. Реализация 241 слова не выполнялась дикторами БК. Все АВ были условно пронумерованы в группах по количеству слов. Так, АВ по 30 слов было семь и они пронумерованы от единицы до семи, АВ по 60 слов было шесть, по 100 слов – пять, по 150 слов – три. В КГ есть диктор Анна. У нее при адаптации на 30 слов набор № 2 дал рост 2,55 процента, а набор № 3 – 1,13 процента (разница более чем в два раза), при адаптации на 100 слов набор № 2 дает рост 3,04, а набор № 3 – всего 0,2 процента (вообще нет слов). И это при начальном распознавании (без адаптации) – 91,29 процента.

Отсюда – вывод, что брать АВ наугад – правильно, но недостаточно эффективно. А как же подобрать АВ поэффективней?

Рассмотрим некий вариант: пусть есть какой-то один определенный диктор, и есть словарь в системе распознавания объемом 1000 слов. Из определенных соображений зададимся целью

брать АВ объемом в 70 слов. Какой набор в 70 слов наилучший? Допустим, что диктор наговорил 1000 слов. Хотя в будущем смысл адаптации именно в том, чтобы любой новый диктор наговаривал не весь словарь, а именно достаточно небольшую АВ. Но для этого ее необходимо определить. Теоретически можно перебрать все наборы по 70 слов из словаря в 1000 слов и затем определить наилучший. Количество таких наборов определяется числом C_{1000}^{70}, что приблизительно составляет 10^{108}. Если непонятно, как предположить, что компьютер за секунду обработает миллиард АВ (т.е. процесс адаптации на АВ плюс распознавание по *контрольной выборке* (КВ)), то для полного перебора всех вариантов понадобится $10^{108}/(1\text{ млрд} \times 3600\text{ с} \times 24\text{ ч} \times 365\text{ дн.}) \approx$ $\approx 10^{91}$ лет. Невозможно! Если брать АВ объемом более 70 слов, ситуация ухудшится.

Тогда нужен иной подход. Например, предлагается взять некоторое количество АВ объемом в тех же 70 слов, чтобы можно было для этого количества выборок выполнить расчеты на компьютерах во временных рамках, которые будут более-менее устраивать при решении данной задачи. Затем АВ, которая дает наибольшее возрастание точности распознавания на КВ, использовать как окончательную для адаптации. Но пока эти рассуждения годятся для одного определенного диктора.

Относительно других дикторов – неизвестно, что дадут те же АВ для них? Предположим ситуацию, когда графики точности распознавания после адаптации в зависимости от АВ двух разных дикторов будут иметь некую схожую конфигурацию (рис. 1).

Из рис. 1 видно, что графики не совпадают, но участки возрастания и убывания у Диктора № 1 и Диктора № 2 соответствуют друг другу. Наибольшая точность распознавания при АВ в 30 слов достигается при наборе № 6, в 60 слов – при наборе № 6, в 100 слов – при наборе № 2, в 150 слов – при наборе № 3.

Тогда достаточно выполнить расчеты для одного диктора, и потом применять его наилучший вариант для всех остальных новых дикторов в

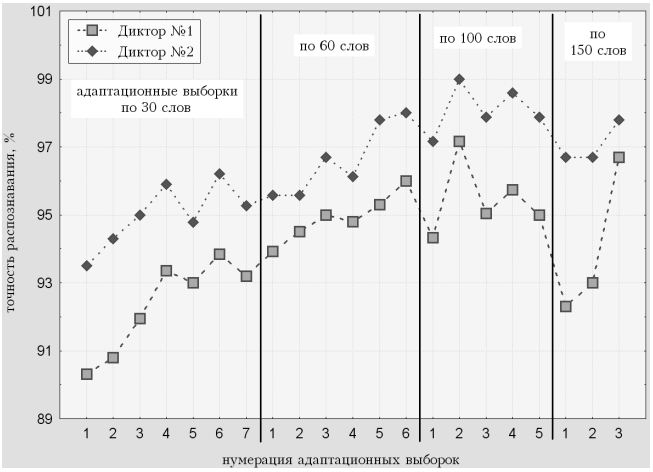

Рис. 1

системе распознавания. Однако реальность опровергает данное предположение. Возвращаясь к результатам упомянутого эксперимента, имеем, например, ситуацию, представленную на рис. 2, где отображена реальная картина для трех дикторов из КГ в 14 дикторов. Конфигурации графиков разнятся. Так, при АВ в 100 слов для диктора Анна лучший набор № 2, Юрий – № 5, Дмитрий – № 2 и № 4. А как поступить? Можно оставить все как есть, и новому диктору предложить наговорить те же 100 слов в АВ наугад. Но каковы будут результаты?

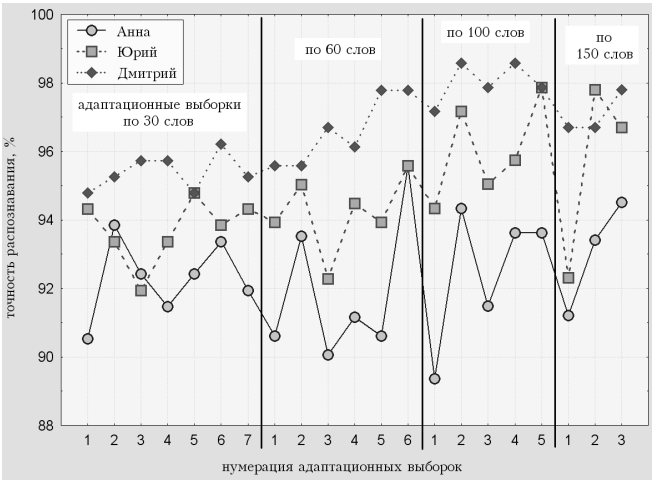

Рис. 2

Другой вариант. По мнению автора, он должен иметь достаточно хорошее практическое применение. В эксперименте было БК из 53 дикторов. По ОВ этих дикторов построена система распознавания, т.с. вычислены ее парамет-

ры. Также есть КГ из 14 дикторов. Понятно, что и БК, и КГ могут быть значительно больше. Очевидно, что это не ухудшит результаты. Дикторы КГ должны наговорить достаточное количество слов для ОВ. Для всех дикторов эта ОВ должна быть одинаковой. Одинаковой для всех должна быть и КВ. Из этой ОВ выбраны различные по количеству и составу наборы для АВ. По каждому такому набору для каждого диктора КГ проводим адаптацию и распознавание по КВ. Далее для каждого определенного набора выполняем подсчет среднего значения точности распознавания после адаптации по всем дикторам из КГ. Набор, дающий лучший результат, выбираем в качестве эталона для адаптации.

Исследования

Графики дикторов Анны и Дмитрия, а также график средних значений по всей КГ изображены на рис. 3.

По средним значениям видно, что лучший набор среди АВ по 30 слов – набор № 3, по 60 слов – набор № 2, по 100 – № 2, по 150 – №№ 2 и 3. В то время, как по Анне – №№ 2, 6, 2, 3 соответственно, по Дмитрию – №№ 6, 5 и 6, 2 и 4, 3 соответственно.

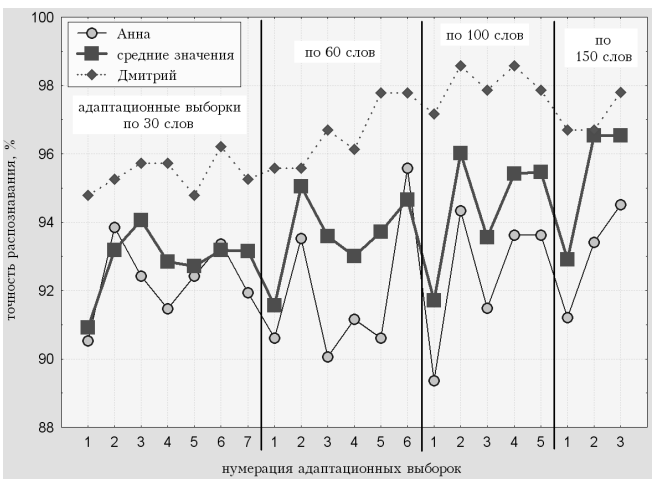

Рис. 3

На рис. 4 изображены графики усредненных значений точности распознавания после адаптации по группам слов для всех дикторов из КГ.

График № 1 – при адаптации вычислены матрицы преобразований для среднего вектора и его дисперсии, график № 2 – только для среднего вектора. Применение матриц преобразования для дисперсии также дает лучший результат.

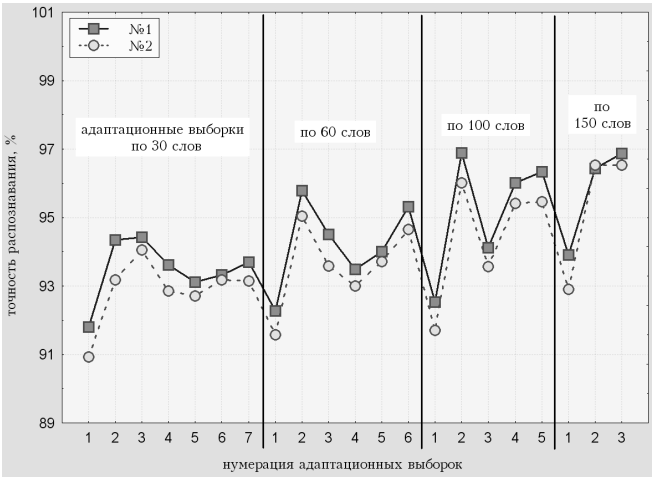

Рис. 4

На рис. 5 изображены графики средне-квадратичных отклонений (СКО) результатов дикторов КГ в каждом определенном наборе АВ. Графики № 1 и № 2 – аналогичны рис. 4. Применение матриц преобразования для дисперсии улучшает результаты, так как разброс становится меньшим.

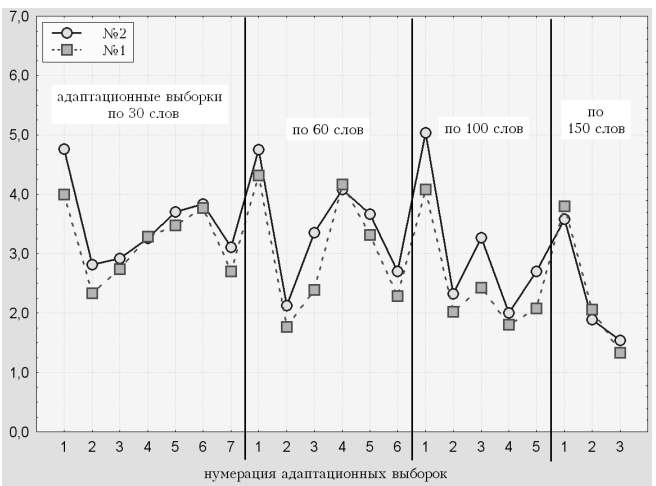

Рис. 5

Из рис. 4 и 5 видно, что лучшие результаты по группам слов имеют и лучшие результаты по СКО, т.е. наблюдается повышенная скученность вокруг лучших результатов, что есть положительным фактором. Из рис. 4 видно, что лучшим будет набор № 2 из АВ в 100 слов (лучше, чем набор № 3 из 150 слов. Если в условии задачи 100 слов в АВ для нового дикто-

ра – утомительны, а 60 слов – оптимальны, то следует выбирать набор № 2 с наименьшим СКО.

Если в группе слов есть одинаковые наибольшие результаты (см. рис. 4, наборы № 2 и 3, график № 2), то тогда в качестве эталона выбирают набор, у которого СКО меньше (набор № 3).

Автор провел дополнительные эксперименты по адаптации, выступив в качестве нового диктора и ОВ из описанного эксперимента. ОВ – 229 слов, КВ – 237 слов. Адаптация проводилась по АВ от 20 до 160 слов, выбранных из ОВ. Наборов по 20 слов было 21, по 30 слов – 20, по 40 – 19, по 50 – 19, по 60 – 18, по 70 – 17, по 80 слов – 16, по 90 – 15, по 100 – 14, по 120 – 12, по 140 – 10, по 160 – 8. На рис. 6 изображен график усредненных значений точности распознавания в зависимости от количества слов в АВ.

Видно, что уже 70–80 слов в АВ, в принципе, достаточно для получения почти наибольшего результата. Хотя 120 – лучше. Аппроксимирующая функция имеет вид

$$y = 73{,}76 - 21{,}55 / e^{0{,}02x} \qquad (1)$$

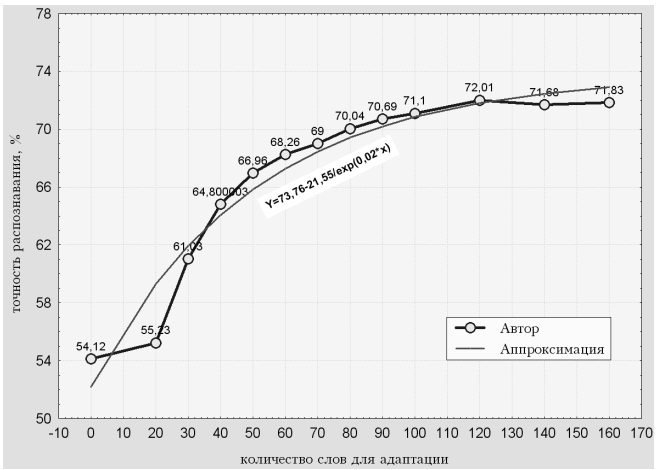

Рис. 6

Коэффициент корреляции $R = 0{,}97$, совпадение достаточно высокое.

Заметим, что здесь точность распознавания ниже, чем у КГ из предыдущего эксперимента. Очевидно, это произошло вследствие иного канала записи, чем при записи голосов КГ.

Заключение. В предложенном способе выбора в качестве эталонного набора для последующей работы с новыми дикторами используется набор, который дает результат лучше, чем средний результат точности распознавания по группе слов всех дикторов, что есть его достоинством (рис. 7). Горизонтальными линиями со значениями точности распознавания на графике изображено именно это среднее.

Недостатком, в определенном смысле, является то, что это – эмпирический подход, требующий громоздкой работы, так как чем больше привлечено дикторов в КГ и наборов адаптационных выборок, тем, очевидно, результат в среднем будет лучше. Но если времени и средств на разработку достаточно, то этот недостаток нивелируется.

По мнению автора, для адаптации следует брать хотя бы 70–80 слов. А если возможно, то лучше 100 слов и даже более.

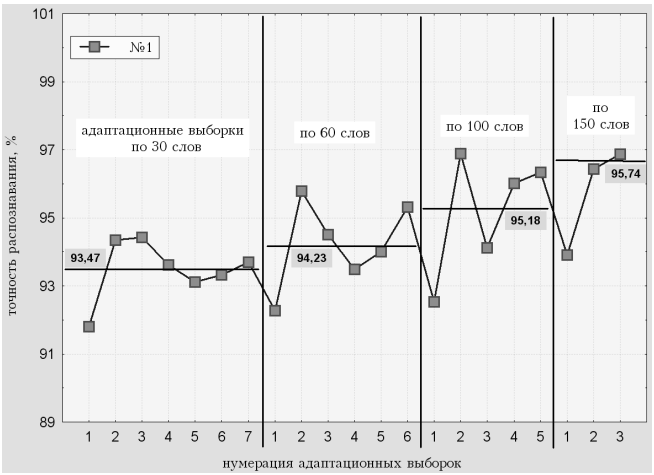

Рис. 7

Отметим также, что, исходя из [5], подбор эталонной АВ отдельно для мужчин и женщин даст положительный результат.

1. *Селюх Р.А.*, *Юхименко О.А.* Прикладне застосування адаптації на голос диктора в системах послівного розпізнавання // Искусственный интеллект. – 2012. – № 3. – С. 185–193.
2. *Юхименко О.А.* Експериментальні дослідження з адаптації до голосу диктора на базі корпусу АКУЕМ / Оброблення сигналів і зображень та розпізнавання образів: Одинадцята Всеукр. міжнар. конф. – Київ, 2012. – С. 51–54.

3. *Селюх Р.А.*, *Юхименко А.А.* Адаптация акустических моделей фонем к голосу диктора на основе метода *MLLR*. Речевые технологии // Народное образование. – 2012. – № 2. – С. 3–11.
4. *Юхименко А.А.* Адаптация к голосу диктора для пофонемного распознавания изолированных слов и спонтанной слитной речи украинского языка // УСиМ. – 2013. – № 4. – С. 85–91.
5. *Сажок М.М.*, *Селюх Р.А.*, *Юхименко О.А.* Адаптація до голосу диктора на основі гендернозалежних акустичних моделей фонем для української мови / Обробення сигналів і зображень та розпізнавання образів: Десята Всеукр. міжнар. конф. – Київ, 2010, С. 59–62.

E-mail: *enomaj@gmail.com*

UDC: 004.934

A.A. Yukhymenko

Selection of Adaptation Sets During Speaker Voice Adaptation

Keywords: recognition system, new announcer, adaptation, adaptation set.

Introduction: There are tasks, when the recognition system must serve to the plenty of users. All these users will be for the system unknown announcers. The system can not good recognize these announcers. So, a new announcer should have a small adaptation set. Thus, if a new announcer pronounce such adaptation set the growth of exactness of recognition will be perceptible.

Purpose: The purpose of work is experimental research of the new announcers speech recognition growth after their adaptation to the recognition system depending on the volumes of the adaptation sets. It is necessary also to probe the dependence of the exactness recognition growth from the high-quality composition of the adaptation sets.

Results: It is set in the numerous experiments, that exactness of the new announcer speech recognition after adaptation is increased on the average at the increase of adaptation set size. But this process is observed only on the average. Experiments shown that the practically identical on volume adaptation sets can give different exactness growth of the recognition substantially. The results of recognition are conducted on every adaptation set. After that the most effective set is selected.

Conclusions: The method of choosing the most effective adaptation sets is offered. This method is empiric and requires a lot of time. For greater efficiency it is necessary to work with the plenty of new announcers and the plenty of the adaptation sets. A conclusion is done about the amount of words, which must be in adaptation sets.

●