УДК 004.8

Д.Ю. Яцина

Харківський національний університет імені В.Н. Каразіна, Харків

НЕЙРОННІ МЕРЕЖІ ДЛЯ РОЗРОБКИ СИСТЕМИ КОНВЕРТАЦІЇ ГОЛОСУ

Система конвертації голосу формулює функцію конвертації специфічних характеристик початкового голосу до відповідних характеристик цільового голосу. В цій статті використовуються такі характеристики голосу: форма вокального тракту, форма збуджувального сигналу (імпульс голосової щілини) та просодичні характеристики (енергія, висота тону). Проведено порівняння функцій конвертації реалізованих за допомогою нейронної мережі радіально-базисних функцій та нейронної мережі загальної регресії. Реалізовано новий метод виявлення аномальних значень в наборі даних.

***Ключові слова**: вокодер з лінійним предикатом, пакетне вейвлет-перетворення, нейронна мережа радіально-базисних функцій та нейронна мережа загальної регресії, метод головних компонент.*

Вступ

Задача зміни голосу полягає у зміненні акустичних параметрів фрази одного оратора (початкового) таким чином, щоб ця фраза сприймалася вимовленою голосом іншого оратора(цільового). Мета нашого дослідження - модифікація індивідуальних та просодичних характеристик початкового голосу до цільового за допомогою нейронних мереж та порівняння найбільш відповідних архітектур.

Сфера застосування даної технології зміни голосу дуже обширна. Починаючи від сфер мистецтва та розваг до безпеки інформаційних систем, а саме: дубляж фільмів, персоніфікація тексту в фразу, караоке додатки, зміна ідентичності голосу з метою приватності та безпеки, системи корекції мовлення пацієнтів з хворобами голосового апарату, синтез вокальних даних на студіях звукозапису. Конвертація голосу може бути застосована до будь-якої задачі де є записане мовлення і потрібно ефективно створити бажану різноманітність голосів без додаткового запису різних вокалістів. Модифікація голосу важлива для проектів інтерпретивної телефонії, в яких іноземці розмовляють різними мовами, а система перекладає їх фрази і синтезує з відповідними голосовими характеристиками. Таким чином можливо почути як іноземець буде розмовляти іншою мовою. Всі ці сфери потребують покращення у якості конвертації голосу, бо ще не досягнута задовільна якість для всезагального застосування даної технології. Нейронні мережі показали високі результати в сферах розпізнавання мовлення [7], тому дослідження зміни голосу за допомогою нейронних мереж актуально.

Задача конвертації голосу достатньо комплексна, велика та багатокомпонентна. Велика кількість різноманітних рішень вже була запропонована. Будь-яка реалізація базується на аналізу записаних звукових даних, де система зміни голосу обчислює характеристики фраз початкового і цільового оратора та формулює функцію конвертації для модифікації характеристик фрази початкового оратора таким чином, щоб ресинтезована фраза звучала ніби вимовленою цільовим оратором.

Записані данні початкового і цільового оратора бувають двох видів:

1) паралельні (набір фраз, кожну з яких вимовляє і початковий, і цільовий оратори);

2) непаралельні (присутні різні фрази початкового оратора і фрази цільового оратора).

Загалом усі методи можливо розподілити на моделюючі та мутаційні [8]. Моделюючі методи конвертують характеристики початкового голосу для кожного короткого сегменту фрази за допомогою завчасно підготовленої функції конвертації, моделюючи характеристики сегментів цільового голосу, які описуються низькою розмірністю. Мутаційні методи конвертують характеристики цілої фрази, тому розмірність цих характеристик дуже велика.

Для конвертації голосу використовуються різноманітні специфічні характеристики оратора, такі як форма вокального тракту, форма збуджувального сигналу і просодичні характеристики. Для опису вокального тракту і збуджувального сигналу використовуються різноманітні вокодери [1]: LPC вокодер, вокодер формант, MFCC вокодер, HNM вокодер та STRAIGHT вокодер. На сьогодні найбільш успішні функції конвертації вокальних трактів реалізовані за допомогою моделей гаусових сумішей (GMM)[3], прихованої марковської моделі (HMM). Але надмірне зглажування параметрів обмежує корисність використання цих підходів і голос звучить надто роботизовано. Існуючі методи конвертації голосу використовують паралельні набори даних і синхронізують фрази ораторів у часі за допомогою алгоритму динамічної трансформації часової шкали[9], який працює із похибками, що погіршують якість зміни голосу.

Наше дослідження присвячене застосуванню нейронних мереж, які спроможні на нелінійні перетворення параметрів, для конвертації голосу з паралельними даними. Ми спробували наступні архітектури

нейронних мереж: нейронна мережа радіальнобазисних функцій(RBFNN) та нейронна мережа загальної регресії(GRNN). Для аналізу/синтезу мовлення ми використали математичну модель голосового тракту кодування з лінійним предиктором (LPC вокодер) [4], яка обчислює два параметри звукового голосового сегменту: форму вокального тракту та форму збуджувального сигналу. Для зменшення розмірності збуджувального сигналу ми застосували дискретне вейвлет-перетворення (DWT) та пакетне вейвлет-перетворення(WPT), порівнявши їх [5]. Ми виконали кластеризацію звуків на голосні та приголосні, що дало покращення в моделюванні збуджувального сигналу. Для покращення якості ми також виявили аномальні сегменти в паралельному наборі даних за допомогою методу головних компонент (ROBPCA) [6].

Запропонована система конвертації голосу

Стисло і спрощено роботу нашої системи можна описати так: фраза початкового оратора розділяється на короткі сегменти, потім кожен сегмент конвертується і набуває індивідуальних та просодичних характеристик цільового оратора, а наприкінці усі конвертовані сегменти склеюються та отримується конвертована фраза. Кожен сегмент по суті описує не слово, не букву, а найелементарнішу неподільну одиницю мовлення – звук. Насправді ж, система зміни голосу має багато комплексних та складних деталей.

Паралельний набір даних. Паралельний набір записаних фраз початкового і цільового оратора - це головне, що ми аналізуємо для побудови правил конвертації голосу. Ми досліджували CMU Arctic dataset [10], який містить близько 1132 фраз записаних 4-ма ораторами: 2 чоловіками та 2 жінками. Будь-який набір даних потрібно обробити перед використанням. Як зазначено в [11] тиша, яка виникає в паузах між словами зайва для побудови функції конвертації, тому ми знаходимо ділянки тиші за допомогою алгоритму виявлення голосової активності(VAD) і видаляємо їх із набору даних. Усі фрази також нормалізуються до одного рівня гучності в діапазоні (-1,1).

Далі розділяємо всі фрази на звукові сегменти довжиною в 30 мс та кроком 15 мс(тобто кожен наступний сегмент охоплює половину попереднього.

Внаслідок цього фраза початкового і цільового оратора буде містити різну кількість сегментів, бо записані фрази цих ораторів мають різну тривалість, адже кожна людина вимовляє звуки з різною швидкістю. Таким чином синхронізація сегментів початкового і цільового оратора у часі необхідна для їх порівняння. Синхронізація та вимірювання у часі виконується за допомогою алгоритму динамічної трансформації часової шкали (DTW) [9] і кожен сегмент описується Мел-кепстральними коефіцієнтами(MFCC). Після цього процесу першому звуковому сегменту початкового оратора(наприклад звук “а”) буду відповідати перший такий же звуковий сегмент (звук “а”) цільового оратора. На рисунку зображено вирівнювання сегментів у часі, а червона лінія показує відповідність початкових сегементів до цільових (рис. 1). Також зображені звукові хвилі фрази до та після синхронізації (рис. 2 і 3).

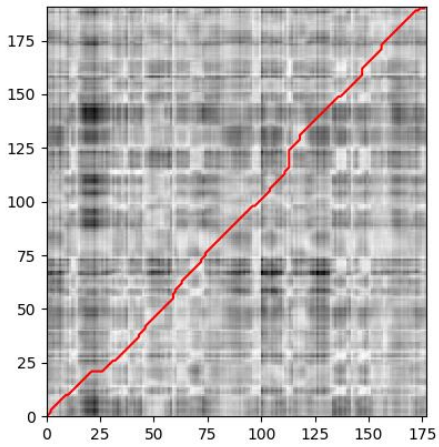

Рис. 1. Синхронізація сегментів початкового вокаліста та сегментів цільового вокаліста

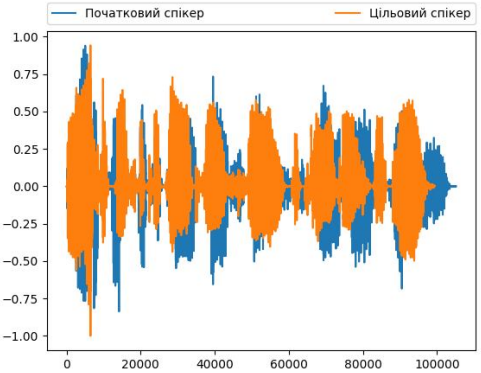

Рис. 2. Звукові хвилі фрази початкового та цільового оратора до синхронізації

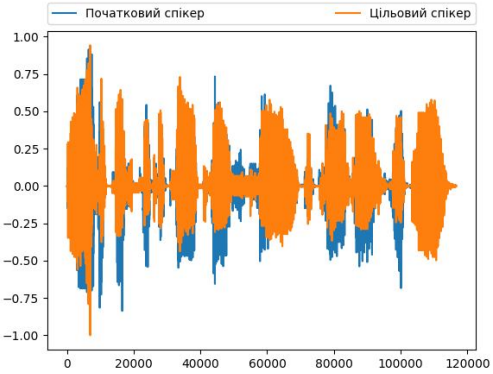

Рис. 3. Звукові хвилі фрази початкового та цільового оратора після сихронізації

Видалення аномальних паралельних сегментів у наборі даних. Оскільки алгоритм динамічної трансформації часової шкали працює з похибками для синхронізації паралельних фраз, то створюються окремі не відповідні сегменти, наприклад звуку “с”

початкового оратора буде відповідати звук "п" цільового оратора. В такому випадку функція конвертації буде навчатись виконувати не правильні прогнозування і такі паралельні сегменти будуть критичними для побудови якісної системи зміни голосу. Очевидно, що кількість таких випадків має бути зменшена то мінімуму.

Ми застосували відстань Махаланобіса для знаходження аномальних сегментів:

$$D_x = \sqrt{(x_i - \mu)^T \cdot C^{-1} \cdot (x_i - \mu)}, \qquad (1)$$

де x_i – об'єднаний паралельний сегмент початкового і цільового оратора, C – коваріаційна матриця. Оскільки розмірність набору даних достатньо велика обчислення відстані Махаланобіса для кожного сегменту потребує потужних ресурсів, тому для зменшення розмірності даних ми застосовуємо метод головних компонент(ROBPCA) [6], який враховує ефект аномальних даних. Наприкінці всі відстані сегментів, які перевищують граничне значення $\sqrt{(\chi^2)_{k;0,975}}$ (де k — розмірність даних) визначаються аномальними і видаляються із тренувального набору даних.

Вибір характеристик. Після обробки набору даних кожен сегмент представляє звукову хвилю. Маємо перейти від звукової хвилі до специфічних характеристик. Для аналізу/синтезу звукових хвиль обрано вокодер кодування з лінійним предиктором (LPC вокодер) 16-го порядку. Протягом процесу аналізу із звукової хвилі отримуємо форму вокального тракту та форму збуджувального сигналу і навпаки протягом процесу синтезу із цих 2-х характеристик отримуємо звукову хвилю [12]. Форма вокального тракту описується 16-ма коефіцієнтами α_k. Щоб отримати збуджувальний сигнал потрібно пропустити звукову хвилю через фільтр:

$$1 + \sum_{\alpha=1}^{order} \alpha_k \cdot z^{-k}. \qquad (2)$$

А щоб синтезувати звукову хвилю потрібно збуджувальний сигнал пропустити через інверсний фільтр:

$$1 \Big/ \left(1 + \sum_{\alpha=1}^{order} \alpha_k \cdot z^{-k}\right). \qquad (3)$$

Для стабільності модифікації форми вокального тракту, LPC коефіцієнти конвертуються в LSF коефіцієнти [6]. Збуджувальний сигнал має довжину всього сегменту (розмірність=480), але такий розмір сигналу достатньо комплексний та потребує багато часу для розробки функції конвертації. Для зменшення його розмірності використовується вейвлет-перетворення. Дискретне перетворення втрачає ефективність у випадку коли фраза містить шуми [2]. Проте пакетне вейвлет-перетворення стійке до шумового середовища. Ми обрали 4-х рівневе пакетне Вейвлет-перетворення, яке розкладає весь збуджувальний сигнал на 16 коротких смуг (розмірність кожної = 480/16=30), які описують збуджувальний сигнал в 16 різних частотних діапазонах. Щоб отримати збуджувальний сигнал знову, ми виконуємо інверсне пакетне вейвлетне-перетворення 16 коротких смуг.

Ми врахували іще один момент, який проявляється у природі збуджувального сигналу. Мовлення складається із суміші голосних та приголосних звуків, а спостереження показали, що форма збуджувального сигналу для більшості приголосних звуків (с, ш, з і т.д.) має стохастичний вигляд, а для голосних (а, о, у і т.д.) - періодичний вигляд [12]. Таким чином конвертувати збуджувальний сигнал для деяких приголосних сегментів не потрібно, бо такий сигнал простіше змоделювати стохастично, а голосні сегменти ми конвертуємо. Тому класифікували кожен сегмент у наборі даних по даному критерію за допомогою нового методу кластеризації спектральних та часових характеристик звукових сегментів [17].

Також для кожного сегмента звуку обчислюємо енергію звукової хвилі та висоту тона (фундаментальна частота), що знадобиться нам для зміни просодичних характеристик.

Функції конвертації характеристик. Кожен сегмент описується вокальним трактом та 16-ма короткими смугами збуджувального сигналу. В сумі для побудови системи треба побудувати 17 функцій конвертацій. Система зміни голосу отримується в два етапи: тренування і тестування. Весь набір даних розділяється на два: тренувальний та тестувальний. Протягом тренування ми навчаємо функції конвертацій прогнозувати характеристики цільового сигналу із характеристик початкового сигналу. Тобто почергово асоціюємо характеристики сегменту початкового оратора з характеристиками відповідного сегменту цільового оратора. Протягом етапу тестування ми трансформуємо сегменти початкового оратора через отримані функції конвертації та оцінюємо якість, порівнюючи характеристики конвертованих сегментів з бажаними. Оскільки перетворення характеристик не є лінійним, нейронні мережі повинні показати свою ефективність в реалізації функцій конвертацій. Ми дослідили наступні архітектури нейронних мереж: Нейронна мережа радіально-базисних функцій (RBFNN) та нейронна мережа загальної регресії (GRNN). Нейронна мережа радіально-базисних функцій – це спеціальний вид нейронної мережі прямого поширення (Feed forward neural network), яка асоціює вхідний простір нелінійно в схований простір, а потім схований простір лінійно асоціює в вихідний простір. Дана нейрона мережа складається з трьох шарів з m вхідними вузлами, p схованими вузлами та з одним чи багатьма вихідними вузлами [13, 14]. Для нашої задачі кількість вихідних вузлів така сама як і кількість вхідних. Нейронна мережа загальної регресії має особливу властивість, що вона не потребує процесу ітера-

тивного тренування для оптимізації параметрів, тому що використана гаусова функція активації слугує обчислювачем необхідних параметрів, а ми тільки вказуємо її центр θ та ширина σ. Ця нейрона мережа складається з трьох шарів з m вхідними вузлами, p схованими вузлами та m вихідними вузлами [15].

Експериментальні результати

Запропонована система зміни голосу складається з багатьох компонентів. Всі компоненти були реалізовані на мові програмування Python, а нейронні мережі були спроектовані за допомогою модуля Tensor Flow. Прослухати приклади конвертованих фраз можливо за посиланням [16].

Для побудови системи конвертації голосу потрібно мінімально 40 паралельних фраз. Тому ми використовували лише частину фраз із набору даних.

Основна оптимізація системи зосереджена на параметрах нейронних мережах. Загалом процес конвертації полягає у побудові 17 нейронних мереж: одна – для конвертації вокальних трактів, а інші 16 – для конвертування 16 смуг декомпозиції збуджувального сигналу. Для радіально-базисної нейронної мережі ми оптимізували розмір схованого шару та параметр ширини σ, а для загальної регресії лише параметр ширини σ на проміжку значень.

Для обчислення похибки використовується наступна функція:

$$\mathrm{Err} = \sum_n \sum_q \left(y_q\left(x^{(n)}\right) - t^{(n)}_{(q)} \right),$$

де $\left(x^{(n)}, t^{(n)}\right)$ - набір даних, $t^{(n)}_{(q)}$ - бажане значення q вузла, коли на вході до мережі вектор $x^{(n)}$.

Опираючись на чисельні експерименти, ми вирішили застосовувати радіально-базисну нейронну мережу для прогнозування форми вокальних трактів, а нейронну мережу загальної регресії для моделювання збуджувального сигналу.

Кількість вузлів у схованому шарі радіально-базисної нейронної мережі, а саме кількість кластерів вхідного набору була визначена в 80 кластерів. А оптимальний параметр ширини σ обирався на проміжку (0.01, 2.5). Мінімальна похибка отримувалася при параметрі ширини σ = 0.51 . (рис. 4).

Нейронна мережа загальної регресії показала переваги в прогнозуванні смуги декомпозиції збуджувального сигналу завдяки можливості узагальнення та усереднення найбільш імовірних виходів. Оскільки для кожного звукового сегменту ми маємо 16 смуг збуджувального сигналу, то ми оптимізуємо 16 нейронних мереж загальної регресії, в сховані шари яких ми кладемо всі відповідні смуги тренувального набору даних і підбираємо експериментальне значення ширини σ для кожної нейронної мережі.

Приклад оптимізації для першої смуги збуджувального сигналу зображено на рис. 5.

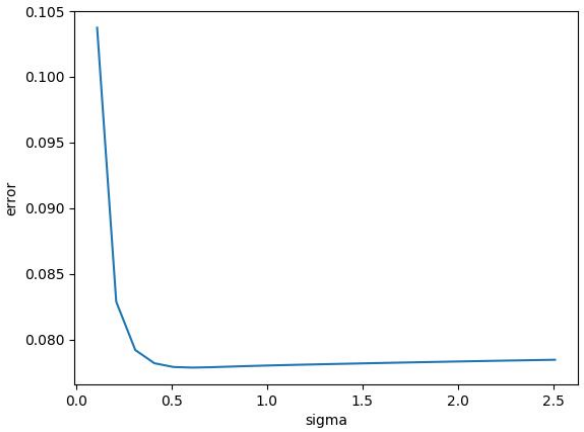

Рис. 4. Оптимізація параметру ширини нейронної мережі радіально-базисних функцій для конвертації вокальних трактів

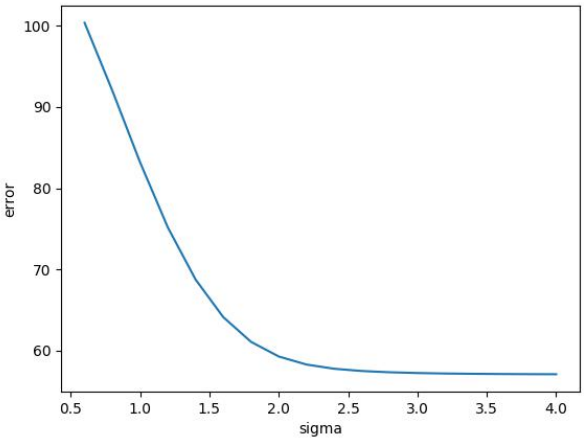

Рис. 5. Оптимізація параметру ширини нейронної мережі загальної регресії для конвертації першої смуги збуджувального сигналу

Для того щоб отримати конвертовані фрази ми для кожного сегменту тестового набору даних виконуємо:

- збір всіх 16 конвертованих смуг збуджувального сигналу та виконуємо інверсне пакетне вейвлет перетворення щоб отримати цілий збуджувальний сигнал
- синтезуємо звукову хвилю за допомогою конвертованої форми вокального тракту та конвертованого збуджувального сигналу
- змінюємо просодичні характеристики кожного конвертованого сегменту: модифікуємо енергію відповідно до сегменту початкового оратора та змінюємо висоту тону, щоб середня висота тону конвертованої фрази відповідала середній висоті тону початкової фрази. Наостанок з'єднуємо всі конвертовані сегменти та отримуємо конвертовані фрази.

Висновки

Ми побудували систему зміни голосу за допомогою нейронних мереж, а також реалізували новітні модифікації в загальній побудові систем синтезу мовлення. Експерименти показали, що для конвер-

тації вокальних трактів найменша похибка була отримана за допомогою нейронної мережі радіально-базисних функцій, а для збуджувального сигналу – за допомогою нейронної мережі загальної регресії. Видалення аномальних сегментів в паралельному наборі даних також покращило якість конвертації голосу. В подальших планах хочемо розробити компонент, який згладжує з'єднанні звукові сегменти, тому що в областях з'єднання сегментів інколи виникають стрибки сигналу, внаслідок цього виникають зайві шуми та похибки.

Також плануємо замінити класичний етап вибору характеристик звукових сегментів на глибинне навчання, що дозволить отримати абстрактне представлення характеристик, які легше обробляти, конвертувати та прогнозувати.

Запропонований спосіб реалізації потребує підготовчих етапів, таких як запис паралельних даних, що ускладнює процес отримання робочої системи. Надалі будемо реалізовувати систему конвертації голосу, яка не потребує паралельних даних. В такому випадку для отримання системи потрібні будь-які фрази цільового оратора і будь-які фрази початкового оратора. Зараз активно з'являються нові вокодери для аналізу/синтезу мовлення, які потребують дослідження. В реальних ситуаціях рідко зустрічаються записи мовлення ораторів без зайвих шумів, тому дослідження з метою отримання вокодерів стійких до шумового середовища актуально. Шляхів покращення роботи системи конвертації голосу дуже багато, всі вони потребують ґрунтовного дослідження.

Список літератури

1. A.N. Chadha, A comparative performance of various speech analysis-synthesis techniques / A.N. Chadha, J.H. Nirmal, P. Kachare // Int. J. Signal Process. Syst. 2 (1) (2014) 17–22.

2. J. Nirmal, Voice conversion using General Regression Neural Network, Applied Soft Computing / J. Nirmal, M. Zaveri, S. Patnaik, P. Kachare. 2014.

3. W. Kain Spectral voice conversion for text-to- speech synthesis / W. Kain, M. Macon // In: Proceeding of International Conference on Acoustics, Speech, and Signal Processing, vol. 1, IEEE, 1998, pp. 285–288.

4. K.S. Rao, Voice conversion by mapping the speaker-specific features using pitch synchronous approach / K.S. Rao // Comput. Speech Lang. 24 (3) (2010). 474–494.

5. S. Desai Spectral mapping using artificial neural networks for voice conversion/ S. Desai, A.W. Black, B. Yegnanarayana, K. Prahallad // IEEE Trans. Audio Speech Lang.Process. 18 (5) (2010) 954–964.

6. Sushant V. Rao Novel Pre-processing using Outlier Removal in Voice Conversion/ Sushant V. Rao, Nirmesh J. Shah, Hemant A. Patil, 2016.

7. S.H. Mohammadi, Voice Conversion Using Deep Neural Networks With Speaker-Independent Pre-training,/ S.H. Mohammadi, A. Kain, 2014.

8. S.H. Mohammadi Transmutative Voice Conversion / S.H. Mohammadi, A. Kain 2013.

9. Holmes, J.N. Speech synthesis and recognition/John Holmes and Wendy Holmes.—2 nd ed, 2001.

10. Набір голосових даних [Електронний ресурс]. – Режим доступу: festvox.org/cmu_arctic.

11. E. Helander. On the impact of alignment on voice conversion performance / E. Helander, H Silén, M Gabbouj, 2008.

12. T.H. Park Introduction To Digital Signal Processing / T.H. Park, 2010.

13. S. Haykin Neural networks and learning machines / Simon Haykin. – 3rd ed.

14. M. Bishop Neural Networks for Pattern Recognition / M. Bishop. 1995.

15. A. Amrouche. Efficient System for Speech Recognition using General Regression Neural Network / A. Amrouche, J.M. Rouvaen, 2008.

16. Приклади конвертації [Електронний ресурс]. – Режим доступу: https://drive.google.com/open?id=0BwP19oq-ytjEZVh0Vl9jOEpnaGM.

17. S. Mondal. Clustering based voiced-unvoiced-silence detection in speech using temporal and spectral parameters/ S. Mondal, A. D. Barman, 2015.

Надійшла до редколегії 16.03.2017

Рецензент: д-р техн. наук, проф. В.А. Краснобаєв, Харківський національний університет імені В.Н. Каразіна, Харків.

НЕЙРОННЫЕ СЕТИ ДЛЯ РАЗРАБОТКИ СИСТЕМЫ КОНВЕРТАЦИИ ГОЛОСА

Д.Ю. Яцина

Система конвертации голоса формулирует функцию конвертации специфических характеристик начального голоса к соответственным характеристикам целевого голоса. В статье используются такие характеристики голоса: форма вокального тракта, форма возбудительного сигнала и просодические характеристики(энергия, высота тона). Проведено сравнение функций конвертации реализованных с помощью нейронной сети радиально-базисных функций и нейронной сети общей регрессии. Реализовано новый метод обнаружения аномальных значений в наборе данных.

Ключевые слова: *вокодер з линейным предикатом, пакетное вейвлет-преобразование, нейронная сеть радиально-базисных функций и нейронной сети общей регрессии, метод главных компонент.*

NEURAL NETWORKS FOR VOICE CONVERSION SYSTEM DESIGN

D.Yu. Yatsyna

Voice conversion system formulates conversion function, which can transform specific parameters of source speaker to taregt speaker. In this paper we used voice parameters: shape of the vocal tract, shape of exciatation signal(glottal pulse) and prosodic features(energy, pitch). We compare Radial Basis Function Neural Network and General Regression Neural Network for parameters conversion. We implement new method for outlier detection in dataset.

Keywords: *LPC-vocoder, Wavelet Packet Transform, RBF Neural Network, General Regression Neural Network, Robust Principal Component Analysis.*