DOI: https://doi.org/10.15407/usim.2018.06.007
УДК 004.934

Т.К. ВІНЦЮК, д-р техн. наук,
Міжнародний науково-навчальний центр інформаційних технологій та систем НАН та МОН України, просп. Глушкова, 40, Київ 03187, Україна

М.М САЖОК, канд. техн. наук, зав. відділом,
Міжнародний науково-навчальний центр інформаційних технологій та систем НАН та МОН України, просп. Глушкова, 40, Київ 03187, Україна,
sazhok@gmail.com

Р.А. СЕЛЮХ, мол. наук. співроб.,
Міжнародний науково-навчальний центр інформаційних технологій та систем НАН та МОН України, просп. Глушкова, 40, Київ 03187, Україна,
vxml12@gmail.com

Д.Я. ФЕДОРИН, мол. наук. співроб.,
Міжнародний науково-навчальний центр інформаційних технологій та систем НАН та МОН України, просп. Глушкова, 40, Київ 03187, Україна,
dmytro.fedoryn@gmail.com

О.А. ЮХИМЕНКО, наук. співроб.,
Міжнародний науково-навчальний центр інформаційних технологій та систем НАН та МОН України, просп. Глушкова, 40, Київ 03187, Україна,
enomaj@gmail.com

В.В. РОБЕЙКО, наук. співроб.,
Міжнародний науково-навчальний центр інформаційних технологій та систем НАН та МОН України, просп. Глушкова, 40, Київ 03187, Україна,
valya.robeiko@gmail.com

АВТОМАТИЧНЕ РОЗПІЗНАВАННЯ, РОЗУМІННЯ ТА СИНТЕЗ МОВЛЕННЄВИХ СИГНАЛІВ В УКРАЇНІ

Проведено аналіз стану та вказано основні шляхи вирішення проблеми автоматичного розпізнавання, розуміння та синтезу українського та багатомовного мовлення, усного перекладу з української та на українську. Узагальнено теоретичні та експериментальні результати доробку українських учених у галузі розпізнавання, розуміння та синтезу звукових образів. Наведено засоби комп'ютерного розпізнавання та синтезу українського мовлення.

Ключові слова: *мовлення, мовленнєвий сигнал, аналіз, розпізнавання, розуміння, синтез.*

Вступ

Розроблення та поширення засобів комп'ютерного розпізнавання та синтезу українського мовлення є найбільш досконалим засобом спілкування людини з машиною — за допомогою голосу. Людина подає усні команди, комп'ютер сприймає їх — розпізнає і розуміє. То є автоматичне розпізнавання та розуміння мовлення. Навпаки, якщо комп'ютер може озвучити (синтезувати) будь-який текст, то маємо справу з автоматичним синтезом мовлення за текстом.

Надалі автоматичне розпізнавання мовлення — це процес автоматичного оброблення

мовленнєвого сигналу, результатом якого є послідовність слів. Автоматичне розуміння мовлення — це більш узагальнений процес обробки мовлення, результатом якого є передаваний сенс (зміст) (нагадаємо, що один і той же сенс можна передати різними послідовностями слів). Аналогічно можна розрізняти простий й осмислений синтез мовлення за текстом. В останньому випадку комп'ютер спершу генерує осмислений текст, що виражає певну думку, а вже потім озвучує (синтезує) його.

Світовий рівень національної науки, техніки та культури значною мірою визначається наявністю розробок щодо комп'ютерних технологій та систем автоматичного розпізнавання та синтезу національного мовлення. Це зумовлено багатьма чинниками: мовлення є найбільш зручним, звичним, доступним і швидким засобом комунікації між людьми, а отже, найбільш придатним для спілкування людини з машинами; мовлення разом з мовою відіграє найсуттєвішу роль в національно-культурному та науково-технічному житті етносу.

В цій статті зроблено аналіз стану та вказані основні шляхи вирішення проблеми автоматичного розпізнавання, розуміння та синтезу українського мовлення, усного перекладу з української та на українську. Викладки будуть зроблені в контексті зв'язків та взаємовпливу національних наук та культур світу.

Специфіка проблеми. Мовленнєвий сигнал характеризується надзвичайними різноманітністю та надлишковістю. Аналоговий мовленнєвий сигнал з виходу мікрофона подається на аналого-цифровий перетворювач (АЦП) і далі у вигляді послідовності чисел подається в комп'ютер. Як правило, аналого-цифрове перетворення виконується в дискретному рівномірному часі з кроком 50 і менше мікросекунд (20 і більше тисяч вимірів миттєвої амплітуди мовленнєвого сигналу в секунду). Кожний вимір робиться з точністю 12—20 біт, тобто розрізняється від 2^{12} до 2^{20} значень миттєвої амплітуди. Отже, загалом мовленнєвий сигнал характеризується потоком більш ніж $12 \times 20000 = 2{,}4 \times 10^5$ біт/с або 3—10 байт/с.

Розглянемо, наприклад, задачу автоматичного розпізнавання окремо вимовлених слів. Нехай у словнику всього $2^8 = 256$ слів і в середньому кожне слово вимовляється за одну секунду. Розпізнати усне слово в цьому випадку означає перейти від початкової інформації об'ємом 3×10^4 байт до кінцевої інформації обсягом один байт про номер слова, який воно має в словнику. Отже, виходить, що $(3 \times 10^4 — 1)$ байт в мовленнєвому сигналі є надлишковими й тільки один байт («крапля в морі») несе корисну інформацію про сенс сказаного. Автоматичне розпізнавання мовлення — «боротьба» з надлишковістю, «вивуджування» «корисної» інформації, граничне стискання інформації.

Але головною «перепоною» вирішення проблеми розпізнавання є надзвичайне розмаїття мовленнєвих сигналів. Навіть два мовленнєві сигнали, що відповідають двом поспіль вимовам одного й того ж слова одним і тим же диктором, завжди є різними: при «накладанні один на одного» вони ніколи не дадуть збігу («Двічі в одну й ту саму воду ввійти неможливо»). В межах однієї мови, якщо не звертати увагу на діалекти, мовленнєві сигнали не тільки відрізняються тим, *що* сказано або яка послідовність слів вимовлена, але й залежать від індивідуальних особливостей голосу, функціонального й емоційного станів того, хто говорить, від способу та манери, темпу та гучності вимовляння, причому темп і гучність змінюються нелінійно в часі. Мовленнєві сигнали звуків змінюються під впливом сусідніх звуків у послідовностях — відбуваються так звані явища коартикуляції. Акустичні характеристики слів варіюються під впливом синтагматичних та фразових наголосів, а також змінюються залежно від інтонації — перелічування, звертання, завершеності, незавершеності, стверджування, питання, оклику тощо.

Загалом у мовленнєвому сигналі є інформація не тільки про те, *що* сказано, але й про те, хто говорить, який його функціональний стан, який темп мовлення тощо. Вся ця інформація виступає як надлишкова і зайва відносно інформації про те, що говориться.

Очевидно, що алгоритми автоматичного розпізнавання мають враховувати основні фактори змінюваності сигналів мовлення, базуватись на моделях параметричного процесу породження

мовленнєвих сигналів, який відображає розмаїття та закономірності генерації та перетворення мовленнєвого сигналу.

Сучасні уявлення про процес породження мовленнєвого сигналу базуються на нуль-полюсній моделі мовленнєвого тракту, яка збурюється голосовим (моделює голосові зв'язки — при генерації голосних звуків та дзвінких приголосних) або шумовим (при генерації аспіративних і фрикативних звуків) джерелами збурення, або їх комбінацією. Нуль-полюсна модель описується різницевим рівнянням не вище 20-го порядку (не більше 10 резонаторів). Параметри моделі та характеристики джерел збурення плавно змінюються в часі з так званою силабічною частотою й тим самим моделюють рух язика, губ, зубів, м'якого піднебіння, частоти коливань голосових зв'язок (періоду основного тону), інтенсивності. Голосове джерело моделюється імпульсом збурення, який має форму, близьку до трикутної, з плавним наростанням й різким заднім фронтом. Імпульс збурення виникає з періодом основного тону, що плавно змінюється в часі й займає десь близько третини періоду. Шумове джерело збурення моделюється генератором дискретного білого шуму.

Сигнали збурення фільтруються нуль-полюсною моделлю мовленнєвого тракту (аналог акустичної труби) й далі випромінюються у довкілля. Сигнал мовлення є результатом динаміки моделі мовотворення. У злитому мовленні мовленнєвий тракт не встигає налаштуватись на певну конфігурацію для окремих фонем, як уже подається команда на перебудову, на генерацію наступної фонеми. Але мовленнєвий тракт має інерційні властивості, його не можна миттєво переналагодити, отже, мовленнєвий сигнал генерується лінійною системою, параметри якої весь час змінюються. Розрізняють стаціонарні та перехідні частки (сегменти) мовленнєвого сигналу.

Для стаціонарних сегментів створюються моделі з більш-менш незмінними параметрами. Інтенсивність мовлення регулюється амплітудою джерел збурення, темп мовлення моделюється зміною довжини стаціонарних сегментів, інтонація — зміною періоду основного тону за певними законами в часі. Коартикуляція відображається в моделі тим, що значення параметрів мовленнєвого тракту для даного звука є залежними від параметрів попереднього й наступного звуків; до того ж значення цих параметрів «рухаються» за інерційними законами. Закономірності темпу мовлення «розігруються» в основному шляхом додержання певних співвідношень довжин стаціонарних сегментів звуків. Щоб генерувати звуки, відповідні фонемі, треба задати характерні значення параметрів моделі мовленнєвого тракту й джерел його збурення, рівно ж задати й закони їх зміни в часі.

Надлишковість і розмаїття мовленнєвих сигналів найкраще ілюструються моделлю мовотворення. Очевидно, що автоматичні розпізнавання та синтез мовлення явно чи неявно мають ґрунтуватись на наших уявленнях про генерацію мовленнєвого сигналу, про основні фактори, що пояснюють його розмаїття, а також дають можливість штучно генерувати мовленнєвий сигнал з наданням йому необхідних індивідуальних та емоційних властивостей.

Хоч усі люди на Землі й мають однакову анатомію і, отже, можна користуватись спільною моделлю мовотворення, все ж кожне національне мовлення характеризується власними фонемним складом, правилами артикулювання та інтонування, словотворення та об'єднання слів у речення тощо. Отрж, кожен етнос, щоб бути залученим до світової науки та культури, повинен досліджувати власні національні мову й мовлення.

Стан проблеми

В Україні проблему почали розробляти десь із середини 60-х років XX ст. Були виконані перші теоретичні розробки (Т.К. Вінцюк, В.С. Кириченко, В.О. Куниця, В.К. Малушенко, Б.Б. Тимофеєв, В.Г. Зайцев). Тоді ж були розроблені перші програми та пристрої, що розпізнавали декілька десятків окремо вимовлюваних слів (Т.К. Вінцюк, Б.Б. Тимофеєв і В.Г. Зайцев). Згодом сформувались наукові школи: під керівництвом Т.К. Вінцюка (Інститут кібернетики, а з 1997 р. — Міжнародний науково-навчальний центр інформаційних технологій та систем), М.П. Деркача (Львівський універ-

Рис. 1. Експериментальна система розпізнавання злитого мовлення на ЕОМ "БЭСМ-6" (1970)

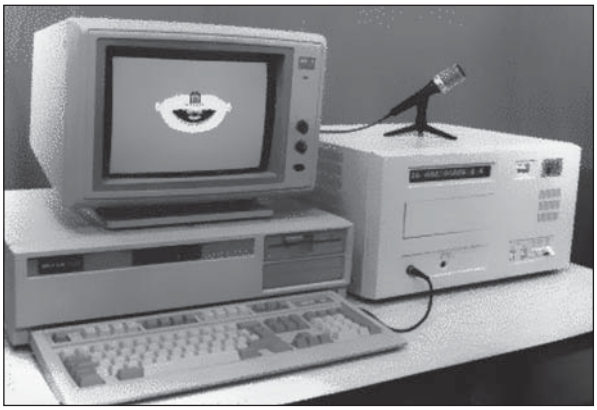

Рис. 2. Система усного діалогу RECH-121 (1986)

ситет), М.Ф. Бондаренка (Харківський інститут радіоелектроніки), О.М. Карпова (Дніпропетровський університет), Т.О. Бровченко та Е.О. Нушікян (Одеський університет).

Українські школи успішно конкурували на всесоюзній арені колишнього СРСР. Всесоюзні семінари «Автоматичне розпізнавання слухових образів» (АРСО) чотири рази проходили в Україні: 1968 р. — у Києві та Каневі, 1974 р. — у Львові, 1982 р. — у Києві та Одесі, 1988 р. — у Києві.

В Україні виконані певні наукові роботи, особливо теоретичного та експериментального плану, щодо моделювання процесів розпізнавання та синтезу мовлення. Добре знані в світі запропоновані в Інституті кібернетики (ІК) АН України загальні алгоритми обробки сигналів з метою їх розпізнавання та синтезу, відомі під назвою ІКДП-метод або Генеративна модель розпізнавання образів [1,2], а також експериментальні системи розпізнавання та розуміння злитого мовлення (рис. 1) [3,4], дослідні зразки систем усного діалогу (СУД) лінії МОВА-RECH (рис. 2) [5,6].

ІКДП-метод ґрунтується на ієрархічній (І) структурі породження (складання або композиції (К)) складних модельних сигналів мовлення й на порівнянні їх шляхом динамічного програмування (ДП) з розпізнаваним сигналом. Підвалини ІКДП-методу були закладені в 1966—1971 р.: спершу для розпізнавання окремо вимовлюваних слів (1968) [7], пізніше узагальнення на розпізнавання злитого мовлення (1971) [8], в тому числі на пофонемне розпізнавання [9, 10] й на смислову інтерпретацію злитого мовлення [11]. Одночасно розв'язувались різні задачі навчання та самонавчання розпізнаванню мовлення [10, 12, 13].

На Заході ця теорія відома як *DTW — Dynamic Time Warping* (динамічне згортання часу) та визнана піонерною в світі. З 1975 р. поширилася модифікація цієї теорії під назвою Приховані Марківські Моделі (*HMM — Hidden Markov Model*) та використовується в наш час у найбільш продуктивних системах розпізнавання [26—28].

Різні експериментальні системи розпізнавання окремо вимовлюваних слів демонструються в Інституті кібернетики АН УСРС з 1966 р. [14], зв'язного мовлення — з 1971 р. [3], смислової інтерпретації — з 1979 р. [15]. ІКДП-метод отримав широке визнання в світі, роботи українських вчених мають послідовників і цитуються в США, Великобританії, Франції, Німеччині, Японії та інших країнах [16—20]. Перші експе-

рименти з комп'ютерного синтезу українського мовлення відносяться до 1966 р. [21].

Розробки систем усного діалогу (СУД) лінії МОВА-*RECH* для практичного використання започатковані в 1978 р. Першою моделлю була СУД «*RECH*-1» (1980) [5]. З того часу розроблено цілу низку моделей: 1, 1001, 121, 2, 3, 111, 1111, 122, 123, 124, 4.

Моделі 1, 1001, 111, 1111, 122, 123, 124 є автономними, які підключаються до будь-якого мікрокомп'ютера. Моделі 2, 3, 4 вбудовуються в комп'ютер. Модель 1 (1980) розпізнає до 256 окремо вимовлюваних слів (усних команд) з надійністю розпізнавання 95 відсотків й синтезує (озвучує) будь-який текст українським чи російським мовленням зі словесною розбірливістю 97 відсотків. У Модель 1 реалізовано навчування на голос та словник користувача. Для цього кожне слово робочого словника має бути вимовлене хоч б один раз. При однократному вимовлянні кожного слова швидкість навчання (налаштування) СУД на розпізнавання усних команд дорівнює 30 слів за секунду. Щоб забезпечити 100 відсотків сприйняття (розпізнавання) усних команд використовується коментований режим введення інформації, за якого правильне чи неправильне розпізнавання коментується поданням усних команд ВІРНО або ПОМИЛКА. Орфографічний текст, поданий на озвучування, синтез, має бути розміченим сильним (—) або слабким (+) наголосами в словах.

Модель 1001 (1984) [22] додатково реалізує смислову інтерпретацію квазізлитого (з паузами між словами) мовлення, наприклад виконує усні завдання на виконання чотирьох арифметичних дій, які задаються природною мовою (порядок слів не є жорстко фіксованим).

Модель 121 (1986) [23] подана на рис. 2 розроблялась для мікрокомп'ютерів класу IBM PC XT/AT згідно з контрактами ЮНЕСКО. Вона є багатомовною (сім мов: українська, російська, англійська, французька, іспанська, німецька та італійська), виконує автоматичне розпізнавання окремо вимовлюваних слів (у словнику $256 \times 3 = 768$ слів) та злитого мовлення, усний переклад з однієї мови на іншу, синтезує (озвучує) будь-який текст на будь-якій із семи мов.

Розпізнавання виконується в реальному часі — затримка відповіді розпізнавання після закінчення мовлення не залежить від його тривалості й дорівнює 0,3 секунди. Максимальна тривалість злитого мовлення при розпізнаванні — 15 секунд. Налаштування на голос користувача, робочий словник і предметну область виконується в режимі навчання розпізнаванню. Кожній парі (диктор, словник), а, отже, й додатковій парі (мовлення-мова, предметна область) відповідає індивідуальний файл мовлення обсягом до 32 Кб. Надійність розпізнавання окремо вимовлюваних слів — 99 відсотків, слів у злитого мовленні для випадку вільного порядку слідування слів («коефіцієнт» розгалужень дорівнює обсягу словника) — 93 відсотків. Модель СУД «Мова-121» двічі успішно продемонстрована у штаб-квартирі ЮНЕСКО в Парижі [24].

В моделі 1111 було опрацьовано автоматичне розпізнавання та смислову інтерпретацію мовленнєвого сигналу для словника в 5000 слів.

Модель «Мова-4» є одноплатною й вбудовується в персональний комп'ютер (1986) Д [25]. Основою цієї моделі є мікропроцесор 1813ВЕ1. Модель розпізнає в реальному часі 300 усних слів-команд, а також озвучує (синтезує) будь-який текст українською або російською мовою. «Мова-4» розроблена на замовлення ВО «Електронмаш» (Київ).

Автоматичні розпізнавання та смислова інтерпретація мовлення

Найбільш робастною системою розпізнавання була б така, яка б запам'ятовувала всі можливі сигнали мовлення як прототипи, а потім, при розпізнаванні, порівнювала б розпізнаваний сигнал зі збереженими прототипами. На жаль, через надзвичайне розмаїття мовленнєвих сигналів цей підхід до розпізнавання не реалізується: немає й не буде такого комп'ютера, здатного запам'ятати всі можливі мовленнєві сигнали й, тим більше, їх порівняти. Проте, на щастя, при всьому розмаїтті, мовленнєві сигнали, пов'язані сильними детермінованими залежностями, які найкраще описати за вико-

ристання допустимих перетворень, які дозволяють перейти від однієї реалізації мовленнєвого образу до іншої. Наприклад, всі реалізації одного й того ж слова при вимовлянні одним і тим же диктором відрізняються нелінійно змінюваним темпом вимовляння. Отже, розмаїття мовленнєвих сигналів, зумовлене змінюваним темпом вимовляння, задається економним описом допустимих перетворень осі часу зі збереженням його прямого ходу.

Загалом є намагання задати (запам'ятати) окремі реалізації, наприклад кожного слова, які оголошуються модельними реалізаціями або прототипами, а потім визначити (описати) правила допустимих перетворень цих прототипів, застосовуючи які, утворюються (генеруються) різноманітні похідні модельні прототипи, наприклад, такі, що відрізняються нелінійно змінюваним темпом та інтенсивністю вимовляння. Нехай вдалося якимось чином побудувати модель, яка економно описує (або дозволяє генерувати) різноманітні модельні сигнали мовлення, які своєю сукупністю більш-менш якісно описують реальне розмаїття мовленнєвих сигналів. Якщо так, то в такий спосіб вирішується проблема пам'яті. Далі необхідно вирішити проблему обчислень — порівняння розпізнаваного сигналу зі генерованими модельними сигналами.

Отже, простий перебір модельних сигналів при порівнянні не підходить: необхідно знайти ефективні шляхи пошуку модельних сигналів, які певною мірою є найбільш схожими з розпізнаваним сигналом. Отож, має бути забезпечений ефективний спрямований пошук оптимальних рішень. Так, модель (алгоритм) розпізнавання має задовольняти дві вимоги: бути і адекватною, і конструктивною. Остання вимога означає економність опису розмаїття сигналів мовлення, а також направлений перебір варіантів у порівнянні сигналів.

В ІКДП-методі множини модельних сигналів мовлення задаються (описуються, генеруються) стохастичними породжувальними автоматними граматиками, а порівняння розпізнаваного сигналу мовлення з генерованими модельними (рівно ж формування відповіді розпізнавання) реалізується направленим перебором варіантів і пошуком оптимального рішення за динамічного програмування. Відповідь розпізнавання визначається згенерованим модельним сигналом, який є найбільш схожим (у певному сенсі) з розпізнаваним сигналом. В ІКДП-методі, підкреслимо, генерація і пошук оптимальних рішень виконуються спрямовано (без повної генерації й перебору всіх варіантів), але прийняті рішення еквівалентні повним генерації та перебору модельних сигналів.

Розглянемо процеси розпізнавання та смислової інтерпретації мовленнєвого сигналу.

Будь-якарозпізнавальнасистемаскладається з аналізатора та розпізнавача-інтерпретатора.

В аналізаторі виконується попередня обробка мовленнєвого сигналу, перехід від первинного опису мовленнєвого сигналу до вторинного (рис. 3). При аналізі з початкового обсягу інформації 8-96 КБайт/с виділяється найбільш суттєва його частина з не більш як 1-10 КБайт/с, яка все ще зберігає інформацію про те сказане. Універсальними ознаками мовленнєвого сигналу, поточні значення яких обчислюються на підставі спостережуваного мовленнєвого сигналу, як правило, виступають миттєві передавальна характеристика мовленнєвого тракту та параметри джерел його збурення або різні їх еквіваленти. Оскільки задача аналізу в цій постановці є некоректною і такою, яка погано піддається формалізації та розв'язанню, і враховуючи, що спостережуваний мовленнєвий сигнал є згорткою сигналів збурення з імпульсним відгуком мовленнєвого тракту, визначаємо робастним такий аналіз мовленнєвого сигналу, який ґрунтується на врахуванні квазіперіодичної структури згортки.

Обчислюватимемо поточні значення таких ознак: мовленнєвий сигнал є квазіперіодичним, шумним чи комбінованим. Якщо сигнал є квазіперіоднчиим, то виділятимемо поточний період основного тону, а саме — обчислюватимемо поточне значення періоду основного тону, а також запам'ятовуватимемо амплітудно-часову форму сигналу на довжині цього періоду. Відзначимо, що виділений поточний період мовленнєвого сигналу є згорткою одного імпульсу джерела збурення з мовленнєвим трактом. Якщо ж

поточний сигнал є шумним, то запам'ятаємо якийсь його фрагмент на стандартній довжині, наприклад 5 мс. Якщо ж сигнал є комбінованим (і квазіперіодичним, і шумним одночасно), то розділимо його на дві частини: низькочастотну до 2,5 кГц та високочастотну (шумову) — за 2 кГц, й запам'ятаємо повністю його низькочастотну, а також і шумову частини.

Очевидно, є сенс розрізняти окрім квазіперіодичного, "шумного" й комбінованого сигналів ще й відсутність сигналу мовлення. Далі введемо міру схожості двох сусідніх періодів й двох сусідніх шумових фрагментів. Якщо, наприклад, виявляється, що наступний період за формою повторює поточний й відрізняється від нього тільки інтенсивністю, тобто ці сусідні періоди мають відносно велику міру схожості, то є сенс запам'ятати тільки один перший період, а наступний за формою є повторенням старого, і для нього досить вказати тільки значення довжини періоду і відповідний множник зміни інтенсивності.

Міри схожості фрагментів мовленнєвого сигналу можуть мати різні вирази. Найбільш уживані міри схожості виражаються через амплітудний спектр, кепстр, автокореляційну функцію, так звані *a*- чи *b*-предиктивні параметри, коефіцієнти відбиття, коди або інші описи порівнюваних фрагментів [1]. Але у будь-якому випадку ці описи обчислюються на підставі виділених і порівнюваних фрагментів мовленнєвого сигналу (квазіперіодів або шумових фрагментів).

Отже, на виході аналізатора мовленнєвого сигналу маємо результат аналізу у формі часової послідовності спостережуваних елементів. Кожен елемент має ознаки:

- тон-шум (елемент є квазіперіодом, комбінованим або паузним);
- довжину (тривалість, наприклад, довжину поточного періоду);
- форму (амплітудно-часова форма квазіперіоду чи шумового фрагменту або еквіваленти: спектр, кепстр, автокореляційну функцію, предиктивні параметри, коефіцієнти відбиття, коди тощо, або форму попереднього елементу, якщо форма повторюється, але в даному випадку вказується множник зміни інтенсивності).

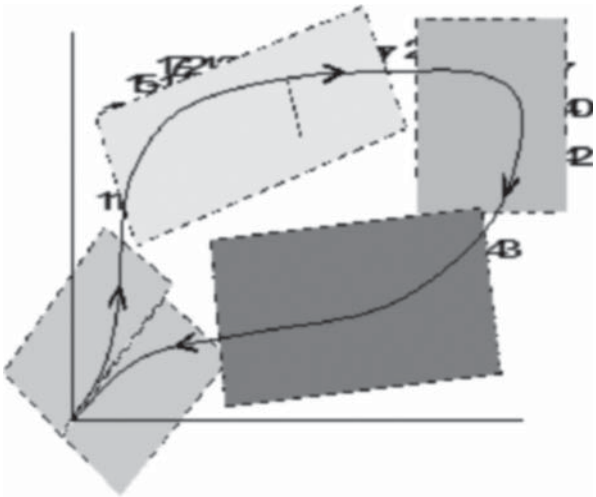

Рис. 3. Опис мовленнєвих сигналів векторними функціями часу — послідовностями векторів-елементів

Опис мовленнєвих сигналів в такий спосіб є зручним для створення бази знань, спільної як для автоматичного розпізнавання, так і для автоматичного синтезу мовлення.

Нехай далі є заданою сукупність сегментів (реалізацій) фонем для даного диктора. Сегментом (реалізацією) фонеми будемо називати *часову* послідовність спостережуваних елементів, яка певним чином «вирізана» дослідником з експериментального матеріалу. Сформуємо навчальну вибірку з сегментів, які відповідають певній фонемі при фіксованому оточенні іншими фонемами (та, що йде перед нею, і та, яка йде слідом). В українському мовленні розрізнятимемо близько 70 різних фонем (серед них є наголошені та ненаголошені голосні), й, отже, буде 70^3 фонем-трійок.

Із усіх реалізацій навчальної вибірки фонеми-трійки можна вибрати одну, найкращу, яку оголосимо прототипом фонеми-трійки і яка найкраще апроксимує всі інші реалізації навчальної вибірки. Отже, скористаємось мірою схожості сегментів-реалізацій, яка є сумою елементарних мір схожості — між двома порівнюваними елементами. Оскільки порівнювані реалізації мають різну кількість елементів (довжину), то у порівнянні їх довжини вирівнюються. То робиться шляхом зміни кількості повторень тих форм-елементів прототипу, які від-

значені аналізатором як такі, що можуть повторюватись. Загалом, у прототипі фонеми-трійки максимально зберігається природа вимовляння фонеми з урахуванням коартикуляції, допустимої зміни темпу та інтенсивності вимовляння, індивідуальних особливостей диктора.

Одночасно визначено певний процес породження (генерації) різних модельних сегментів фонеми-трійки: зберігаючи порядок слідування форм-елементів, окремі з яких повторюємо в певних межах, вказаних у прототипі фонеми-трійки проти кожної форми-елемента.

Далі розглянемо різні практичні задачі розпізнавання мовлення. Почнемо з розпізнавання окремо вимовлюваних слів. Нехай дано словник. Кожне слово задане своїм орфографічним текстом. Від орфографічного тексту слова переходимо до однієї чи декількох фонетичних його транскрипцій. Далі, виходячи з фонетичних транскрипцій слова, складаємо його прототипи шляхом об'єднання у послідовності відповідних прототипів фонем-трійок. При розпізнаванні пред'явлена реалізація слова порівнюється з перетворюваними прототипами його. При перетвореннях прототипів слова зберігається порядок послідовності форм-елементів прототипу та варіюється в дозволених межах повторюваність форм-елементів.

Процес порівняння та пошук найкращої міри схожості реалізується методами динамічного програмування. Розпізнавана реалізація відноситься до того слова, перетворений прототип якого дав найбільшу інтегральну схожість з розпізнаваним сигналом.

Аналогічно розглядаємо орфографічні тексти речень, їх фонетичні еквіваленти, а також творимо прототипи речень із прототипів слів та розглядаємо їх допустимі перетворення. Як і у випадку розпізнавання окремо вимовлених слів, процес перебору допустимих речень і можливих границь між словами досягається методами динамічного програмування. Відповіддю розпізнавання злитого мовлення є те допустиме речення, перетворений прототип якого виявився найбільш схожим з пред'явленим для розпізнавання сигналом мовлення. Найбільш простою задачею є розпізнавання злитого мовлення у випадку вільного порядку слів. Коли ж врахувати не тільки лексику, а й синтаксис та семантику мовлення, то на порядок слів накладаються додаткові обмеження. У цьому випадку процеси розпізнавання ускладнюються, вимагають великих обсягів пам'яті та швидкодії комп'ютерів.

Найбільш складною є задача смислової інтерпретації злитого мовлення. В межах фіксованої предметної області необхідно конструктивно задати структури, які породжують всі допустимі послідовності слів, що виражають один і той самий смисл, і це потрібно робити для всіх смислів, які можуть передаватись в межах предметної області. Для цього пропонується ієрархічна структура автоматних породжувальних граматик. На вищому рівні ієрархії генеруються всі можливі допустимі речення та тексти, що виражають один й той самий смисл, для всіх можливих при діалозі передаваних смислів.

Далі, опускаючись на нижчі рівні ієрархії та звертаючись до бази фонемних знань, для кожного з допустимих речень згідно фонетичних текстів та транскрипцій слів синтезуються допустимі модельні сигнали злитого мовлення. Ці сигнали складаються з модельних сегментів фонем-трійок. Синтезовані модельні сигнали порівнюються з розпізнаваним сигналом, результат порівняння використовується як зворотний зв'язок для генерації та направленого пошуку перетворених модельних сигналів, які є найбільш схожими з розпізнаваним сигналом в межах предметної області. Ці найбільш схожі модельні сигнали далі аналізуються: з'ясовується, яким послідовностям слів вони відповідають, і який зміст ці послідовності передають.

Базовою технікою для обчислень залишається те ж динамічне програмування, яке стає багатоступеневим, ієрархічним. У процесі смислової інтерпретації на мовному рівні використовуються спискові структури, наприклад LISP-структури, на рівні мовлення та на акустичному рівні — бази знань щодо фонем-трійок. Запропонована методика орієнтована на використання індивідуального файлу мовлення (бази знань щодо фонем-трійок), який формується в режимі навчання-самонавчання

розпізнаванню мовлення за мовленнєвою навчальною вибіркою.

Обговорювана техніка аналізу, розпізнавання та смислової інтерпретації мовлення потребує солідної комп'ютерної підтримки — кількох гігабайтів оперативної пам'яті та сотень мегафлопсів швидкодії.

Автоматичний синтез мовлення за текстом

Ця проблема є зворотною відносно розпізнавання. За експертними оцінками, вона є простішою, ніж розпізнавання відносно 10:1, 100:1, а то й 1000:1. При синтезі мовлення вже не є такою гострою проблема врахування розмаїття сигналів мовлення, і на передній план виходять аспекти моделювання індивідуальності синтезованого мовлення, надання йому натуральності та якості звучання.

Для моделювання синтезу індивідуального мовлення за текстом скористаємось індивідуальним файлом фонем-трійок, сформованим на підставі навчальних вибірок диктора в режимі навчання розпізнаванню.

Вхідний текст для синтезу мовлення відрізняється від звичайного орфографічного тексту тим, що в ньому додатково розставлені наголоси (сильні «—» або слабкі «+») в словах або синтагматичні розділові знаки . | , |_ | : | ; | ! |) |?|„|..|.

Приклад вхідного тексту:

До-брий де-нь! З Ва-ми гово-рить маши-на. Віта-ю Ва+с.

До+брий де-нь: З Ва-ми гово-рить машина. Бу-дьте здоро-ві! Ха-й Ва+м щасти-ть. Учітеся, брати-мої+.

Далі вхідний текст за допомогою автоматичного транскриптора трансформується у фонемний текст із фонетичних слів (слово чи сукупність кількох слів, об'єднаних одним сильним наголосом). При цьому, наприклад, прийменники приєднуються до наступних слів і утворюють із ними єдині фонетичні слова. У свою чергу, фонемний текст розбивається на синтагми (інтонаційно об'єднані послідовності слів), а кожна синтагма — на ритмогрупи (це підпослідовність з фонетичних слів, серед яких тільки одне слово з сильним наголосом).

Приклад фонемного тексту:

До+брий де-нь: ### Зва-ми гово-рить маши-на.

Бу-дьте здоро-ві! ## Ха-й ва+м шчасти-ть.

Учі-тесьа, брати- мойі+.

В цьому тексті шість синтагм (символом # позначена фонема-пауза): перша-синтагма «# До+брий де-нь:» складається з однієї ритмогрупи «# До+брий де-нь»; друга синтагма «### Зва-ми гово-рить маши-на.» має три ритмогрупи «### Зва-ми», «гово-рить» і «маши-на»; третя синтагма «## Бу-дьте здоро-ві!» складається з двох ритмогруп «## Бу-дьте» і «здорові»; четверта синтагма — з двох ритмогруп «Ха-й ва+м» і «шчасти-ть», п'ята синтагма «### Учі-тесьа,» — з однієї ритмогрупи «### Учі-тесьа»; шоста синтагма «брати- мойі+.» — також з однієї ритмогрупи.

Поняття синтагми та ритмогрупи використовуються модулями ритміки та інтонування. Модуль ритміки обчислює коефіцієнти-множники збільшення-зменшення «стандартної» довжини-тривалості для кожної фонеми-трійки, яка використовується в даному фонемному контексті, що озвучується. Коефіцієнти-множники визначаються процедурами-функціями, які залежать від поточної фонеми, що розглядається в поточному оточенні з сусідніх фонем, від її позиції в фонемному слові, в ритмогрупі, в синтагмі, від типу синтагми, який визначається розділовим знаком, а також від місця фонеми в слові відносно наголошеної голосної. За окремими правилами обчислюється значення коефіцієнта-множника для фонем — наголошених голосних. Найбільш поширений спосіб задання-опису ритмічних правил — логіко-табличні база знань і процедура обчислень.

В модулі інтонування або, як його інакше називають, просодики та енергетики, обчислюються поточні значення періоду основного тону — для дзвінких фонем, а також коефіцієнти-множники збільшення-зменшення інтенсивності кожної поточної форми-елемента в прототипі розглядуваної фонеми-трійки й з урахуванням «рекомендованого» модулем ритміки

збільшення-зменшення тривалості прототипа фонеми-трійки в даному контексті. Інтонаційна база знань має 30 типових інтонаційних контурів: 10 типів синтагм, по три типи ритмогруп на кожну синтагму. В кожній синтагмі розрізняють три типи ритмогруп: ядерна ритмогрупа (для українського мовлення вона за замовчуванням останньою в синтагмі), перед'ядерна ритмогрупа й початкова ритмогрупа, яка є першою в синтагмі. Якщо в синтагмі тільки одна ритмогрупа, то вона завжди є ядерною, якщо дві, то перша є початковою, а друга — ядерною. Типові інтонаційні контури «розігруються» на ритмогрупах: вибирається один контур за типом поточної синтагми і типом поточної ритмогрупи.

Кожний інтонаційний контур задається шістьма числами-коефіцієнтами, які вказують на відносне збільшення мінімальної частоти основного тону (або зменшення його максимального періоду). Шість чисел інтонаційного контуру визначають кусочно-лінійну (без розривів) зміну частоти основного тону для всієї довжини ритмогрупи, яка скоригована модулем ритміки: перший сегмент ламаної належить до перед'ядра ритмогрупи, другий, третій та четвертий сегменти — до ядра ритмогрупи, яким є сильнонаголошена фонема в ритмогрупі (з «—»-наголосом), п'ятий — до післяядра ритмогрупи.

Фонетичний транскриптор, модулі ритміки та інтонування є загальними для розглядуваної мови-мовлення, індивідуальними є акустичні бази знань про прототипи фонем-трійок.

Результати роботи фонетичного транскриптора, модулів ритміки та інтонування «апелюють» до бази знань про прототипи фонем-трійок певного диктора, голос котрого вирішено синтезувати.

Синтез мовлення ведеться посинтагменно, ритмогрупа за ритмогрупою, фонемне слово за фонемним словом, фонема —за фонемою. Для кожної поточної фонеми і її оточення вибирається відповідний прототип фонеми-трійки. Кожна чергова форма-елемент цього прототипу зчитується з бази знань, визначається кількість повторень цієї форми-елемента (згідно «вказівок» модуля ритміки), обчислюється інтенсивність та тривалість (періоду основного тону для дзвінких фонем) кожного повторення цієї ж форми-елемента (згідно корекцій модуля інтонування), і далі утворювані в такий спосіб «повторювані» й скориговані за тривалості та інтенсивності форми-елементи прототипу фонеми-трійки один за одним, у стик, передаються через цифроаналоговий перетворювач на динамік, де фізично озвучуються.

Апеляція до індивідуальної бази знань з прототипів-трійок диктора дозволяє надавати синтезованому мовленню достатні розбірливість і якість звучання, моделювати бажану індивідуальність мовлення.

За автоматичного синтезу мовлення повною мірою використовують лінгвістичні, фонетичні, акустичні наукові знання про національну мову та мовлення.

Прикладні розробки

Дещо про використання комп'ютерних систем і технологій, які базуються на автоматичному розпізнаванні та синтезі мовлення. Про реалізовані або можливі використання цих засобів багато сказано в науково-технічній літературі. У 2000—2010 рр. у Міжнародному науково-навчальному центрі інформаційних технологій та систем в межах ДНТП "Образний комп'ютер" розроблено портативні пристрої усномовної інформатики. В їх числі: засоби усного діалогу для комп'ютерів і АРМ на їх основі, автоматичний фонетичний стенограф (рис. 4), усні словники-перекладачі, в тому числі з української і на українську (рис. 5), цифровий диктофон з голосовим управлінням (рис. 6).

Інші приклади використання мовленнєвих технологій це — автоматичний друк та редагування текстів під диктування, усномовні довідниково-інформаційні системи, годинники та ваги, що говорять, управління телевізором за допомогою голосу, усномовні комп'ютерні технології навчання тощо.

Стенограф записує мовленнєві сигнали фонетичними транскрипціями (перетворення звук—текст).

Перекладач виконує усний переклад з української мови на англійську в межах обраної

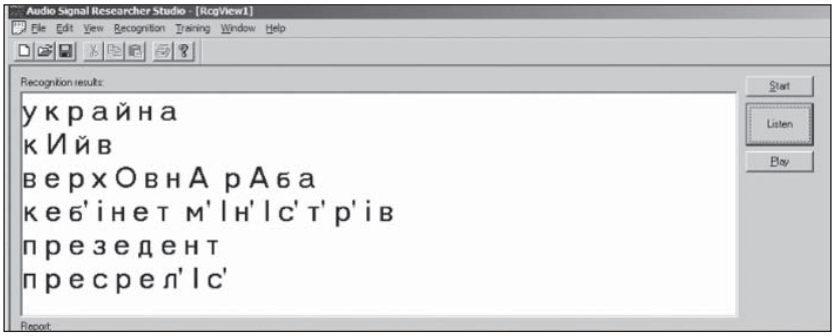

Рис. 4. Автоматичний фонетичний стенограф

Рис. 5. Портативний усний словник-перекладач

предметної області. Після вимовляння користувачем українською мовою слова або фрази результат автоматичного розпізнавання та перекладу озвучується українською і/або англійською. Є макетний пристрій на 300 слів/фраз на основі мікропроцесора цифрового оброблення сигналів (ЦОС) *ADSP*-2188*N*. В базовому пристрої використовується мікропроцесор ЦОС типу *BF*561, який за своїми технічними характеристиками дає можливість оперувати словниками до 10 тис. слів і більше.

Вокофон виконує функції запису та відтворення аудіоінформації, іменування та розмітка інформації, що записується, пошук інформації за ключовими словами, вимовленими користувачем, виконуються в режимі керування голосом. Є макетний пристрій на основі мікропроцесора цифрового оброблення сигналів (ЦОС) *ADSP*-2188*N*.

Слід звернути особливу увагу на можливість організації комп'ютерної допомоги людям з вадами зору та слуху у зв'язку з реалізацією автоматичного аналізу, розпізнавання, розуміння та синтезу мовлення. Окрім «банального» автоматичного перетворення в текст, який читається-сприймається глухими людьми, актуальним є перетворення мовленнєвого сигналу в зображення, які «читаються» людьми, або які перетворюють текст (автоматично читають текст) в мовлення та «портрет-говорун», що сприймаються людьми з вадами слуху [32].

Технології автоматичного розпізнавання та синтезу мовлення за текстом — де засоби комп'ютерної допомоги для сліпих. Тексти, що

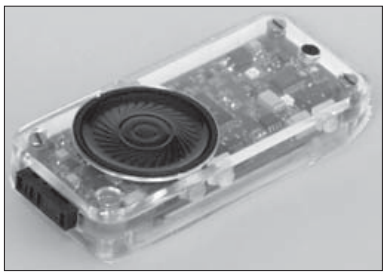

Рис. 6. Вокофон

висвічуються на моніторі комп'ютера, перехоплюються синтезатором мовлення, і незряча людина чує-«бачить» текст; клавіші, які він натискає, «називають себе голосом»; комп'ютер «читає синтезованим і бажаним голосом» електронну книгу, а на моніторі синхронно з мовленням жестикулює «портрет-говорун». Усе ж найбільшу допомогу згадувані засоби надають при навчанні дітей з названими проблемами.

Засоби автоматичного розпізнавання та синтезу мовлення особливо ефективні тоді, коли вони використовуються в комплексі з іншими засобами людино-машинної взаємодії (графічної, за допомогою зображень, малюнків тощо).

Сьогодення

Упродовж останнього десятиліття в загальних теоретичних підходах спостерігається певна збалансованість генеративних і дискримінативних моделей. Трансд'юсерне представлення дало змогу узагальнити методи комбінування генеративних моделей і їх оптимізації, що привело до краще обґрунтованих і більш гнучких конструкцій систем розпізнавання.

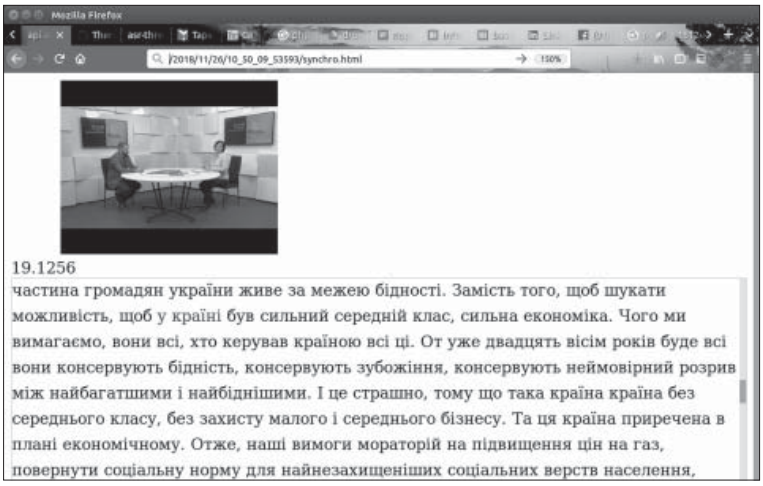

Рис. 7. Результат перетворення відеозапису на текст, починаючи з 19 сек., відтворений у браузері

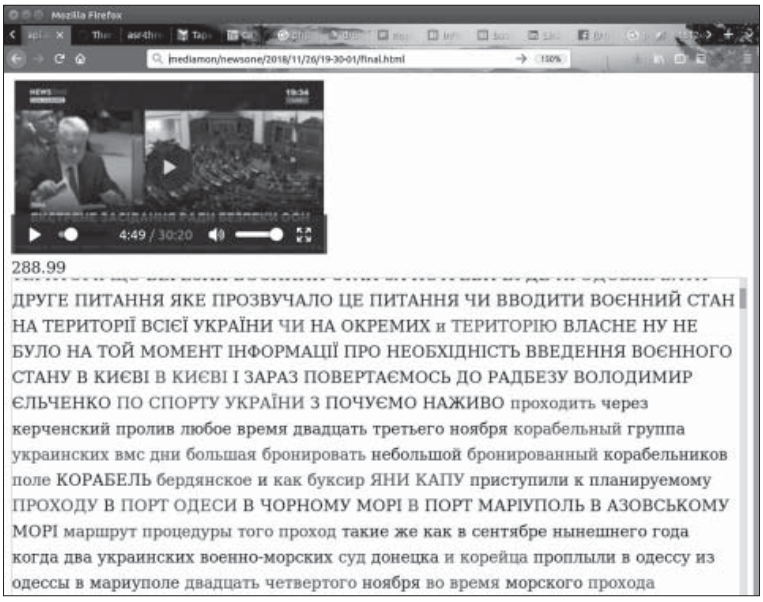

Рис. 8. Моніторинг каналу телебачення. Текст, який вважається системою українськомовним, відтворено великими літерами

Поширилось застосування багатошарового перцептрона в межах підходу, відомого як *Deep Learning* або *DNN*, для апроксимації в просторі вторинного опису сигналу областей, відповідних формам-елементам [33]. Хоч такий метод і не дає прямої можливості проводити процедури адаптації на голос диктора, на відміну від сумішей нормальних законів (*GMM*), він має певні теоретичні переваги (наприклад, не схильний до локальності), а системи з використанням поєднання *GMM/DNN* демонструють помітне поліпшення надійності розпізнавання

Намітилась тенденція використання не лише слів, а і класів еквівалентності слів, що представляють поняття, які певною мірою узагальнюють семантичні, синтаксичні та фонетичні властивості слів. Це дає змогу генерувати осмислені еталонні тексти в процесі розпізнавання зі значною економією оперативної пам'яті.

Новітні прикладні системи характеризуються охопленням більш розмаїтих сигналів мовлення, ширшої лексики, прогнозуванням знаків пунктуації, виокремленням метаданих і моделюванням реальної багатомовності. При створенні експериментальних систем поруч із

реалізаціями власних методів і алгоритмів використовується інструментальні засоби, розроблені міжнародною спільнотою та надані у відкритий доступ [29, 30]. Архітектура систем базується як на використанні ПК, так і на портативних пристроях (планшетах, смартфонах) та в "хмарі" на основі віддаленої взаємодії клієнт-сервер.

На сучасному етапі розвиваються певні технології та системи.

Серія досліджень, сконцентрованих на поглибленому аналізі окремих рівнів ієрархії мовленнєвих образів, дала змогу краще моделювати такі особливості української мови, як відносно вільний порядок слів і високу флективність, і створити експериментальні системи розуміння спонтанного мовлення та усного перекладу в межах предметних областей [34, 35].

Система автоматизації стенографування фонограм засідань (2008) вперше продемонструвала доцільність використання автоматичного перетворення мовлення на текст для створення стенограм [36].

Автоматизація замовлення квитків (2010) забезпечує розпізнавання початкового та кінцевого пункту відправлень, дати, кількості та типу квитків, здійснює голосові пояснення у діалозі та підтвердження замовлення [37].

Система диктування Диригент (2012) дає змогу вводити інформацію в комп'ютер голосом через мікрофонну гарнітуру в реальному часі, харак-теризується взаємодією з користувачем, покриває до 95 відсотків лексики [38].

Система *WebSten* (2015) перетворює довільні досить якісні записи спонтанного мовлення на текст [39]. У найновіших модифікаціях послівна надійність складає 70—90 відсотків залежно від наявності зокрема, шумів, завад, а також більше однієї мови (рис. 7 і 8). Пропонується як веб-сервіс з можливістю редагування тексту.

Експериментальна система моніторингу телерадіоефіру *MediaAudit* (2015) забезпечує більше 90 відсотків релевантності знайдених сюжетів, підтримує використання шаблонів під час пошуку, дає змогу прослуховувати аудіозаписи синхронно з розпізнаним текстом [39, 40].

Підсистему розпізнавання дикторів за голосом (з 2017) створено на основі моделювання індивідуальних особливостей мовця. З її використання визначаються метадані: хто говорить із відомих системі осіб, моменти переходу черги говорити до іншої людини. Реалізовано на ПК, портативному пристрої та в архітектурі клієнт-сервер [39].

Технологія, зворотна розпізнаванню мови — синтез мови за текстом — початково реалізована в системі «Текстофон» [31]. Вона забезпечує озвучення довільних україномовних текстів. Доступні чоловічий і жіночий голоси, можливо регулювання швидкості відтворення мови. Висока натуральність і розбірливість досягається шляхом обробки великих обсягів записів диктора. Встановлюється на ПК або сервер [39].

Перспективи розвитку

Завдяки стрімкому розвитку технологій в останні роки з'явилися перспективи вирішення складніших задач, таких як розпізнавання мовлення в шумах, автоматичне стенографування засідань та моделювання усного діалогу між людиною і комп'ютером. Пропонуємо й надалі розвивати напрямок функційного моделювання інтелектуальної, головне підсвідомої, діяльності людини та всього живого, що пов'язане зі сприйняттям слухових образів.

Висновки

Повне розв'язання задачі розуміння мовленнєвого сигналу рівносильне створенню комп'ютерів настільки ж інтелектуальних, як і людина. Вченими НАН України зроблено вагомий внесок у світовий розвиток галузі розпізнавання, розуміння та синтезу звукових образів.

За останні десятиліття в мовленнєвих технологіях здійснено значний прогрес завдяки збільшенню потужності обчислювальних ресурсів. Досягнуте підвищення робастності для окремих завдань у розпізнаванні значно наблизило рівень сприйняття мови комп'ютером до людського.

СПИСОК ЛІТЕРАТУРИ

1. *Винцюк Т.К* Анализ, распознавание и интерпретация речевых сигналов, Киев: Наук. думка, 1987б 264 с.
2. *Винцюк Т.К.* Сравнительный теоретический анализ ИКДП- и НММ-методов рпспознавания речи, Автоматическое распознавание слуховых образов: Тез. докл. 15-го Всесоюз. Семинара, Таллинн : Ин-т кибериетики АН Эстонии, 1989, С. 18—24.
3. *Винцюк Т.К, Гаврилюк О.Н., Пучкова И.Г.* Алгоритми распознавания слов и фраз и результаты их моделирования, Автоматическое распознавание слуховых образов: Тр. 8-го Всесоюз. Семинара, Львов : Изд-во Львовского ун-та, 1974, Ч. 3, с. 33—37.
4. *Винцюк Т.К., Гаврилюк О.Н., Куляс А.И., Шинкаж А.Г.* Система реального времени для распознавания слов и слитной речи, Автоматическое распознавание слуховых образов. Тбилиси: Мецниереба, 1978, с. 176—178.
5. *Винцюк Т.К,, Лобанов Б.М., Шинкаж А.Г.* Система распознавания речи и система усного диалога СРД «Речь» на основе микро ЭВМ, Автоматическое распознавание образов, Киев: ИК АН УССР, 1982, С. 516—521.
6. *Vintsiuk T.K.* Speech Dialogue Systems of the RECH Series , Proc. First Intern. Conf. on Information Technology for Image Analysis and Pattern Recognition, Lviv, 1990, Vol. 1, p. 367—370.
7. *Винцюк Т.К.* Распознавание слов устной речи методами динамического программирования, Кибернетика, 1968, № 1, с. 81—88.
8. *Винцюк Т.К.* Поэлементное распознавание непрерывной речи, составленной из слов заданного словаря, Там же, 1971, № 2, с. 133—143.
9. *Вінцюк Т.К.* Пофонемне розпізнавання зв'язної мови. Вихідні передумови і постановка задачі, Автоматика, 1972, № 6, с. 40—49.
10 *Вінцюк Т.К.* Пофонемне розпізнавання зв'язної мови. Алгоритми розпізнавання, навчання та самонавчання, Там же, 1973, № 1, с. 63—72.
11. *Винцюк Т.К.* Проблема автоматического понимания речи, Распознавание образов, Киев : ИК АН УССР, 1977, с. 28—34.
12. *Винцюк Т.К.* Обучение поэлементному распознаванию речи, Распознавание образов и конструирование читающих автоматов, 1969, Вып. 2, с. 23—35.
13. *Винцюк Т.К* Алгоритм определения эталонных элементов слова по совокупности его реализаций, Тр. Акуст. ин-та, 1970, Вып. 12, с. 163—168.
14. *Винцюк Т.К* Распознавание ограниченного набора речевых сигналов, Распознавание образов и конструирование читающих автоматов, 1966, Вып. 1, с. 135—149.
15. *Биатов КМ., Винцюк Т.К.* Система смысловой интерпретации слитной речи, Автоматическое распознавание слуховых образов 1982, Киев : ИК АН УССР, 1982, с. 365—368.
16. *Lienard J.S.* Le processus de la communication parlee, Paris etc.: Masson, 1977, 190 p.
17. *Bridle J.S., Brown M.D., Chamberlain R.M.* Continuous Connected Word Recognition using Whole Word Templates, The Radio and Electronic Eng., 1983, 53, № 4, p. 167—175.
18. *Ney H.* Dynamic Programming as a Technique for Pattern Recognition, Proc. 6th Intern. Conf. on Pattern Recognition, Munich, 1992, p. 1119—1125.
19. *Levinson S.E.* Structural Methods In Automatic Speech Recognition, Proc. of the IEEE, 1985, 73, №11, p. 1625—1650.
20. *Tscheschner W.* Probleme der automatischen Sprachverarbeitung aus heutiger Sich, Nachrichtentechnik, Electronic, 1979, 29, № 1,p. 26—29.
21. *Винцюк Т.К.* Распознавание некоторых классов речевых сигналов: Автореф. дисс. канд. техн. Наук, Киев, ИК АН УССР, 1967, 24 с.
22. *Винцюк Т.К* Смысловая интерпретация пословно произносимых фраз в СРД «Речь-1001», Автоматическое распознавание слуховых образов, Каунас, 1986, 4.1, с. 15—116.
23. *Final* Report on the UNESCO Contract SC/RP 261060.8 «Development of the Multilingual (including English, Russian languages) Speech Dialogue System for Micro-Computer», Kyjiv : Institute of Cybernetics, 1988, 97 p.
24. *Свідчення* досягнень радянської науки (Інформація ТАРС із Парижу), Рад. Україна, 17 грудня 1987 року.
25. *Система* речевого диалога СРД «Речь-4» для микроЭВМ «Поиск-2» (Отчет о НИР), Киев : ИК АН УССР, 1990, 171 с.
26. *L. Rabiner, B.-H. Juang.* Fundamentals of speech recognition. Prentice-Hall Int., 1993.
27. *Sadaok*i Furui. 50 years of progress in speech and speaker recognition. In Proc. of 10th Int. Conf. "Speech and Computer", Patras, Greece, 2005, p. 1—9.
28. *Daniel Jurafsky, James H. Martin.* Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. (2nd edition, 2014)
29. *Gales M., Young S.* "The Application of Hidden Markov Models in Speech Recognition." Foundations and Trends in

Signal Processing, 2007, 1(3), p. 195—304.
30. *Povey D.* "The Kaldi Speech Recognition Toolkit", Povey D., Ghoshal A., Boulianne G. et. al, IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.
31. *Вінцюк Т., Людовик Т., Сажок М., Селюх Р.* Автоматичний озвучувач українських текстів на основі фонемно-трифонної моделі з використанням природного мовного сигналу, Праці 6-ї Всеукраїнської міжнародної конференції "Оброблення сигналів і зображень та розпізнавання образів" — УкрОбраз'2002, Київ, 2002.
32. *Крак Ю, Вінцюк Т, Кириченко М., Гаращенко Ф., Бармак О.* Розробка комп'ютерних технологій моделювання та керування візуальними образами людського обличчя при синтезі мовлення, Мат-ли Шостої Всеукр. міжнар. конф. «Оброблення сигналів і зображень та розпізнавання образів» (УКРОБРАЗ'2002), 8—12 жовтня 2002р., Київ: Видання УАОІРО,2002, с. 23—26.
33. *Dahl G., Dong Yu, Li Deng, Acero A.* "Context-Dependent Pre-Trained Deep Neural Networks for Large Vocabulary Speech Recognition", IEEE Trans. Speech and Audio Proc., Special Issue on Deep Learning for Speech Processing, 2011.
34. 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.
35. Mykola Sazhok, Valentyna Yatsenko, Taras Vintsiuk. Interpretation of Continuous Ukrainian Pronunciation for Spoken Dictionary-Interpreter. — Proceedings of the 12th International Conference on Speech and Computer — SpeCom'2007, Moscow, 2007, pp. 170—175.
36. *Пилипенко В.В., Робейко В.В.* Автоматизированный стенограф украинской речи, Искусственный интеллект. Донецк: 2008. № 4.
37. *Пилипенко В.В., Биднюк С.А., Селюх Р.А., Пилипенко А.В.* Построение сценариев формализованного устного диалога на примере заказа билетов на железнодорожные поездаУСиМ, 2013, № 4, с. 71—75.
38. *Sazhok M., Robeiko V., Fedoryn D.* Distinctive features for Ukrainian real-time speech recognition system, Мат-ли XII Всеукр. міжнар. конф. «Оброблення сигналів і зображень та розпізнавання образів» (УКРОБРАЗ'2014), 2014 р., Київ: Видання УАОІРО, 2014.
39. Сажок Н.Н. Речевые информационные технологии и системы, УСиМ, 2017, № 2, с. 38—45.
40. *Сажок Н. Н., Робейко В.В., Федорин Д.Я., Селюх Р.А.* Система преобразования телерадиовещания в текст для украинского языка, УСиМ, 2015, № 6, с. 66—73.

Стаття надійшла 05.12.2018

REFERENCE

1. *Vintsiuk T.K.* Analysis, recognition and interpretation of speech signals, Kiev: Nauk. dumka, 1987, 264 p (In Russian).
2. *Vintsiuk T.K.* "Comparative theoretical analysis of ICDP and HMM methods of speech recognition", Automatic recognition of auditory images: Proc. report 15th All-Union. Workshop, Tallinn: Institute of Cybernetics, Estonian Academy of Sciences, 1989, pp. 18—24 (In Russian).
3. *Vintsiuk T.K., Gavrilyuk ON, Puchkova II.G.* "Algorithms for the recognition of words and phrases and the results of their simulation", Automatic recognition of auditory images: Tr. 8 All-Union. Seminar, Lviv: Publishing House of Lviv University, 1974, Part 3, pp. 33—37 (In Russian).
4. *Vintsiuk T.K., Gavrilyuk O.N., Kulyas A.I., Shinkazh A.G.* "Real-time system for word recognition and continuous speech", Automatic recognition of auditory images. Tbilisi: Metspiereba, 1978, pp. 176—178 (In Russian).
5. *Vintsiuk T.K., Lobanov B.M., Shinkaz A.G.* "Speech recognition system and oral dialogue system SRD "RECH" on the Basis of a Micro Computer", Automatic Pattern Recognition, Kiev: EC of the Ukrainian SSR, 1982, pp. 516—521 (In Russian).
6. *Vintsiuk T.K.* "Speech Dialogue Systems of the RECH Series", Proc. First Intern. Conf. on Information Technology for Image Analysis and Pattern Recognition, Lviv, 1990, Vol. 1, pp. 367—370.
7. *Vintsiuk T.K.* "Speech recognition by dynamic programming methods", Cybernetics, 1968, 1, pp. 81—88.
8. *Vintsiuk T.K.* "Item-by-element recognition of continuous speech made up of words from a given vocabulary", Cybernetics, 1971, 2, pp. 133—143 (In Russian).
9. *Vintsiuk T.K.* "Phoneme recognition of coherent language. Initial prerequisites and problem statement", Automation, 1972, 6, pp. 40—49 (In Ukrainian).
10. *Vintsiuk T.K.* "Phoneme recognition of coherent language. Recognition, learning and self-learning algorithms". Automation, 1973, 1, pp. 63—72 (In Ukrainian).
11. Vintsiuk T.K. "The problem of automatic speech understanding, Pattern Recognition", Kiev: EC of the Ukrainian Academy of Sciences, 1977, pp. 28—34 (In Russian).

12. *Vintsiuk T.K.* "Learning element-by-speech recognition, Pattern Recognition and Design of Reading Automata", 1969, 2, pp. 23—35 (In Russian).
13. *Vintsyuk T.K.* "Algorithm for determining the reference elements of a word from the totality of its realizations", Tr. Acoustic inst., 1970, 12, pp. 163—168 (In Russian).
14. *Vintsyuk T.K.* "Recognition of a limited set of speech signals, Pattern recognition and design of reading machines", 1966, 1, pp. 135—149 (In Russian).
15. *Biatov K.M., Vintsiuk T.K.* "System of semantic interpretation of continuous speech", Automatic recognition of auditory images 1982, Kiev: IC of the Ukrainian Academy of Sciences, 1982, pp. 365—368 (In Russian).
16. Lienard J.S. "Le processus de la communication parlee", Paris etc.: Masson, 1977, 190 p.
17. *Bridle J.S., Brown M.D., Chamberlain R.M.* "Continuous Connected Word Recognition using Whole Word Templates", The Radio and Electronic Eng., 1983, 53, 4, pp. 167—175.
18. *Ney H.* "Dynamic Programming as a Technique for Pattern Recognition", Proc. 6th Intern. Conf. on Pattern Recognition, Munich, 1992, pp. 1119—1125.
19. *Levinson S.E.* "Structural Methods In Automatic Speech Recognition", Proc. of the IEEE, 1985, 73, 11, pp. 1625—1650.
20. *Tscheschner W.* "Probleme der automatischen Sprachverarbeitung aus heutiger Sich", Nachrichtentechnik, Electronic, 1979, 29 (1), pp. 26—29.
21. *Vintsiuk T.K.* Recognition of certain classes of speech signals: author. diss. Cand. tech. Sciences, Kiev, IC of the Academy of Sciences of the USSR, 1967, 24 p.
22. *Vintsiuk T.K.* "Semantic interpretation of word-by-word phrases in the RDS "Speech-1001", Automatic recognition of auditory images, Kaunas, 1986, 4.1, pp. 15—16 (In Russian).
23. *Final* Report on the UNESCO Contract SC/RP 261060.8 «Development of the Multilingual (including English, Russian languages) Speech Dialogue System for Micro-Computer», Kyiv : Institute of Cybernetics, 1988, 97 p.
24. An indication of the achievements of Soviet science (Information TARS iz Parizhu), Rad. Ukraine, December, 17 1987. (In Ukrainian).
25. The system of speech dialogue of the SRD "Speech-4" for the Poisk-2 microcomputer (Research Report), Kiev: EC of the Ukrainian Academy of Sciences, 1990, 171 p (In Russian).
26. *L. Rabiner, B.-H. Juang.* Fundamentals of speech recognition. Prentice-Hall Int., 1993.
27. *Sadaok*i Furui. "50 years of progress in speech and speaker recognition". In Proc. of 10th Int. Conf. "Speech and Computer", Patras, Greece, 2005, pp. 1—9.
28. *Daniel Jurafsky, James H. Martin.* Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. (2nd edition, 2014)
29. *Gales M., Young S.* "The Application of Hidden Markov Models in Speech Recognition." Foundations and Trends in Signal Processing, 2007, 1(3), pp. 195—304.
30. *Povey D., Ghoshal A., Boulianne G. et. al,* "The Kaldi Speech Recognition Toolkit", IEEE 2011, Workshop on Automatic Speech Recognition and Understanding, 2011.
31. *Vintsyuk T., Lyudovyk T., Sazhok M., Selyukh R.* "The automatic speaker of Ukrainian texts on the basis of a phoneme-trifon model using the natural speech signal". Proceedings of the 6th All-Ukrainian International Conference "Processing Signals and Images and Pattern Recognition" - UkrObraz ' 2002, Kyiv, 2002. (In Ukrainian).
32. *Krak Yu, Vintsyuk T, Kirichenko M., Garaschenko F., Barmak O.* "Development of computer technologies for modeling and controlling visual images of a human face in the synthesis of speech", Mat-ly of the Sixth Allukr. international conf. "Processing of Signals and Images and Pattern Recognition" (UKROBRAZ'2002), October 8—12, 2002, Kyiv: Publications of UaIROO, 2002, pp. 23-26 (In Ukrainian).
33. *Dahl G., Dong Yu, Li Deng, Acero A.* "Context-Dependent Pre-Trained Deep Neural Networks for Large Vocabulary Speech Recognition", IEEE Trans. Speech and Audio Proc., Special Issue on Deep Learning for Speech Processing, 2011.
34. Vasylieva N, Sazhok M., Vintsiuk T., Chollet G. "Acoustic-Phonetic Model Application for Syllable Speech Recognition Output Post-Processing". Proceedings of the 12th International Conference SpeCom'2007, Moscow, 2007, pp. 182—187.
35. Sazhok M., Yatsenko V., Vintsiuk T. "Interpretation of Continuous Ukrainian Pronunciation for Spoken Dictionary-Interpreter". Proceedings of the 12th International Conference on Speech and Computer – SpeCom'2007, Moscow, 2007, pp. 170-175.
36. *Pilipenko V.V., Robeiko V.V.* Automated stenographer of Ukrainian speech, Artificial Intelligence. Donetsk: 2008, 4 (In Russian).
37. *Pylypenko V.V., Bidnyuk S.A., Selyukh R.A., Pylypenko A.V.* Formalized Scenarios Building for Speech Dialog Systems on the Example of a Ticket Train Service, *Upravlausie sistemy i masiny*, 2013, 4, pp. 71—75 (In Russian).
38. *Sazhok M., Robeiko V., Fedoryn D.* Distinctive features for Ukrainian real-time speech recognition system, Proceedings of XII Vseukr. international conf. "Processing signals and images and image recognition » (UKROBRAZ), 2014., Kyiv: Vydannya UAOIRO, 2014.

39. *Sazhok M.M.* "Speech information technologies and systems", *Upravlausie sistemy i masiny*, 2017, 2, pp. 38—45 (In Russian).
40. *Sazhok N.N., Robeiko V.V., Fedoryn D.Ya., Selyukh R.A. "Broadcast Speech-to-Text System for the Ukrainian Language". Upravlausie sistemy i masiny, 2015, 6, pp. 66—73 (In Russian).*

Received 05.12.2018

Taras Vintsuk, Doctor of Technical Sciences, Professor, Head of the Department,
International Research and Training Center for Information Technologies
and Systems of the NAS and MES of Ukraine, Academician Glushkov ave., 40, Kyiv, 03187, Ukraine

Mykola Sazhok, PhD in Techn. Sciences, Head of the Department,
International Research and Training Center for Information Technologies
and Systems of the NAS and MES of Ukraine, Academician Glushkov ave., 40, Kyiv, 03187, Ukraine
sazhok@gmail.com

Ruslan Selyukh, researcher, International Research and Training Center
for Information Technologies and Systems of the NAS and MES of Ukraine,
Academician Glushkov ave., 40, Kyiv, 03187, Ukraine
vxml12@gmail.com

Dmytro Fedoryn, researcher, International Research and Training Center
for Information Technologies and Systems of the NAS and MES of Ukraine,
Academician Glushkov ave., 40, Kyiv, 03187, Ukraine
dmytro.fedoryn@gmail.com

Oleksandr Yukhymentko, researcher, International Research and Training Center
for Information Technologies and Systems of the NAS and MES of Ukraine,
Academician Glushkov ave., 40, Kyiv, 03187, Ukraine
enomaj@gmail.com

Valentyna Robeiko, researcher, International Research and Training Center
for Information Technologies and Systems of the NAS and MES of Ukraine,
Academician Glushkov ave., 40, Kyiv, 03187, Ukraine
valya.robeiko@gmail.com

AUTOMATIC RECOGNITION, UNDERSTANDING AND SYNTHESIS OF SPEECH SIGNALS IN UKRAINE

Introduction. Speech is the most convenient, habitual, accessible and fast mean of communication between people and, therefore, is the most suitable for communication between human beings and machines. This makes topical the capability to develop automatic speech recognition and synthesis systems for the national science, technology and culture.

Purpose. The purpose is to analyze the state and outline the main ways of solving the problems of automatic recognition, understanding and synthesis for Ukrainian speech and spoken translation from Ukrainian Sign Language to Ukrainian language.

Methods. Modeling the spoken intellectual human activity using the analysis-by-synthesis approach accomplished with the experimental research and approbation in real application conditions.

Results. Methods and algorithms proposed and adapted to the specific hardware/software platforms allow the speech information systems developing meeting the growing expectations of potential users. The described contemporary spoken information systems demonstrate more generalization and less sensitivity to speaker and domain during analysis and high naturalness of synthesized speech signal. Due to these achievements, the processes of spoken information input and retrieval can be partially or fully automated, particularly, for Ukrainian.

Conclusion. For decades, methods and algorithms based on Generative Model are shown their productivity for speech technologies and systems that makes them widely applicable nowadays. The internationally recognized Ukrainian research school benefits from its history and traditions, demonstrates steady development and readiness to solve prospective problems related to multilingual, multimodal and acoustically adverse environments.

Keywords**:** *speech, speech signal, analysis, recognition, understanding, synthesis.*

Т.К. Винцюк, д-р техн. наук, профессор, зав. отделом,
Международный научно-учебный центр информационных технологий и систем
НАН Украины и МОН Украины, просп. Академика Глушкова, 40, Киев 03187, Украина

Н.Н Сажок, канд. техн. наук, зав. отделом,
Международный научно-учебный центр информационных технологий и систем
НАН Украины и МОН Украины, просп. Академика Глушкова, 40, Киев 03187, Украина,
sazhok@gmail.com

Р.А. Селюх, мл. научн. сотруд.,
Международный научно-учебный центр информационных технологий и систем
НАН Украины и МОН Украины, просп. Академика Глушкова, 40, Киев 03187, Украина,
vxml12@gmail.com

Д.Я. Федорин, мл. научн. сотруд.,
Международный научно-учебный центр информационных технологий и систем
НАН Украины и МОН Украины, просп. Академика Глушкова, 40, Киев 03187, Украина,
dmytro.fedoryn@gmail.com

А.А. Юхименко, мл. научн. сотруд.,
Международный научно-учебный центр информационных технологий и систем
НАН Украины и МОН Украины, просп. Академика Глушкова, 40, Киев 03187, Украина,
enomaj@gmail.com

В.В. Робейко, научн. сотруд.,
Международный научно-учебный центр информационных технологий и систем
НАН Украины и МОН Украины, просп. Академика Глушкова, 40, Киев 03187, Украина,
valya.robeiko@gmail.com

АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ, ПОНИМАНИЕ И СИНТЕЗ РЕЧИ В УКРАИНЕ

Вступление. Речь является наиболее удобным, привычным, доступным и быстрым средством общения между людьми и, следовательно, наиболее подходящим для общения между человеком и машиной. В этом состоит актуальность разработки автоматических систем распознавания и синтеза речи для национальной науки, техники и культуры.

Цель. Цель данной статьи — проанализировать состояние и наметить основные пути решения проблем автоматического распознавания, понимания и синтеза украинской речи и устного перевода с украинского и на украинский языки.

Методы. Моделирование разговорной интеллектуальной деятельности человека с использованием подхода «анализ через синтезом» с экспериментальными исследованиями и апробацией в реальных условиях применения.

Результаты. Методы и алгоритмы, предложенные и адаптированные к конкретным аппаратным/программным платформам, позволили разработать речевые информационные системы, отвечающие растущим ожиданиям потенциальных пользователей. Описанные современные речевые информационные системы демонстрируют большее обобщение и меньшую чувствительность к диктору и предметной области при анализе и высокую естественность синтезированного речевого сигнала. Благодаря этим достижениям процессы ввода и поиска устной информации могут быть частично или полностью автоматизированы, в частности для украинского языка.

Заключение. На протяжении десятилетий методы и алгоритмы, основанные на Генеративной модели, показали свою производительность для речевых технологий и систем, что сделало их широко применимыми в наши дни. Всемирно признанная украинская научная школа черпая энергию из своей истории и традиций, демонстрирует устойчивое развитие и готовность решать будущие задачи, возникающих в связи с многоязычнием, мультимодальностью и помехоустойчивостью.

***Ключевые слова**: речь, речевой сигнал, анализ, распознавание, понимание, синтез.*