Дослідження зв'язку акустичного, фонемного та лексичного рівнів для розпізнавання злитого українського мовлення

Н.Б. Васильєва, М.М. Сажок

Відділ розпізнавання та синтезу мовлення,
Міжнародний науково-навчальний центр інформаційних технологій та систем,
м. Київ, Україна
{ ninel ; mykola } @uasoiro.org.ua

Abstract

This paper describes the development of multilevel multidecision systems for speech signal to text conversion, based on phonemes and syllables. The entire phoneme variety is extracted to select the training set. and control sets to estimate the parameters of acoustic recognition models. The estimation of acoustic model parameters is based on mono-speaker speech corpus. The factors compensating the inconsistency of acoustic and linguistic component model scales are analyzed and their values are explored. A way to convert phonetic decoder output to word sequences is described. The results of experimental research and future plans are discussed.

Keywords: multilevel speech recognition, syllable, control sets, continuous speech.

1. Вступ

Системи пофонемного розпізнавання зазвичай оперують алфавітом фонем (контекстно залежних або контекстно незалежних), з яких складаються мовленнєві образи слів. Потім на слова накладаються обмеження їх слідування шляхом побудови лінгвістичної моделі (ЛМ) або граматик. При збагаченні лексики зростають обсяги робочого словника, суттєво ускладнюються граматика або ЛМ, а це призводить до зменшення продуктивності системи розпізнавання.

Якщо використовувати замість слів мовленнєві образи складів або морфем, то збагачення лексики не призведе до помітного зростання робочих словників та ускладнення граматики чи ЛМ. При цьому постає проблема переходу від послідовностей складів (морфем) до послідовностей слів, оскільки помилка розпізнавання складу або морфеми може спричиняти ситуацію, коли їх послідовностям не можливо безпосередньо співставити послідовність слів.

В попередній роботі [1] досліджувалась надійність розпізнавання фонем і складів двох видів. Для проведення експериментальних досліджень використовувався однодикторний мовленнєвий корпус злитого мовлення. Велику увагу приділено створенню навчальної вибірки (НВ): вибору початкового текстового корпусу, алгоритму вибору текстів, оброблення "Жадібним" алгоритмом вибраних текстів, запису мовленнєвої НВ. Алфавіт корпусу НВ налічував близько 51 тис. фонем-трифонів у 18 тис. реченнях. Обсяг словника – 47,5 тис. слів. Загальна кількість реалізацій слів у цій НВ – 184,9 тис. Записано близько 36 годин запису акустичної бази навчальної вибірки. Також був описаний алгоритм вибору НВ для ізольованих слів. Розглядалися два словника: словник УМІФ та словник частотних слів. Обсяг словника НВ ізольованих слів склав ~ 13 тис. слів та після запису більш як 12 годин мовлення.

Графіки частоти фонем-трифонів у різних джерелах (текстовий корпус, словник УМІФ та частотний словник) та отримані відповідні НВ наведені на рис. 1. Тут зокрема можна побачити, що при роботі "Жадібного" алгоритму, кількість елементів, що зустрілися один раз, збільшилася в декілька разів для кожної НВ. Також із рисунка випливає, що частота фонем-трифонів загалом відповідає розподілу Ципфа–Мандельброта як для вхідних корпусів, так и після роботи "Жадібного" алгоритму.

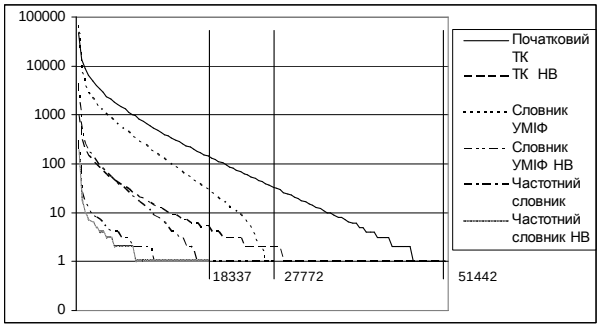

Рисунок 1: Розподіл фонем-трифонів по частотності в текстових вибірках

Експерименти проводилися на різних контрольних вибірках (КВ). Перша КВ формувалася за принципом частотності фонем-трифонів, що використовуються. "Частотна" КВ складалася із 3 тис. речень, обсяг словника мав 3 225 слів, загальна кількість реалізацій яких – 8 987 слів. Отримана КВ має 3,6 годин запису. Друга КВ вибиралася випадковим чином, з тих самих текстів, з яких вибирався текст НВ. "Випадкова" КВ складалася із 2 тис. речень, обсяг словника мав 10 013 слів, загальна кількість реалізацій яких – 22 864 слів. Отримана КВ має 4,3 годин запису. Третя КВ була вибрана з текстів, які не використовувалися для вибору НВ. Для цього із сайту української Вікіпедії [2] випадковим чином вибраний зв'язний текст (1,2 тис. речень). Обсяг словника складав 7,3 тис. слів. Загальна кількість реалізацій слів – 16 тис.

Процедура розпізнавання проводилась за допомогою декодерів *HTK* [3] і *Julius* [4] на трьох КВ: «Частотній», «Випадковій» та «Вікіпедії». В якості елемента робочого словника брались: фонеми (всього 59), відкриті склади (всього 7 270), склади, поділені за правилами складоподілу (всього 10 200) та цілі слова.

Метою даної роботи є дослідити зв'язки між акустичним, фонетичним і лексичним рівнями математичної моделі розпізнавання мовленнєвого сигналу для реалізації більш гнучкої схеми розпізнавання, що передбачає розподіл праці між фахівцями в галузях акустики, лінгвістики та інформатики.

В наступному розділі описується зв'язок між акустичним і фонетичним рівнями. Потім досліджуються параметри, які компенсують невідповідності шкал акустичної та лінгвістичної складової моделі розпізнавання. Четвертий розділ присвячено побудові моделі переходу від фонетичного до лексичного рівня. У висновках обговорюються результати досліджень та плануються подальші дослідження.

2. Зв'язок між акустичним і фонетичним рівнями

Акустичний і фонетичний рівень пов'язані через генеративні моделі фонем, які породжують еталонний сигнал згідно з деякими обмеженнями. Для схеми розпізнавання, що оминає фонемний рівень і переходить одразу до слів, ці обмеження задаються детерміновано у вигляді граматик або визначаються статистичною лінгвістичною моделлю, яка прогнозує поточне слово за деякою кількістю попередніх слів [5]. У випадку фонетичного стенографа, який безпосередньо не звертається до лексики, обмеження на порядок слідування фонем можуть бути взагалі відсутні або випливати зі складово-морфемної композиції включаючи статистику слідування фонем або складів/морфем [1].

Проводилося оцінювання параметрів акустичних моделей з використанням програмного інструментарію *HTK* та *Julius*. Акустичні моделі формувалися на основі контекстно-незалежних фонем, оскільки їх алфавіт невеликий, а отже, для статистичних оцінок необхідна менша база акустичних сигналів, ніж для складів і фонем-трифонів, яких більше в тисячі разів. Порівняно з попередньою роботою, в якій акустична модель будувалася лише на базі злитого мовлення, в даній розглядалася також модель фонем побудована як на злитому мовлені, так і на ізольованих словах. Це допомогло покращити розпізнавання для «Частотної» КВ, яка давала найгірші результати з обраних КВ.

У Таблиці 1 відображено фонемну помилку розпізнавання при використанні різної композиційності акустичних моделей для трьох навчальних вибірок.

Для кожної з 57 фонем української мови і двох фонем-пауз отримані моделі, які мають кожна три стани та від 4 до 36 сумішей нормальних законів в залежності від частотності. Акустичні моделі для розпізнавання будувалися як з повним урахуванням наголосу, так без урахування наголосу. Ніхто з відомих дослідників для Західних мов не розрізняє в алфавіті системи розпізнавання наголошені та ненаголошені фонеми. Для української мови таке спрощення навряд чи є прийнятним, оскільки доволі значна частка слів відрізняється лише позицією наголосу.

«Випадкова» КВ характерна тим, що частотність фонем у ній більш характерна для довільно взятого фрагменту предметної області. У «Частотній» КВ фонемне розмаїття представлено найбільш широко у якомога стислому обсязі. КВ «Вікіпедія» представляє іншу предметну область, в якій містяться фонемні контексти, не властиві навчальній вибірці.

Поруч із акустичною моделлю, що враховує наголошені та ненаголошені голосні, було створено модель, яка зовсім не розрізняє наголошені та ненаголошені відповідні голосні: *а*1 та *а*, *е*1 та *е*, *и*1 та *и*, *і*1 та *і*, *о*1 та *о*, *у*1 до *у*.

При оцінці надійності розпізнавання на акустичному рівні використовувались показники фонемної помилки (англійською, *PER – Phoneme Error Rate*) та фонемної некоректності (*PIR – Phoneme Incorrectness Rate*) [3].

За наведеними в таблиці результатами видно, що порівняно з попередніми дослідженнями [1], застосування акустичної бази навчальної вибірки ізольованих слів на додачу до злитого мовлення призводить до зменшення фонемної помилки. Це можна пояснити тим, що акустична база ізольованих слів збільшує кількість реалізацій кожної фонеми в контексті паузи, а наявність коротких синтагм (що характерно для природного людського мовлення) сприяє покращенню результатів розпізнавання.

Таблиця 1: Пофонемна помилка % при розпізнаванні ряду контрольних вибірок з урахуванням та без урахування наголосу на основі граматик вільного порядку слідування фонем (А) та бі-грамної статистичної моделі (Б), вільного порядку слідування відкритих складів (В) та складів за правилами складоподілу (Г)

КВ	А	Б	В	Г
"Випадкова"	25,60	24,80	23,06	**21,34**
"Випадкова" (без наголосу)	20,96	18,22	**17,00**	18,36
"Частотна"	26,70	27,68	23,22	**22,33**
"Частотна" (без наголосу)	21,56	20,05	**17,49**	18,11
"Вікіпедія"	30,76	**28,23**	28,53	29,35
"Вікіпедія" (без наголосу)	24,52	**21,28**	22,02	22,88

3. Компенсування невідповідності шкали акустичної та лінгвістичної складових моделі розпізнавання

Декодер намагається знайти послідовність слів або їх компонент $\mathbf{q}_{1:L} = \mathbf{q}_1,...,\mathbf{q}_L$, які найбільш правдоподібно генерують послідовність векторів, що спостерігаються $\mathbf{Y}_{1:T} = \mathbf{y}_1,...,\mathbf{y}_L$, виходячи з інтегральної міри схожості:

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

де α та β – коефіцієнти, які компенсують невідповідності шкали акустичної моделі (АМ) та лінгвістичної моделі (ЛМ), які є компонентами математичної моделі автоматичного розпізнавання мовленнєвого сигналу. Тому на першому етапі проводилися експерименти з метою емпірично підібрати параметри α та β, рекомендований діапазон яких складає 0—20 та 0—(–20) відповідно [3], [4].

Такі експерименти проводились для пофонемного та поскладового розпізнавання. На рис. 2–3 зображено показники %*PER* та %*PIR* пофонемного розпізнавання «Випадкової» КВ при змінах коефіцієнта β в п'яти точках (0, –5, –10, –15, –20) для α, що дорівнює 0, 5 та 10.

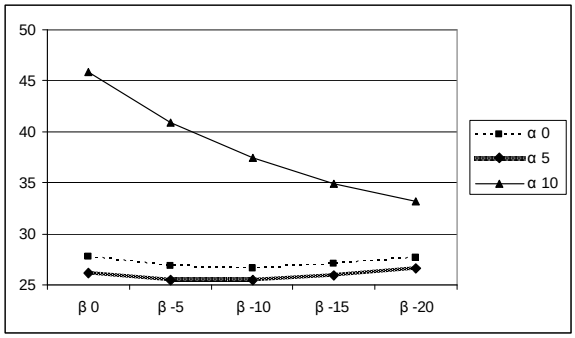

Рисунок 2: Показники PER розпізнавання (%) для злитого мовлення на “Випадковій” КВ

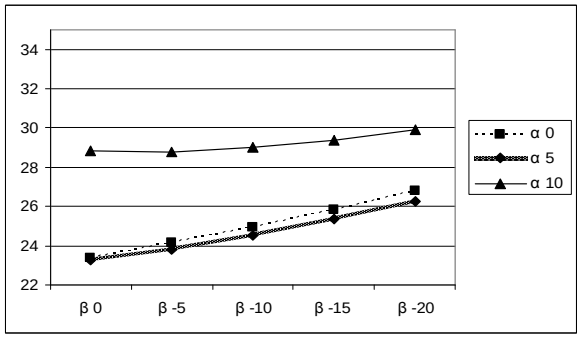

Рисунок 3: Показники *PIR* розпізнавання (%) для злитого мовлення на “Випадковій” КВ

Зменшення *PER* відбувається головним чином за рахунок скорочення кількості фонем, які помилково вставлені декодером. Ріст некоректності обумовлений зменшенням правильно розпізнаних елементів. Із рисунків випливає, що найменша фонемна помилка досягається при значеннях параметрів $\alpha = 5$ та $\beta = -5$. Показник *PIR* дає можливість стверджувати, що надійність виросла за рахунок скорочення числа вставок.

При розпізнаванні в умовах вільної граматики слідування складів та для інших КВ спостерігається подібна картина.

Загалом, важливість розглянутих параметрів α та β беззаперечна. Остаточне рішення про використання тих чи інших значень залежить від того, що важливіше: не втратити фонеми, що потенційно можуть збігатися з еталонами, або позбутися якомога більшої кількості зайвих елементів.

4. Моделювання переходу від фонетичного до лексичного рівня

Перетворення відповіді розпізнавання фонемного стенографа на послідовності слів все ще є найменш дослідженим питанням у схемі багаторівневого розпізнавання.

Відповіддю розпізнавання фонемного стенографа є послідовність фонем, що супроводжується інформацією про оцінку тривалості фонем та значенням критерію, що відображає впевненість системи щодо кожної окремо взятої фонеми. Цю послідовність фонем ми апроксимуємо результатом допустимих спотворень деякої фонемної транскрипції, що викликані як особливостями вимови та коартикуляцією, так і недосконалістю власне декодера й акустичної моделі. Останнє зумовлено зокрема тим, що при оцінці параметрів АМ вноситься суб’єктивізм експерта на етапі формування фонемних транскрипцій слів.

На рис. 4 наведено приклад генеративної моделі фонетичного рівня, що допускає спостереження еталонної фонеми у вигляді:

$$(q_{i-1}, q_i, q_{i+1}),\ q_i \in Q \cup \varnothing,$$

де Q – множина елементів, що включають у себе ім’я фонеми та (опційно) статистики тривалості та критерію. Для зручності ці моделі можуть бути розкладені на ряд простіших із подальшим їх поєднанням у граф динамічного програмування згідно з відомим транскрипціями слів при навчанні або з урахуванням допустимих обмежень при обробленні результату фонетичного стенографа [6].

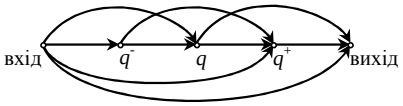

Рисунок 4: Узагальнена структура моделі фонетичного рівня фонеми q.

Оцінка параметрів моделей фонемного рівня полягає в застосуванні максимізації математичного сподівання (або критерію) на графах реалізацій для якомога більшої навчальної вибірки.

У попередніх роботах [6], не висвітлювалось питання, як сформувати початкову множину моделей фонемного рівня. Найбільш ефективний спосіб ініціалізації моделей фонетичного рівня повинен зводити до мінімуму їх кількість, не втрачаючи при цьому потенційних кандидатів. Для цього пропонується використовувати інформацію про найбільшу відповідність між спостережуваною та еталонною послідовностями слів. Оскільки пофонемна помилка в середньому складає близько 20% (Таблиця 1), то в середньому чотири з п’яти еталонних і спостережуваних фонем будуть збігатися.

Подібний випадок розглянуто в Таблиці 2, де наводиться приклад формування початкових моделей для реалізації речення «ми не глухі». Напівжирним шрифтом виділено імена еталонних фонем, яким гіпотетично відповідає спостережувана фонема з таким же іменем.

Таблиця 2: Ініціалізація гіпотез моделей фонетичного рівня

	Спостереження			Еталон	
№	Довжина	Критерій	Ім’я	Ім’я	Гіпотези
1	51	-22.15	*pau*	***pau***	*pau*
2	16	-30.19	*m*	***m***	*m*
3	9	-34.10	*y1*	***y1***	*y1*
4	8	-30.01	*n*	***n***	*n*, *n*+*y*1
5	3	-33.17	*y1*		
6	4	-32.71	*e*	***e***	*y*1-*e*, *e*
7	8	-28.27	*h*	***h***	*h*
8	6	-31.31	*l*	***l***	*l*, *l*+*l*
9	7	-35.32	*l*	*u*	*l*, *
10	21	-27.98	*kh1*	***kh1***	*kh*1, *l*-*kh*1
11	11	-29.78	*i1*	***i1***	*i*1, *i*1+*i*
12	18	-26.95	*i*		
13	15	-24.31	*t1*	*sp*	*sp*, *i*-*sp*

Моделі фонетичного рівня дають змогу сформувати множину гіпотез послідовностей фонем. Серед усіх гіпотез нас цікавлять лише ті, яким можуть відповідати одна або більше послідовностей слів. Інші гіпотези, очевидно, є неперспективними. З’ясувати це можливо лише звернувшись до лексичного рівня.

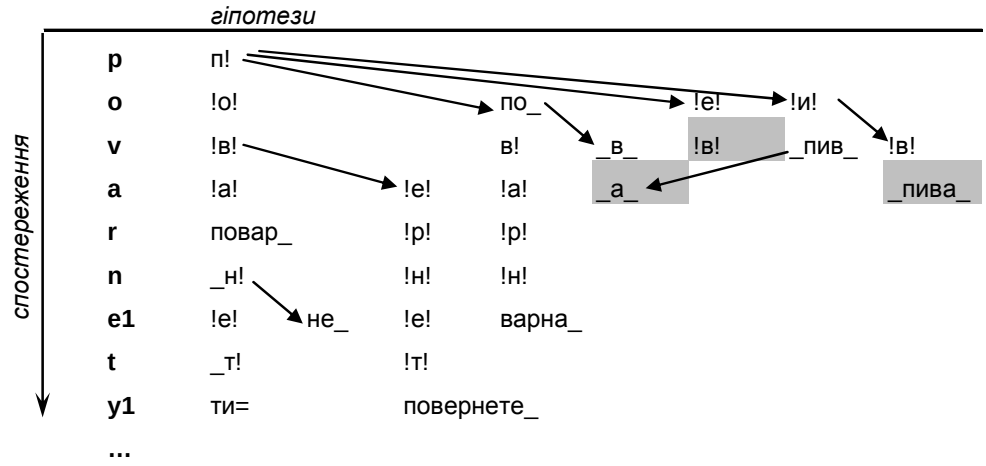

Рисунок 5. Приклад графа багатозначного переходу від фонемного до лексичного рівня.

На рис. 5 наведено граф переходу від фонетичного рівня до лексичного на прикладі фрагменту фонемної транскрипції, що відповідає слову «повернéте». На основі допустимих гіпотез послідовностей фонем у вузлах графу формуються гіпотези початків слів. Спеціальні символи використовуються як мнемонічний атрибут для позначення сегментів, що відповідають гіпотетичним границям слів, що допускають (_) або не допускають (=) продовження слова. Гіпотетичне слово виписується повністю, як тільки воно накопичиться шляхом конкатенації сегментів, що не можуть бути завершальними у слові (!).

Допустимі переходи між вузлами випливають з логіки атрибутів відповідних сегментів і накопичення орфографічного написання слова. Наприклад, від сегмента, що не допускає закінчення слова, перехід до початку слова не можливий. На рисунку затінено ті вузли, з яких не існує допустимих переходів.

На момент закінчення фрагменту з рис. 5 випливає, що з трьох траєкторій відновлюються послідовності повністю завершених слів: «повар не ти», «повернете» і «по варна ти». Яка з цих гіпотез має більшу вірогідність, вирішується на лексичному, синтаксичному та семантичному рівнях. З цих рівнів лише лексичний є достатньо опрацьованим у вигляді лінгвістичної моделі, що будується на основі передбачення слова за його попередніми одним, двома і більше словами [5].

5. Висновки

Робота спрямована на виключення композитних моделей слів з акустичного рівня декодера. Це позначилось на посиленні уваги до зв'язків між рівнями розпізнавання.

Згідно з проведеними експериментами, використання вільної граматики слідування складів показує кращі результати для предметної області, що відповідає навчальній вибірці. В іншому випадку виправдане застосування бі-грамної моделі обмеження порядку слідування фонем.

Велику увагу в даній роботі було приділено експериментальному дослідженню параметрів декодера, які компенсують невідповідності шкали акустичної та лінгвістичної складової моделі розпізнавання. Сумарне збільшення значення параметрів призводить до скорочення кількості розпізнаних фонем, що з одного боку зменшує коректність розпізнавання, а з іншого – підвищує надійність.

Для досліджень часткового врахування наголосу, потрібно обрати фонеми, для яких будуть розрізнятися наголошений та ненаголошений варіанти. Такими фонемами можуть бути *е* та *и*, оскільки їх ненаголошені реалізації на слух дуже схожі, а отже внесок ненаголошених реалізацій цих фонем у відповідні моделі з наголосом не є бажаним.

Планується застосувати статистичні обмеження послідовностей не лише для фонем, а і для складів, що має призвести до зменшення помилки розпізнавання. Залишається недослідженим вплив деяких параметрів декодування на надійність та швидкість. Зокрема, будуть розроблятися підходи до використання автоматично виведених сегментів та зменшення алфавіту складів, що має прискорити розпізнавання.

Досліджений багатозначний перехід між фонетичним і лексичним рівнем має бути вбудовано в багаторівневий декодер при генеруванні перспективних гіпотез послідовностей фонем. При цьому одночасно формуються послідовності слів, серед яких можуть у свою чергу зустрітися неперспективні послідовності ще до завершення декодування всього сегменту мовлення.

6. Література

[1]. Н. Васильєва. Використання граматик вільного порядку слідування фонем і складів для пофонемного розпізнавання злитого мовлення. Штучний інтелект. – Донецьк, 2011, № 4, 80-86 ст.

[2]. http://uk.wikipedia.org

[3]. Young S.J. et al., The HTK Book Version 3.4, Cambridge University, 2006.

[4]. A. Lee, T. Kawahara. "Recent Development of Open-Source Speech Recognition Engine Julius". Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, 2009, pp. 131-137.

[5]. Mark Gales, Steve Young. The Application of Hidden Markov Models in Speech Recognition. Foundations and Trends in Signal Processing Vol. 1, No. 3 (2007), pp/ 195–304.

[6]. N. Vasylieva, M. Sazhok T. Vintsiuk, G. Chollet. Acoustic-Phonetic Model Application for Syllable Speech Recognition Output Post-Processing. Proceedings of the 12th International Conference SpeCom'2007, Moscow, 2007, pp. 182-187.