# 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 преградой 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.

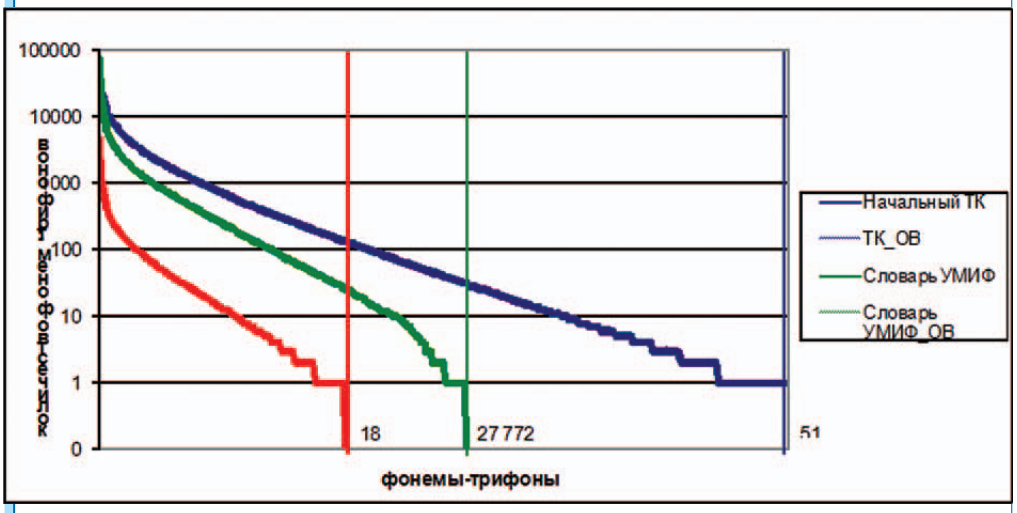


*Рис. 1.* Распределения фонем-трифонов по встречаемости в текстовых выборках

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: a, o, e, y, i, p, c, x, E, X, H, B, A, O, P, M, T;
- 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* 5*H 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:

- уставание 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 украино-язычной 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 $\mathbf{q}_{1:L} = q_1,\ldots,q_L$, которые наиболее правдоподобно генерируют последовательность наблюдаемых веторов $\mathbf{Y}_{1:T} = \mathbf{y}_1,\ldots,\mathbf{y}_L$, исходя из интегральной меры схожести:

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

где $\alpha$ и $\beta$ — коэффициенты, компенсирующие несоответствие шкалы акустической и лингвистической составляющих модели распознавания. Поэтому на первом этапе проводились эксперименты с целью эмпирически подобрать параметры $\alpha$ и $\beta$, 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 $\beta$ в пяти точках (0, – 5, – 10, – 15, – 20) для $\alpha$, 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.

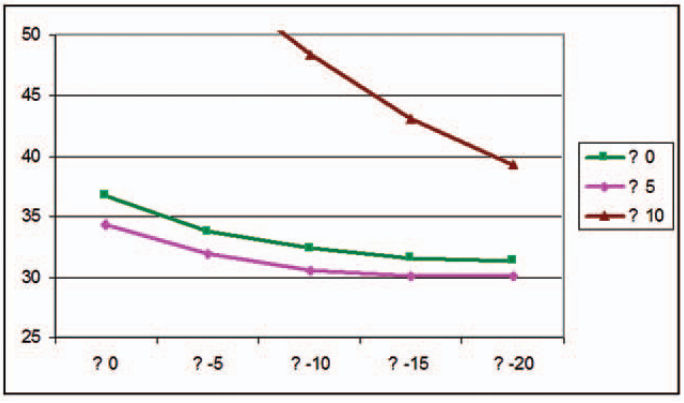


*Рис. 2.* Показатели *PER* распознавания (%) для слитной речи на «частотной» КВ

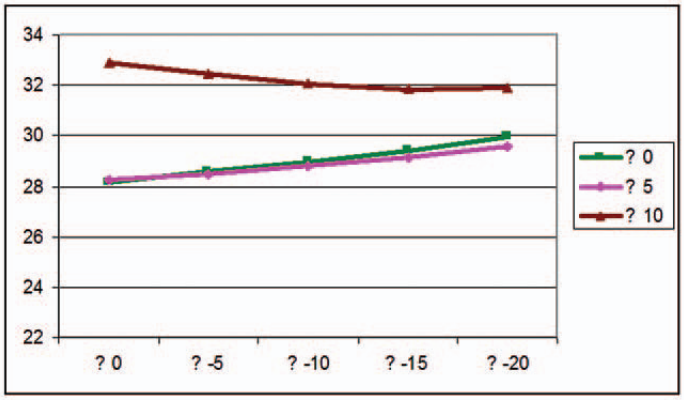


*Рис. 3.* Показатели *PIR* распознавания (%) для слитной речи на «частотной» КВ

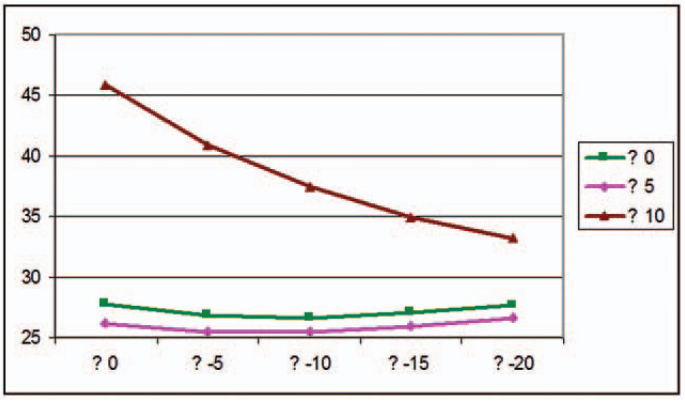


*Рис. 4.* Показатели *PER* распознавания (%) для слитной речи на «случайной» КВ

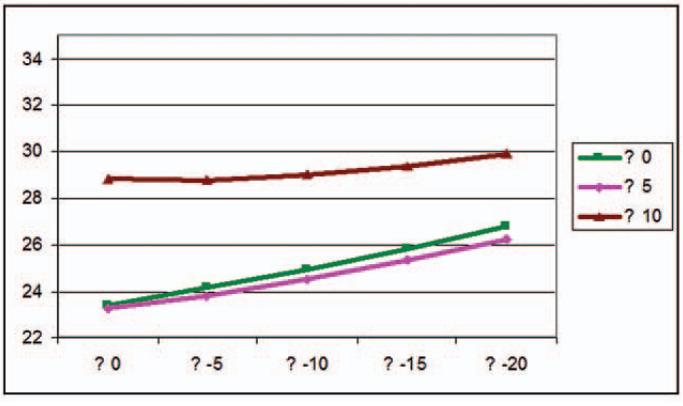


*Рис. 5.* Показатели *PIR* распознавания (%) для слитной речи на «случайной» КВ

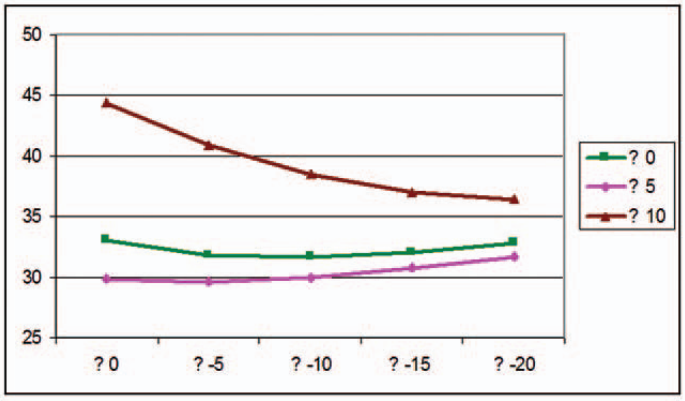


*Рис. 6.* Показатели *PER* распознавания (%) для слитной речи на КВ «Википедия»

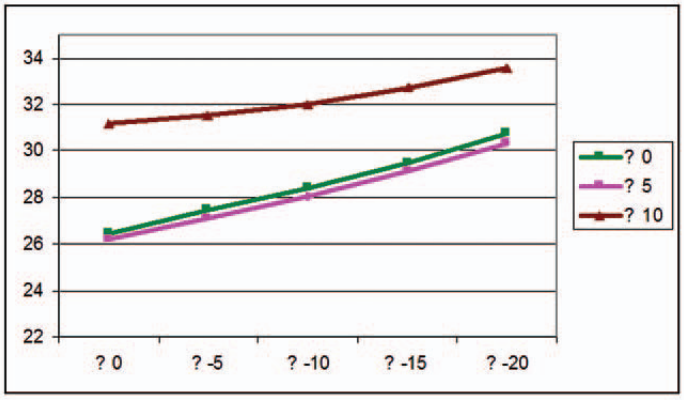


*Рис. 7.* Показатели *PIR* распознавания (%) для слитной речи на КВ «Википедия»

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 | |
|---|---|---|---|---|---|---|
| | *HTK* | *Julius* | *HTK* | *Julius* | *HTK* | *Julius* |
| «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 $\alpha$ и $\beta$, 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.* Modelyuvannya bahatorivnevoho poskladovoho rozpiznavannya 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 orforhafichnoho tekstu na fonemnyy. Shl. Donets'k, 2011, № 4. P. 117–126.

5. *Goncharov E., Kochetov Yu*. Povedenie veroyatnostnykh zhadnykh algoritmov dlya mnogostadiynykh 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 3nd 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.vassilleva@gmail.com; ninel@uasoiro.org.ua*