ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ТА КОМП'ЮТЕРНА ТЕХНІКА

УДК 681.327.12

В. В. Ковтун[1]

ОЦІНЮВАННЯ ОСНОВНОГО ТОНУ У АВТОМАТИЗОВАНІЙ СИСТЕМІ РОЗПІЗНАВАННЯ МОВЦЯ КРИТИЧНОГО ЗАСТОСУВАННЯ

[1]Вінницький національний технічний університет

Запропоновано метод оцінювання трендів основного тону, який, на відміну від існуючих, використовує оптимізовану із застосуванням дерева переходів факторіальну приховану Марковську модель для формування трендів основного тону узагальнюючи при цьому інформацію від детекторів станів основного тону, на основі глибокої та рекурентної нейромереж, що дозволило спрогнозувати оцінки станів основного тону, використовуючи довготривалу інформацію з пакетів фреймів мовного сигналу, описати часову динаміку основного тону та зменшити вплив шумів у мовному сигналі на якість оцінок основного тону. Створено методи оцінювання станів основного тону на основі глибокої та рекурентної нейромереж та метод оцінювання трендів основного тону на основі факторіальної прихованої Марковської моделі (ФПММ). Проведено дослідження для оптимізації параметрів запропонованих методів для використання у складі автоматизованої системи розпізнавання мовця критичного застосування (АСРМКЗ). Зокрема, результати досліджень дозволяють рекомендувати нормовані за потужністю кепстральні ознаки як базові для оцінювання основного тону запропонованими методами, застосовувати в роботі методів пакети фреймів тривалістю 10 фреймів, будувати описані у методах нейромережі, використовуючи на прихованих шарах 1024 нейрони та використовувати 68 станів для опису основного тону. Результати проведених досліджень залежності якості розпізнавання мовців АСРМКЗ від рівня відношення сигнал/шум (ВСШ) у вхідному мовному матеріалі та оцінками основного тону, отриманими в результаті роботи створених методів, параметри яких оптимізовано з урахуванням результатів проведених досліджень, показали, що для всіх рівнів ВСШ найточніші оцінки основного тону дає ФПММ-метод, забезпечуючи імовірність правильного розпізнавання мовців АСРМКЗ на рівні 96…99 % для вибраної тестувальної вибірки.

Ключові слова: автоматизована система розпізнавання мовців критичного застосування, основний тон, глибока нейромережа, рекурентна нейромережа, факторіальна прихована Марковська модель.

Вступ

Використовуючи засоби цифрової обробки інформації, мовний сигнал, генерований людиною, розглядають як послідовність електричних коливань, отримуваних від акустично-електричного перетворювача під впливом акустичних мовних коливань на його вхід. Таке представлення мовного сигналу є дискретним і містить як інформаційну складову, описувану скінченною множиною можливих звуків та їх комбінацій, так і корельовану з нею контекстну складову, тобто інформацію про фізичний, емоційний стан людини, акустичний стан її оточення і, зокрема, інформацію про індивідуальні особливості процесу мовотворення, яку можна використовувати для розпізнавання особи мовця.

Мова людини складається з трьох чітко виражених за джерелом породженням груп сигналів-звуків: вокалізовані, невокалізовані та вибухові. Вокалізовані звуки мови (голосні, дзвінкі приголосні) утворюються у наслідок перетворення безперервного повітряного потоку з легень при його проходженні крізь низки у імпульсний потік, який далі збуджує мовний тракт людини. Квазіперіодичні параметри імпульсного потоку визначаються індивідуальними параметрами коливань низок і можуть бути виділені із запису вокалізованого фрагменту мовного сигналу. Невокалізовані звуки (глухі приголосні) являють собою широкосмуговий шум із неперервним спектром, утворюваний

вихровим повітряним потоком, який проходить певним звужуванням мовного тракту людини. Ці звуки зазвичай передують перерозподілу енергії при закритому мовному тракті, коли повітря спочатку стискається, а потім різко звільняється, утворюючи вибухові або фрикативні звуки.

Імпульси повітряного потоку, створювані коливаннями низок під час формування вокалізованих звуків з достатньою точністю вважають періодичними. Цей період T_0 називають періодом основного тону, а обернену до нього величину f_0 — частотою основного тону. Якщо низки тонкі або напружені, то період їх коливань буде коротким, а частота, відповідно, високою, і навпаки — якщо низки грубі або ненапружені, то період їх коливань буде довгий, а частота — низькою. Зазвичай перший варіант притаманний дітям і жінкам, а другий — чоловікам. Втім дослідження [1] показують, що частота основного тону для всіх голосів лежить в діапазоні 70…400 Гц, змінюючись у часі протягом виголошення мовного повідомлення і має свій індивідуальний діапазон та динаміку зміни для кожної людини, який зазвичай є трохи більшим октави. Вищенаведені обставини зумовлюють актуальність параметрів основного тону в процесі розв'язання задачі розпізнавання мовця.

Втім, застосування основного тону як інформативної для розпізнавання особи мовця ознаки у автоматизованій системі розпізнавання мовців критичного засновування (АСРМКЗ) вимагає забезпечення її стійкості до визначених типів шумів та можливості адаптації методу її виділення до інших типів шумів. Відомі методи оцінювання основного тону досліджують або гармонійні структури мовного сигналу у частотному просторі, або періодичні явища у часовому просторі, або поєднують ці підходи, пропустивши мовний сигнал крізь гребінку фільтрів, які перекривають частотний діапазон, де зустрічається основний тон. Більшість методів виділення основного тону на основі аналізу мовного сигналу у частотному просторі оцінюють спектр мовного сигналу, припускаючи, що піки енергії спектру відповідають гармонікам основного тону [2], [3]. Наприклад, метод SAFE [4] дає імовірнісну оцінку наявності гармонік основного тону на інтервалах мовного сигналу, де спостерігаються піки енергії спектра рівня відношення сигнал/шум. У методі PEFAC [5] використовується нелінійна амплітудна компресія для ослаблення вузькосмугових шумів та аналізуються енергетичні піки отриманого після фільтрації спектру щодо наявності гармонік основного тону. Інший підхід до виявлення основного тону базується на факті періодичності його гармонік під час аналізу мовного сигналу у часовому просторі. Так метод RAPT [6] описує мовний сигнал у часовому просторі за допомогою нормалізованої автокореляційної функції та аналізує її піки на предмет відповідності гармонікам основного тону, а метод YIN [7] використовує інформацію від квадратно різницевої функції на основі нормалізованої автокореляційної функції з тією ж метою. Розширенням цього підходу є паралельний аналіз у часовій області сигналів з виходів гребінки фільтрів, на вхід якої подається мовний сигнал. У методі [8] період основного тону описується статистично на основі кореляції положення піків сигналів з виходів гребінки фільтрів із використанням математичного апарату прихованих Марковських моделей для генерації остаточного рішення, а метод [9] доповнює ідею попереднього методу врахуванням інформації про зашумленість частотних смуг гребінки фільтрів шляхом введення відповідних вагових коефіцієнтів, які враховуються при побудові підсумкової корелограми. Окремо варто загадати про метод [10], де рішення щодо виявлення періоду основного тону приймається персептроном, який аналізує нормалізовані автокореляційні функції, що відповідають мовним сигналам. Слід зазначити, всім описаним методам бракує точності виділення основного тону. Джерелом похибки є спотворення представлення мовного сигналу в наслідок присутності шумів, особливо низькочастотних, які накладаються на частотний діапазон де спостерігається основний тон, та недоліки процедури ідентифікації гармонік основного тону пов'язані з недосконалістю використовуваних методів прийняття рішень.

Постановка задачі дослідження

Отже, *метою статті* є врахування інформації про основний тон при розпізнаванні мовців АСРМКЗ, специфіка якої полягає у забезпеченні прогнозованості результату розпізнавання у визначених умовах експлуатації системи, які в основному визначаються рівнем відношення сигнал/шум. Аналіз актуальних методів оцінювання основного тону показав їх суттєву шумозалежність та низьку адаптивність до умов експлуатації та недосконалість процедури розпізнавання гармонік основного тону, зумовлену недосконалістю використовуваних методів прийняття рішень. Автори пропонують використати глибоку нейромережу для оцінювання станів основного тону, особливістю якої є суттєва гнучкість у процесі прийняття рішень та можливість узагальнювати

великі обсяги вхідних даних. Для врахування динаміки основного тону пропонується використати рекурентну нейромережу, якій за рахунок зворотних зв'язків властивий ефект пам'яті щодо попередньо прийнятих рішень, що дозволяє встановити тенденцію динаміки основного тону, яка менше залежатиме від рівня шумів у аналізованому фреймі. Нарешті, для узагальнення інформації від глибокої та рекурентної нейромереж пропонується використати факторіальну приховану Марковську модель, оптимізовану для імовірнісної оцінки станів основного тону, у якій враховуються зважені оцінки станів основного тону, отримані від глибокої та рекурентної нейромережі, а вагові коефіцієнти оцінок встановлюються на етапі навчання моделі в залежності від того, яку частотну смугу описує коефіцієнт стану, та який рівень шумів у аналізованому мовному сигналі. Враховуючи чуттєву ресурсоємність використовуваних елементів системи оцінювання основного тону доцільним є оптимізація їх параметрів.

Нейромережеві моделі оцінювання основного тону

Основний тон — це індивідуальна характеристика голосового джерела людини, яка проявляється на вокалізованих інтервалах мови у частотному діапазоні від 60 Гц до 400 Гц. Розіб'ємо цей частотний діапазон на N частотних смуг, які описуватимуться змінними стану основного тону $s^{(1)},\ldots,s^{(N)}$. Орієнтуючись на представлення октави у логарифмічній шкалі 24 смугами, згаданий частотний діапазон перекриємо 67 смугами, тобто частота квантування n-ї смуги дорівнюватиме $60\cdot 2^{\frac{n-1}{24}}$ Гц. Також використаємо стан $s^{(0)}$ для детектування невокалізованих мовних інтервалів та пауз. Змінна стану $s_t^{(i)}$ набуде значення, рівне 1, якщо у момент часу *t* у *i*-й частотній смузі детектується основний тон, або 0 у протилежному випадку. В результаті підсумкова кількість змінних стану N дорівнюватиме 68.

Теорія цифрової обробки сигналів передбачає розбиття множини детермінованих значень мовного сигналу S_{in} на блоки-фрейми тривалістю t, які формують множину фреймів

$$X=\left[x_0,x_1,\ldots,x_{I-1}\right], \tag{1}$$

де I — кількість фреймів, на яку розбито фонограму мовного сигналу. Зазначимо, що основний тон може бути присутнім більш ніж у одному фреймі і інформація від інших фреймів може впливати як на точність оцінювання основного тону у поточному фреймі, так і використовуватися безпосередньо для оцінювання динаміки основного тону. Отже, необхідно одночасно аналізувати p сусідніх фреймів мовного сигналу. Сформулюємо концепцію представлення інформації для виділення основного тону у вигляді пакету послідовних фреймів $x^{(i)}=\left[x_{i-p}^T,\ldots,x_{i-1}^T,x_i^T,x_{i+1}^T,\ldots,x_{i+p}^T\right]$, $i=0,1,\ldots,M-1$, де M — кількість пакетів фреймів, а параметр p визначає кількість фреймів у пакеті. В результаті на основі множини фреймів (1) отримаємо множину пакетів фреймів

$$X_{mb}=\left[x_{\mu 0},x_{\mu 1},\ldots,x_{\mu(M-1)}\right], \tag{2}$$

де $\left\{\mu 0,\mu 1,\ldots,\mu(M-1)\right\}\subset\left\{0,1,\ldots,I-1\right\}$ — вектори ознак, які виділяються з відповідних фреймів.

Представимо (2) у матричному вигляді для всієї множини фреймів (1)

$$X_{mb}=\begin{bmatrix} x_{\mu 0-p} & \cdots & x_{\mu(M-1)-p} \\ \vdots & & \vdots \\ x_{\mu 0} & \cdots & x_{\mu(M-1)} \\ \vdots & & \vdots \\ x_{\mu 0+p} & \cdots & x_{\mu(M-1)+p} \end{bmatrix}, \tag{3}$$

де номера фреймів, які при формуванні пакетів виявляють меншими нуля, вважаємо рівними нулю, так само як номера фреймів, які перевищують $I-1$ вважаємо рівними $I-1$. Матриця (3) визначає структуру вхідного компонента навчальної вибірки автоматизованої системи оцінювання основного тону.

В загальному випадку оцінювання основного тону можна розглядати як процес виявлення залежності між значеннями інформативних ознак, виділених з мовного сигналу, та значеннями змінних стану основного тону у будь-який момент часу звучання фонограми. Таке формулювання задачі оцінювання основного тону зумовлює вибір глибокої нейромережі прямого поширення сигналу (Deep Neural Network, DNN) [11] як базового елемента прийняття рішень у представлених далі дослідженнях. Щоб дати імовірнісну оцінку процесу оцінювання основного тону на основі DNN використаємо метод перехресної ентропії

$$L(y,x,\Theta)=-\sum_{n=0}^{N} y_n \ln f_n(x), \tag{4}$$

де $y=(y_0,\ldots,y_N)^T$ — бажаний вихідний вектор класифікатора, x — множина векторів інформативних ознак, виділених із вхідного пакету фреймів, представленого у вигляді (2), $f_n(\cdot)$ — фактичне значення на виході n-го нейрону вихідного шару нейромережі, а змінна Θ описує параметри, значення яких встановлюються у процесі навчання класифікатора. У нейронах прихованих шарів використаємо сигмоїдні функції активації ϕ, а у нейронах вихідного шару використаємо SOFTMAX функції активації ψ для отримання імовірнісних вихідних значень. Для вхідного сигналу X_{mb} значення на виході l-го шару Θ^l опишемо як

$$\Theta^l = f^l\left(W^l\Phi^l\right)=\left[\theta_0^l,\theta_1^l,\ldots,\theta_{M-1}^l\right]=\begin{bmatrix}\theta_{10}^l & \theta_{11}^l & \ldots & \theta_{1(M-1)}^l\\ \theta_{20}^l & \theta_{21}^l & \ldots & \theta_{2(M-1)}^l\\ \vdots & \vdots & \vdots & \vdots\\ \theta_{q_l0}^l & \theta_{q_l1}^l & \ldots & \theta_{q_l(M-1)}^l\end{bmatrix}, \tag{5}$$

де $W^l=\begin{bmatrix}w_{10}^l & w_{11}^l & \ldots & w_{1q_{l-1}}^l\\ w_{20}^l & w_{21}^l & \ldots & w_{2q_{l-1}}^l\\ \vdots & \vdots & \vdots & \vdots\\ w_{q_l0}^l & w_{q_l1}^l & \ldots & w_{q_lq_{l-1}}^l\end{bmatrix}$ є матрицею міжнейронних зв'язків, $\Phi^l=\begin{bmatrix}1 & 1 & \ldots & 1\\ \theta_0^{l-1} & \theta_1^{l-1} & \ldots & \theta_{M-1}^{l-1}\end{bmatrix}$ є матрицею вхідних впливів, $\theta_\lambda^0=x^{\mu\lambda}$, q_l — кількість нейронів на шарі l за виключенням нейрону зміщення, w_{jk}^l — вага зв'язку між k-м нейроном $l-1$ шару та j-м нейроном l шару, f^l — функція активації нейронів l-го шару.

З урахуванням описаних вище структурних особливостей DNN для оцінювання основного тону процес формування вихідного сигналу нейромережі $y_{DNN}(t)$ у момент часу t при подачі на її вхід вектора параметрів x представимо так:

$$\begin{aligned} y_{DNN}(t)&=\psi\left(\left(W^{0,j}\right)^T h_j(t)\right);\\ h_j(t)&=\phi\left(v_j(t)\right);\\ v_j(t)&=\left(W^{j,j-1}\right)^T h_{j-1}(t);\\ h_1(t)&=\phi\left(\left(W^{1,i}\right)^T x_i(t)\right), \end{aligned} \tag{6}$$

де $W^{j,j-1}$ — матриця ваг міжнейронних зв'язків між $j-1$-м та j-м прихованими шарами, h_j — вектор вихідних значень нейронів j-го шару, $W^{0,j}$ — матриця ваг міжнейронних зв'язків між останнім прихованим шаром DNN та вихідним шаром і $W^{1,3}$ — матриця ваг міжнейронних зв'язків між вхідним шаром DNN та першим прихованим шаром.

Навчальна вибірка DNN для оцінювання основного тону містить множину векторів ознак x_t для фреймів *t*, на які у часовому просторі розбито фонограму мовного сигналу, у вигляді (3) та множину бажаних векторів станів основного тону s_t розмірністю $N+1$ елемент кожен. Для навчання DNN використовуватимемо міні-пакетний (mini-batch) стохастичний метод градієнтного спуску у якому фактична оцінка кожного з міні-пакетів визначалася в результаті узагальнення множин елементів навчальної вибірки відповідно до (4). На виході навченої DNN отримаємо апостеріорні імовірності для кожного *i*-го стану основного тону $P\left(s_t^{(i)}\middle|x_t\right)$, подавши на вхід DNN множину векторів інформативних ознак, яка представляє пакет фреймів, до якого увійшли фрейми $x_1,\ldots,x_t$. Для оцінювання станів основного тону на основі отриманих на виході нейромережі розподілів імовірності $P\left(s_t^{(i)}\middle|x_1,\ldots,x_t\right)$ застосуємо алгоритм Вітербі. Цей алгоритм на основі даних про правдоподібність і перехідну імовірність обчислює ціну генерації оптимальної послідовності. Правдоподібність $p\left(x_t\middle|s_t^{(i)}\right)$ кожного фрейму пропорційна апостеріорній імовірності $P\left(x_t\middle|s_t^{(i)}\right)$, отриманій на виході нейромережі, і обернено пропорційна імовірності $P\left(s^{(i)}\right)$

$$p\left(x_t\middle|s_t^{(i)}\right)\propto\frac{P\left(s_t^{(i)}\middle|x_t\right)}{P\left(s^{(i)}\right)}, \tag{7}$$

де апріорна імовірність $P\left(s^i\right)$ і перехідна матриця обчислюються на основі навчальних даних. Враховуючи, що в навчанні нейромережі використовувався мовний матеріал, у якому окрім вокалізованих інтервалів були присутні невокалізовані інтервали та паузи, тривалість яких зазвичай суттєво більша, апріорна імовірність стану відсутності основного тону $P\left(s^{(0)}\right)$ виявляється вищою за імовірності решти станів основного тону, що може призвести до помилки округлення під час оцінювання основного тону. Тому пропонується ввести ваговий мультиплікативний параметр $\alpha\in(0,1]$ для змінної імовірності стану відсутності основного тону $P\left(s^{(0)}\right)$, значення якого встановлюється на основі співвідношення тривалості вокалізованих інтервалів до загальної тривалості мовного сигналу. Отримані в результаті роботи алгоритму Вітербі значення станів основного тону $s_t^{(i)}$ для фреймів мовного сигналу конвертуємо у частоти $\hat{f}_0$, відповідно до попередньо сформованої частотної сітки, використовуючи правило

$$\hat{f}_0=\arg\max_{\hat{f}_0'}P\left(\hat{f}_0'\middle|\gamma_{iv},p\left(x_t\middle|s_t^{(i)}\right),X\right), \tag{8}$$

$\forall i,v=0,1,\ldots,N-1,\ \forall t=0,1,\ldots,I-1,\ X=\left[x^0,x^1,\ldots,x^{I-1}\right]$, де γ_{iv} — імовірність переходу зі стану $s^{(i)}$ у стан $s^{(v)}$.

Відмітимо, що описаний виразом (2) спосіб подачі вхідної інформації у DNN поряд із простотою реалізації та можливістю оцінювання динаміки основного тону має суттєвий недолік — необхідно одночасно аналізувати p сусідніх фреймів мовного сигналу. Це означає, що на вхідному шарі DNN має бути $p\cdot N_{in}$ нейронів, де N_{in} — розмірність вектора ознак, який характеризує *t*-й фрейм мовного сигналу. Отже, чим більше фреймів аналізується, тим краще оцінюється динаміка основного тону, але, відповідно, стрімко збільшуємо розмірність вхідного шару DNN, чим зумовлюється ускладнення процесу її навчання аж до його повної неможливості.

Спробуємо позбавитися виявленого недоліку, використавши замість DNN рекурентну нейромережу (Recurrent Neural Network, RNN) [12]. Цей вид нейромереж відрізняється від класичних багатошарових нейромереж прямого поширення інформації наявністю зворотних зв'язків, які дозволяють зберігати попередньо отриману інформацію і аналізувати пов'язані порції вхідних даних, що дозволяє застосовуватися ці нейромережі для оцінювання просодичних характеристик мовного

сигналу та інформації про динаміку основного тону на множині фреймів мовного сигналу. Архітектурні особливості RNN передбачають відмінний від запропонованого у вигляді (2), (3) спосіб подачі вхідної інформації, який можна описати так:

$$\begin{aligned}
&X_{mb}^{0} = \left[x_{\mu 0-p}, \ldots, x_{\mu(M-1)-p} \right];\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&X_{mb}^{p-1} = \left[x_{\mu 0-1}, \ldots, x_{\mu(M-1)-1} \right];\\
&X_{mb}^{p} = \left[x_{\mu 0}, \ldots, x_{\mu(M-1)} \right];\\
&X_{mb}^{p+1} = \left[x_{\mu 0+1}, \ldots, x_{\mu(M-1)+1} \right];\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&X_{mb}^{2p} = \left[x_{\mu 0+p}, \ldots, x_{\mu(M-1)+p} \right],
\end{aligned} \tag{9}$$

де параметр p, як і попередньо, визначає кількість фреймів у пакеті аналізу, яка для множини (8) дорівнює $2p+1$.

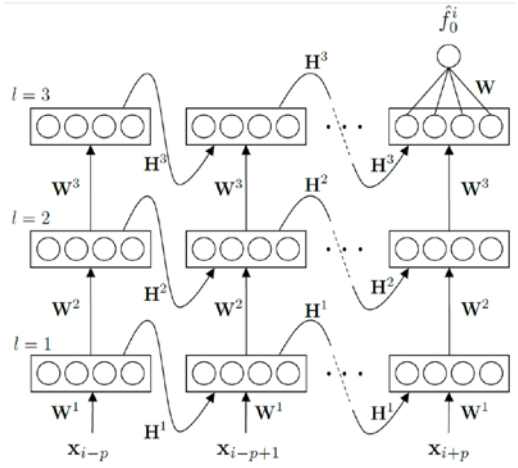

Рис. 1. Архітектура RNN для оцінювання основного тону

Архітектура створеної RNN, показана на рис. 1, передбачає, що сигнал на виході шару l нейромережі для x_i фрейму з пакету фреймів n сформується так:

$$\begin{aligned}
&\theta_n^l = f\left(W^l \phi_n^l + H^l \phi_{n-1}^{l+1} \right);\\
&\phi_n^l = \left[1, \left(\theta_n^{l-1} \right)^T \right]^T;\\
&\theta_n^0 = \left[1, \left(x_{i-p+n} \right)^T \right]^T,
\end{aligned} \tag{10}$$

де W^l — матриця ваг міжнейронних зв'язків між $l-1$ та l шарами нейромережі, а H^l — матриця ваг зворотних міжнейронних зв'язків, які подають вихідні сигнали нейронів l-го шару на їх вхід. Структура матриць W^l та H^l аналогічна описаній у (5).

Вихідний сигнал нейромережі $y_{RNN} = \left(y_1, \ldots, y_n \right)^T$ у момент часу t опишемо так:

$$\begin{aligned}
&y_{RNN}(t) = \psi\left(\left(W^{0,j} \right)^T h_j(t) \right);\\
&h_j(t) = \phi\left(v_j(t) \right);\\
&v_j(t) = \left(W^{j,j-1} \right)^T h_{j-1}(t) + \left(H^{j,j} \right)^T h_j(t-1);\\
&h_1(t) = \phi\left(\left(W^{1,i} \right)^T x_i(t) \right),
\end{aligned} \tag{11}$$

де $W^{j,j-1}$ — матриця ваг міжнейронних зв'язків між $j-1$-м та j-м прихованими шарами, h_j — вектор вихідних значень нейронів j-го шару, $H^{j,j}$ — матриця ваг зворотних зв'язків у межах j-го шару (якщо нейрони мають зворотні зв'язки самі із собою, то матриця $H^{j,j}$ є діагональною, а якщо у шарі немає зворотних зв'язків, то $H^{j,j} = 0$), $W^{0,j}$ — матриця ваг міжнейронних зв'язків між останнім прихованим шаром RNN та вихідним шаром і $W^{1,3}$ — матриця ваг міжнейронрих зв'язків між вхідним шаром RNN та першим прихованим шаром.

Для навчання RNN застосуємо метод зворотного поширення помилки у часі (Backpropagation Through Time, BPTT). Вихідні шари, створених для оцінювання основного тону DNN та RNN ідентичні, що уніфікує процес інтерпретації вихідних сигналів RNN із вищеописаним для DNN у вигляді (7), (8).

Метод оцінювання трендів основного тону

Після отримання за допомогою DNN та RNN оцінок станів основного тону для відповідних пакетів фреймів виникає потреба у інтеграції отриманих імовірнісних результатів у вигляді найімовірніших трендів основного тону. Враховуючи, що створюваний метод оцінювання основного тону передбачається застосовувати у АСРМКЗ, оцінюючи тренди основного тону варто передбачити можливість узагальнення інформації від кількох незалежних процедур оцінювання основного тону, наприклад, від DNN та RNN на основі однакових вхідних даних. Таке узагальнення підвищить надійність оцінки основного тону, що є ключовим параметром методів, які застосовуються у АСРМКЗ. Враховуючи необхідність імовірнісного оцінювання інформації від кількох незалежних її джерел, пропонується застосувати математичний апарат факторіальних прихованих Марковських моделей (Factorial Hidden Markov Model, FHMM) [13] для її аналізу, графічну інтерпретацію якого для двох Марковських мереж, які описують процеси оцінювання станів основного тону DNN та RNN відповідно, наведено на рис. 2.

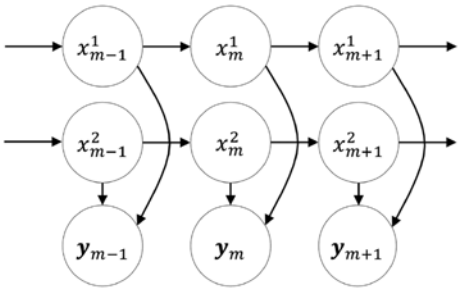

Рис. 2. Графічна інтерпретація FHMM для трендів основного тону з узагальненням інформації від двох Марковських мереж

Нехай приховані змінні $\left(x_m^1, x_m^2\right)$ описують стани основного тону від двох джерел інформації для вектора інформативних ознак, представленому змінною спостереження y_m. Типовим для FHMM є припущення, що значення y_m не залежить від усіх x_m^1 та x_m^2. Узагальнимо множини прихованих змінних та змінних спостереження для всіх аналізованих фреймів мовного сигналу I у вигляді множин $X = \bigcup_{m=1}^{I} \left\{x_m^1, x_m^2\right\}$ та $Y = \bigcup_{m=1}^{I} \left\{y_m\right\}$, відповідно. Тоді підсумкову імовірнісну оцінку тренду основного тону FHMM представимо у вигляді

$$p(X,Y) = p\left(x_1^1\right) p\left(x_1^2\right) p\left(y_1 \middle| x_1^1, x_1^2\right) \prod_{m=2}^{I} p\left(x_m^1 \middle| x_{m-1}^1\right) p\left(x_m^2 \middle| x_{m-1}^2\right) p\left(y_m \middle| x_m^1, x_m^2\right). \quad (12)$$

Апріорні імовірності та транзитні матриці прихованих змінних отримаємо за результатами оцінювання станів основного тону DNN та RNN окремо. Для запобігання отримування нульової імовірності застосуємо згладжування Лапласа під час обчислення кожної змінної спостереження. Використовуючи оцінки апостеріорних імовірностей, обчислимо імовірності викидів, застосувавши правило Байєса:

$$p\left(y_m \middle| x_m^1, x_m^2\right) = \frac{p\left(x_m^1, x_m^2 \middle| y_m\right) p\left(y_m\right)}{p\left(x_m^1\right) p\left(x_m^2\right)}, \quad (13)$$

де імовірність $p\left(y_m\right)$ є сталою для всіх векторів ознак.

Використовуючи обчислені на основі (12) та (13) імовірності, визначимо найімовірнішу послідовність станів (тренд) основного тону, застосувавши алгоритм дерева переходів (Junction Tree Algorithm, JTA) [14], згідно з яким спочатку направлену графічну модель FHMM з рис. 2 перетворимо на ненаправлену, на основі вузлів якої побудуємо дерево переходів для визначення найімовірнішого поширення. Обчислювальна складність для створеного дерева оцінюється у $\left(2 \times 68^3 \times I\right)$ операцій. Враховуючи, що основний тон є сигналом, який породжується голосовим джерелом людини, отриманий в результаті тренд основного тону варто згладити для усунення різких перепадів рівня за допомогою, наприклад, вікна усереднення з тривалістю у три відліки.

Експерименти та аналіз результатів

Для встановлення якості запропонованих методів оцінювання основного тону та визначення інформативності цієї ознаки у складі АСРМКЗ застосовувався матеріал мовного корпусу TIMIT [15] (в якості джерела фонограм із записами мовних сигналів) та матеріал бази NOISEX-92 [16] (в якості джерела записів шумів). Навчальну вибірку сформовано на основі мовного матеріалу із 250 висловлювань кожного з 50 мовців-чоловіків та 50-мовців-жінок із мовного корпусу TIMIT, до яких підмішувалися шуми типів «Babble», «F16», «Factory1», «Leopard» із бази NOISEX-92, утворюючи множини фонограм з рівнем відношення сигнал/шум (Signal-to-Noise Ratio, SNR) [-15, -10, -5, 0, 5, 10, 15] дБ. Першу тестувальну вибірку сформовано на основі мовного матеріалу з 100 висловлювань із мовного корпусу TIMIT від кожного з тих самих 50 мовців-чоловіків та 50 мовців-жінок, мовний матеріал яких увійшов до навчальної бази, до яких підмішувалися шуми типів «Machinegun», «Pink», «Volvo», «White» із бази NOISEX-92, утворюючи множини фонограм із SNR = [-15, -10, -5, 0, 5, 10, 15] дБ. Під час формування другої тестувальної вибірки взято по 100 висловлювань із мовного корпусу TIMIT від тих самих 40 мовців-чоловіків та 40-мовців-жінок, мовний матеріал яких увійшов до навчальної бази, та від 10 мовців-чоловіків та 10-мовців-жінок, яких не було у начальній вибірці. Далі до фонограм другої тестувальної вибірки підмішувалися шуми типів «Babble», «F16», «Factory1», «Leopard» із бази NOISEX-92 утворюючи множини фонограм із SNR = [-15, -10, -5, 0, 5, 10, 15] дБ. Під час формування третьої тестувальної вибірки мовний матеріал першої тестувальної вибірки об'єднування для кожного мовця у єдиний файл зі всіма SNR, із якого випадковим чином вибиралося 500 фреймів, які об'єднувалися у підсумкову фонограму з динамічним SNR. В результаті перша тестувальна вибірка дозволяє визначити якість оцінювання основного тону в умовах впливу невідомих шумів, друга тестувальна вибірка дозволяє оцінити інформативність основного тону для задачі розпізнавання мовців АСРМКЗ за умови спроби ідентифікації мовців, які не мають такого права, а третя тестувальна вибірка дозволяє встановити якість оцінювання основного тону в умовах присутності у мовному сигналі шумів з динамічним SNR.

В якості критерію для встановлення якості запропонованих методів оцінювання основного тону на основі DNN та RNN і узагальненого методу оцінювання тренду основного тону використаємо загальновизнану метрику [17], до складу якої входить груба помилка визначення основного тону (Gross Pitch Error, GPE) та точна помилка визначення основного тону (Fine Pitch Error, FPE). GPE-параметр оцінюється рівнянням

$$GPE = N_{GPE} N_v^{-1}, \tag{14}$$

де N_v — кількість вокалізованих фреймів у тестовому мовному сигналі, N_{GPE} — кількість вокалізованих фреймів у тестовому мовному сигналі, для яких різниця між визначеним оцінюваним методом значенням періоду основного тону $1/\hat{f}_0$ та визначеним експертом істинним періодом основного тону $1/f_0$ перевищує 0,625 мс. FPE-параметр оцінюється рівнянням $FPE = N_v - N_{GPE}$ і використовується для визначення FPE-математичного сподівання μ_{FPE} та FPE-дисперсії σ_{FPE}:

$$\mu_{FPE} = \frac{1}{FPE} \sum_{i=1}^{FPE} \Delta_i; \tag{15}$$

$$\sigma_{FPE} = \sqrt{\frac{1}{FPE} \sum_{i=1}^{FPE} \left(\Delta_i - \mu_{FPE}\right)^2}, \tag{16}$$

де i — номер фрейму у тестовому мовному сигналі, а параметр $\Delta_i = \left| \hat{f}_0^{(i)} - f_0^{(i)} \right|$ є абсолютною оцінкою точності детектування частоти основного тону.

Фонограма з висловлюванням розбивалася на фрейми тривалістю 25 мс з перекриванням у 5 мс, а потім, враховуючи специфіку записів мовного матеріалу у базі TIMIT, перші 400 та останні 200 фреймів видалялися, як такі, що не містили мовної інформації. Далі до даних у фреймі застосовувалося короткочасне перетворення Фур'є (Short-Time Fourier Transform, STFT) з 1024 точковим швидким перетворенням Фур'є для оцінювання щільності спектральної потужності у частотному та часовому просторах, яка використовувалася для виділення таких інформативних ознак як

Мел-частотні кепстральні коефіцієнти (Mel-Frequency Cepstral Coefficients, MFCC) [18], нормовані за потужністю кепстральні коефіцієнти (Power Normalized Cepstral Coefficients, PNCC) [19] та коефіцієнти спектрально-темпоральних рецептивних полів (Spectro-Temporal Receptive Field, STRF) [20], на основі яких формувалася конфігурація простору ознак для представлення фреймів мовного сигналу для подальшого нейромережевого аналізу. Дані із перших 513 частотних діапазонів STFT $(0 \le \omega \le \pi)$ кожного з фреймів використовуються як базові для міні-пакетного аналізу (Mini-Batch Analysis, MBA), який лежить в основі метода міні-пакетного градієнтного спуску (Mini-Batch Gradient Descent, MBGD) [31] для навчання DNN та RNN з параметрами, описаними у [21]. Зокрема, у структурі глибоких нейромереж формувалися п'ять прихованих шарів із 1024 (ReLU для DNN, tanh для RNN) нейронами на кожному і тривалість міні-пакету становила 200 фреймів. Для запобігання перенавчанню параметр випадкового виключення нейронів (Random Unit Dropout, RUD) дорівнював 50 %, а для пришвидшення навчання нейромереж застосовувався метод нормалізації вибірки (Batch Normalization) [21] з рівнем навчання 0,9.

Спочатку встановимо якість запропонованих методів для оцінювання основного тону на основі експериментів з даними першої та другої тестувальних вибірок, результати яких показано на рис. 3.

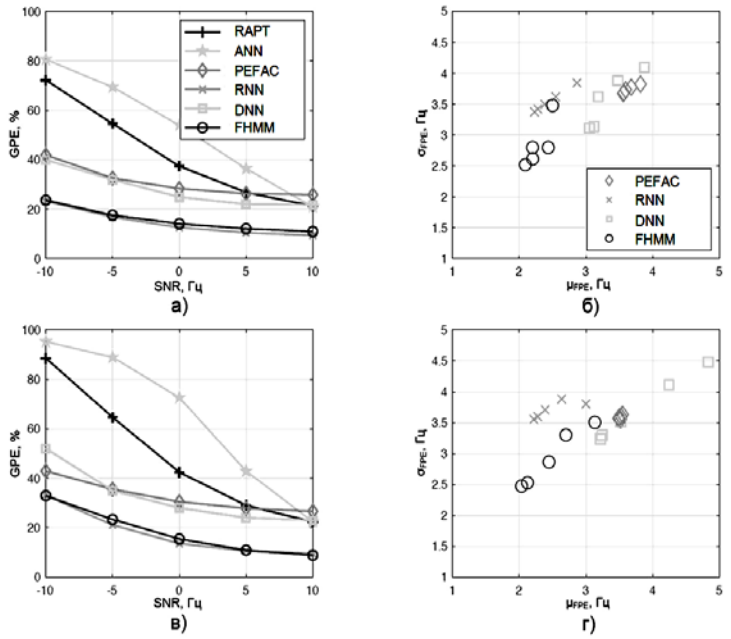

Рис. 3. Якісне оцінювання авторських методів детектування основного тону: а — у GPE/SNR -метриці за даними першої тестувальної вибірки; б — у μ_{FPE}/σ_{FPE} -метриці за даними першої тестувальної вибірки; в — у GPE/SNR -метриці за даними другої тестувальної вибірки; г — у μ_{FPE}/σ_{FPE} -метриці за даними другої тестувальної вибірки

В експериментах окрім авторських методів оцінювання основного тону на основі глибокої (DNN) та рекурентної (RNN) нейромереж та і узагальненого методу оцінювання тренду основного тону (FHMM) реалізовано два ефективних методи оцінювання основного тону на основі теорії цифрової обробки сигналів — RAPT і REFAC, а також авторський метод оцінювання основного тону, у якому замість глибокої нейромережі використано класичний тришаровий персептрон (ANN). На рис. 3а і 3б вищезгадані методи показують свою результативність у GPE/FPE - та μ_{FPE}/σ_{FPE} -метриках, відповідно, на основі даних першої тестувальної вибірки, у якій мовний матеріал навчальної вибірки піддавався впливу невідомих за характером і рівнем шумів. На рис. 3в і 3г вищезгадані методи показують свою результативність у GPE/FPE- та μ_{FPE}/σ_{FPE} -метриках, відповідно, на основі даних другої тестувальної вибірки, у якій необхідно оцінити основний тон мовців, яких не було у навчальній вибірці.

Наведені на рис. 3 результати експериментів показують, що найвищу результативність показав FHMM метод для обох тестувальних вибірок. Втім, як видно з рис. 3г найефективніші методи показують близькі за якістю результати під час аналізу невідомого за джерелом походження мовного матеріалу, що можна пояснити, зокрема, недостатньою інформативністю вибраної конфігурації простору ознак. Зазначимо, що узагальнювальних можливостей персептрону (ANN) не вистачило для ефективного прийняття рішень щодо оцінок основного тону. Гіпотеза щодо доцільності аналізу пакетів фреймів з метою виявлення довготривалих тенденцій основного тону підтвердилася наведеними на рис. 3б результатами, де RNN безпосередньо та з їхньою інтеграцією у межах FHMM-моделі показують явно кращі порівняно із іншими методами результати із чіткими границями кластерів, що говорить про стійкість цих методів до присутності у аналізованому мовному сигналі невідомих акустичних шумів, властивих першій тестувальній вибірці.

Перевагою заснованих на машинному навчанні методів (DNN, RNN, FHMM) є можливість їх гнучкої адаптації на параметричному рівні, що зумовлює доцільність відповідних досліджень,

результати яких наведено на рис. 4. FHMM-система, яка показала за результатами попереднього дослідження найкращі результати щодо оцінювання основного тону інтегрувалася у склад АСРМКЗ [22] та використовувалася для розпізнавання мовців за даними другої тестувальної вибірки за мовним матеріалом із SNR = 0 дБ та за даними третьої тестувальної вибірки, для мовного матеріалу якої властивий динамічний за рівнем та характером шум. Якісною характеристикою результатів АСРМКЗ є середня імовірність помилки розпізнавання $1-P_+$, обчислена як відношення кількості випадків помилкового розпізнавання (сплутування або пропуск невідомого мовця) до загальної кількості проведених експериментів.

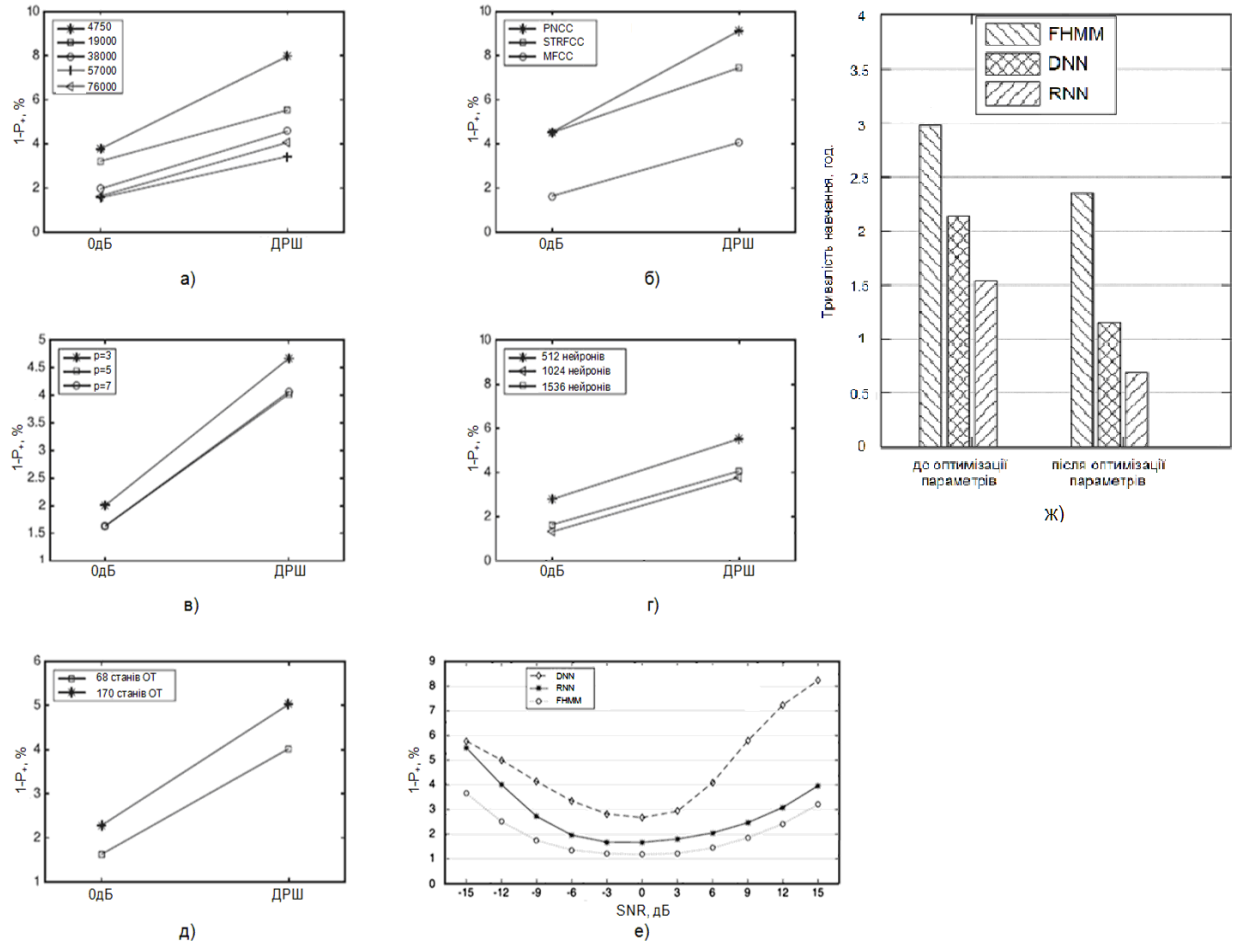

Рис. 3. Залежність імовірності помилки розпізнавання мовців АСРМКЗ $1-P_+$ від параметрів блоку оцінювання основного тону за умови відсутності шумів (0 дБ) / динамічного рівня шумів (ДРШ): а — в залежності від кількості лінгвістичних одиниць, які здатна ідентифікувати АСРМКЗ; б — в залежності від конфігурації простору ознак; в — в залежності від кількості фреймів у пакеті аналізу; г — в залежності від кількості нейронів на прихованих шарах; д — в залежності від масштабу сітки станів основного тону; е — в залежності від класифікатора для прийняття рішень щодо станів основного тону; ж — тривалість навчання класифікаторів до та після оптимізації їх параметрів

У використаній АСРМКЗ поряд з основним тоном для розпізнавання мовців використовується і інша просодична інформація, зокрема тривалість вимови сталих лінгвістичних одиниць — трифонів, які зустрічаються у парольних фразах. Для опису цієї інформації використовується GMM-HMM апарат [9]. На рис. 4а показано залежність якості розпізнавання мовців АСРМКЗ від виду тестувальних даних та кількості лінгвістичних одиниць, які розпізнає GMM-HMM складова АСРМКЗ. Виявляється, що для опису мовного матеріалу тестувальних вибірок достатньо ідентифікувати 57000 лінгвістичних одиниць. Ця інформація дозволяє суттєво спростити обчислювальну складність процесу навчання АСРМКЗ. На рис. 4б наведено залежність якості розпізнавання мовців АСРМКЗ від виду тестувальних даних та конфігурації простору базових ознак, узагальнюючи які методи оцінюють стани основного тону. Це дослідження є особливо важливим, адже результати на рис. 3г, отримані на основі MFCC, показали низьку інформативність такого представлення мовних сигналів у випадку опрацювання мовного матеріалу від невідомого мовця. У дослідженні приймали участь PNCC-ознаки, за своїм походженням стійкіші до впливу шумів у мовному сигна-

лі, та STRFCC-ознаки, які моделюють роботу слухової системи людини при розпізнаванні мовців. Результати досліджень дозволяють рекомендувати PNCC-ознаки як базові для оцінювання основного тону. На рис. 4в показано залежність якості розпізнавання мовців АСРМКЗ від виду тестувальних даних та кількості фреймів у пакеті, яка дорівнює $2p$. Оптимальною вилося значення $p = 5$. На рис. 4г показано залежність якості розпізнавання мовців АСРМКЗ від виду тестувальних даних та кількості нейронів на прихованих шарах нейромереж. Оптимальною виявилася кількість нейронів на прихованих шарах, рівна 1024. На рис. 4д показано залежність якості розпізнавання мовців АСРМКЗ від виду тестувальних даних та деталізації оцінок основного тону, яка виражається у кількості станів. Як виявилося, 68 станів основного тону для задачі розпізнавання мовця АСРМКЗ достатньо. І нарешті, на рис. 4е показано залежність якості розпізнавання мовців АСРМКЗ від SNR у вхідному мовному матеріалі другої тестувальної вибірки та оцінками основного тону, отриманими за DNN, RNN та FHMM методами, параметри яких оптимізовано з урахуванням вищенаведених досліджень. Для всіх рівнів SNR найкращі результати показала система, яка використовувала інформацію від FHMM-методу, що корелюється з показаними на рис. 3 результатами.

Висновки

Основний тон є однією з головних індивідуальних характеристик мови людини, який описує вплив мовного джерела (низок) у процесі мовотворення, що зумовило широке використання цієї характеристики для задачі розпізнавання мовця. Існуючі методи базуються на класичній теорії обробки сигналів і не забезпечують достатнього рівня адаптації до впливу різного виду шумів у мовному матеріалі, що, відповідно, знижує якість отримуваних з їх допомогою оцінок основного тону. Методи на основі теорії машинного навчання показують зіставні результати.

Отже, вперше запропоновано метод оцінюванні трендів основного тону, який на відміну від існуючих, використовує оптимізовану із застосуванням дерева переходів факторіальну приховану Марковську модель для формування трендів основного тону, узагальнюючи інформації від детекторів станів основного тону на основі глибокої та рекурентної нейромереж, що дозволило прогнозувати оцінки станів основного тону, використовуючи довготривалу інформацію з пакетів фреймів мовного сигналу, описати часову динаміку основного тону та зменшити вплив шумів у мовному сигналі на якість оцінок основного тону.

До практичної цінності проведених досліджень можна віднести сформульовані методи оцінювання станів основного тону на основі глибокої та рекурентної нейромереж та метод оцінювання трендів основного тону на основі факторіальної прихованої Марковської моделі. Проведено дослідження для оптимізації параметрів запропонованих методів для використання у складі АСРМКЗ, зокрема, результати досліджень дозволяють рекомендувати PNCC-ознаки як базові для оцінювання основного тону, застосовувати в роботі методів пакети фреймів із 10 фреймів, будувати нейромережі, використовуючи на прихованих шарах 1024 нейрони та використовувати 68 станів для опису основного тону. Результати проведених досліджень якості розпізнавання мовців АСРМКЗ від SNR у вхідному мовному матеріалі та оцінками основного тону, отриманими за DNN, RNN та FHMM методами, параметри яких оптимізовано з урахуванням проведених досліджень, показали, що для всіх рівнів SNR найточніші оцінки основного тону дає FHMM-метод, забезпечуючи імовірність правильного розпізнавання мовців на рівні 96...99 %, як це видно з рис. 4е.

До недоліків створених методів можна віднести значну обчислювальну складність базового математичного апарату глибоких нейромереж, що продемонстровано результатами дослідження, показаними на рис. 3ж. Це зумовлює спрямування подальших досліджень у напрямку оптимізації кількості станів для опису основного тону зі збереженням інформативності цієї ознаки для розпізнавання мовців АСРМКЗ із застосуванням актуальних методів факторного аналізу.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] B. S. Atal, “Automatic speaker recognition based on pitch contours,” *J. Acoust. Soc. Amer.*, vol. 52, pp. 1687-1697, 1972.

[2] D. J. Hermes, “Measurement of pitch by subharmonic summation,” *J. Acoust. Soc. Amer.*, vol. 83, p. 257-264, 1988.

[3] M. R. Schroeder, “Period histogram and product spectrum: New methods for fundamental-frequency measurement,” *J. Acoust. Soc. Amer.*, vol. 43, pp. 829-834, 1968.

[4] W. Chu, and A. Alwan, “SAFE: A statistical approach to F0 estimation under clean and noisy conditions,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 933-944, 2012.

[5] S. Gonzalez, and M. Brookes, "PEFAC-A pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 2, pp. 518-530, 2014.

[6] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech Coding Synth.*, vol. 495, pp. 497-518, 1995.

[7] A. De Cheveigne, and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 111, pp. 1917-1930, 2002.

[8] M. Wu, D. L. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. Speech, Audio Process.*, vol. 11, no. 3, pp. 229-241, 2003.

[9] Z. Jin, and D. L.Wang, "HMM-based multipitch tracking for noisy and reverberant speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 5, pp. 1091-1102, 2011.

[10] B. S. Lee, and D. P. W. Ellis, "Noise robust pitch tracking by subband autocorrelation classification," in *13th Annual Conference of the International Speech Communication Association*, 2012. doi 10.7916/D86M3H3S.

[11] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," *Proc. of AISTATS*, pp. 315–323, 2011.

[12] O. Vinyals, S. V. Ravuri, and D. Povey, "Revisiting recurrent neural networks for robust ASR," *Proc. ICASSP*, pp. 4085–4088, 2012.

[13] Z. Ghahramani, and M. Jordan, "Factorial hidden Markov models," *Mach. Learn*. vol. 29, pp. 245-273, 1997.

[14] M. Jordan, Z. Ghahramani, and T. Jaakkola, "An introduction to variational methods for graphical models," *Mach. Learn*. vol. 37, pp. 183-233, 1999.

[15] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Commun.*, vol. 9, no. 4, pp. 351-356, 1990.

[16] A. Varga, and H. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247-251, 1993.

[17] L. Rabiner, M. Cheng, and A. Rosenberg, "A comparative performance study of several pitch detection algorithms," *IEEE Transactions on Acoustics, Speech and Signal Proc.*, vol. 24, no. 5, pp. 399-418, 1976.

[18] R. H. Mohd, M. Zamil, and B. K. Mohd, "Speaker identification using MFCC coefficients," in *3rd international conference on electrical and computer engineering (ICECE)*, 2004.

[19] C. Kim, and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 24, pp. 1315-1329, 2012.

[20] J. C. Wang, C. H. Lin, and E. T. Chen, "Spectral-temporal receptive fields and MFCC balanced feature extraction for noisy speech recognition," *Asia-Pacific Signal and Information Processing Association (APSIPA)*, 2014. https://doi.org/10.1007%2Fs11042-016-3335-0.

[21] K. Han, and DeL. Wang, "Neural network based pitch tracking in very noisy speech," *IEEE/ACM Transactions on Audio, Speech and Language Proc.*, vol. 22, no. 12, pp. 2158-2168, 2014.

[22] М. М. Биков, та В. В. Ковтун, «Оцінювання надійності автоматизованих систем розпізнавання мовців критичного застосування,» *Вісник Вінницького політехнічного інституту*, № 2, с. 70-76, 2017.

Рекомендована кафедрою комп'ютерних систем управління ВНТУ

Ковтун В'ячеслав Васильович — канд. техн. наук, доцент, доцент кафедри комп'ютерних систем управління, e-mail: kovtun_v_v@vntu.edu.ua .

Вінницький національний технічний університет, Вінниця

V. V. Kovtun[1]

Pitch Estimation for Automated Speaker Recognition System for Critical Use

[1]Vinnytsia National Technical University

The article proposes a method for pitch trend estimation, which, unlike existing ones, uses a factorial hidden Markov model optimized with the junction tree algorithm for pitch trend estimation, generalizing information from pitch state detectors based on deep and recurrent neural networks, with which it is allowed precisely to predict a pitch trend using long-term information from speech frames packets, describe the dynamics of the pitch in the time domain and reduce the noise influence on the quality of pitch estimates. Methods for estimating pitch states based on deep and recurrent neural networks and a method for estimating the pitch trend based on the factorial hidden Markov model (FHMM) are developed. A study was carried out to optimize the parameters of the proposed methods for use as part of the automated speaker recognition system for critical use (ASRSCU). In particular, the results of the research make it possible to recommend power-normalized cepstral characteristics as the basis for estimating the pitch by the proposed methods, to apply frames packets with a duration of 10 frames, to use 1024 neurons in the hidden layers of neural networks that implement the proposed methods, and to use 68 states to describe the pitch. The results of the conducted researches of the dependence of the quality of speakers recognition by the ASRSCU from the level of the signal-to-noise ratio (SNR) in the input speech material and the pitch estimates obtained as a result of the work of the

created methods, the parameters of which are optimized taking into account the results of the conducted studies, showed that for all levels of SNR the exact pitch estimate is provided by the FHMM method, showing the correct speakers recognition probability by the ASRSCU at a level of 96…99 % for the selected test sample.

Keywords: automated speaker recognition system for critical use, pitch, deep neural network, recurrent neural network, factorial hidden Markov model.

Kovtun Viacheslav V. — Cand. Sc. (Eng.), Assistant Professor of the Chair of Computer Control Systems, e-mail: kovtun_v_v@vntu.edu.ua

В. В. Ковтун[1]

Оценивание основного тона в автоматизированной системы распознавания диктора критического применения

[1]Винницкий национальный технический университет

Предложен метод оценки трендов основного тона, который в отличие от существующих, использует оптимизированную с применением дерева переходов факториальную скрытую Марковскую модель для формирования трендов основного тона, обобщая информацию от детекторов состояний основного тона на основе глубокой и рекуррентной нейросетей, что позволило спрогнозировать оценки состояний основного тона, используя долговременную информацию из пакетов фреймов речевого сигнала, описать динамику основного тона во времени и уменьшить влияние шумов в речевом сигнале на качество оценок основного тона. Созданы методы оценки состояний основного тона на основе глубокой и рекуррентной нейросетей и метод оценки трендов основного тона на основе факториальной скрытой Марковской модели (ФСММ). Проведено исследование для оптимизации параметров предложенных методов для использования в составе автоматизированной системы распознавания диктора критического применения (АСРДКП). В частности, результаты исследований позволяют рекомендовать нормированные по мощности кепстральные признаки как базовые для оценки основного тона предложенными методами, применять при работе методов пакеты фреймов продолжительностью 10 фреймов, использовать 1024 нейрона в скрытых слоях нейросетей, которые реализуют предложенные методы, и использовать 68 состояний для описания основного тона. Результаты проведенных исследований зависимости качества распознавания дикторов АСРДКП от уровня отношение сигнал/шум (ОСШ) во входном речевом материале и оценками основного тона, полученными в результате работы предложенных методов, параметры которых оптимизированы с учетом результатов проведенных исследований, показали, что для всех уровней ОСШ самые точные оценки основного тона обеспечивает ФСММ-метод, показывая вероятность правильного распознавания дикторов АСРДКП на уровне 96…99 % для выбранной тестовой выборки.

Ключевые слова: автоматизированная система распознавания дикторов критического применения, основной тон, глубокая нейросеть, рекуррентная нейросеть, факториальная скрытая Марковская модель.

Ковтун Вячеслав Васильевич — канд. техн. наук, доцент, доцент кафедры компьютерных систем управления, e-mail: kovtun_v_v@vntu.edu.ua