DOI 10.31891/2307-5732-2018-267-6(1)-123-127
УДК 621.382.2

А.С. КАШТАЛЬЯН
Хмельницький національний університет

СЕГМЕНТАЦІЯ ТА КЛАСТЕРИЗАЦІЯ АУДІО СИГНАЛУ НА ОСНОВІ ПРИХОВАНОЇ МАРКІВСЬКОЇ МОДЕЛІ

Аналіз аудіо сигналу є одним з важливих завдань з огляду на постійно зростаючі об'єми аудіо та відео інформації. Стаття присвячена задачі сегментації та кластеризації аудіо сигналу.. Сегментація аудіо сигналу передбачає визначення границь, на яких змінюються характеристики, та виділення однорідних ділянок. Багато існуючий методів ґрунтуються на детектуванні точок зміни характеристик для пошуку границь сегментів. В статті розглядається об'єднаний підхід до сегментації і кластеризації, що ґрунтується на підході без вчителя в моделях послідовних даних. Приховані харківські моделі є вдалим вибором для моделювання аудіо сигналу, що представляє собою послідовність даних. В статті досліджено метод сегментації та кластеризації на основі такого типу моделей. Представлено існуючий ітеративний ЕМ алгоритм (Expectation-maximization algorithm) для прихованих моделей Маркова та розширення методу для напівприхованих моделей Маркова та їх застосування до моделювання аудіо сигналу його сегментації та кластеризації. В запропонованому методі кожний прихований стан харківської моделі відображає кластер сегментів. Початково модель ініціалізується з завідомо більшою кількістю кластерів, ніж існує за апріорною оцінкою, з метою зменшення ймовірності потрапляння неоднорідних сегментів в один кластер. Сегментація виконується з використанням алгоритму Вітербі в кожному кластері. Наступним кроком є зменшення числа кластерів шляхом їх об'єднання. Кластери об'єднуються відповідно до відношення правдоподібності. Новий клас представляється новою моделлю, параметри якої формуються за ЕМ алгоритмом. Сегментація переоцінюється з новою топологією прихованої харківської моделі, яка містить на один кластер менше ніж попередня. Процес ітеративно повторюється до досягнення максимуму оціночної функції. В статті представлені результати експерименту сегментації та кластеризації аудіо сигналу з різним типом змісту з неоднорідними ділянками.

Ключові слова: моделі Маркова, апостеріорна ймовірність, алгоритм Вітербі, послідовність прихованих станів.

A. KASHTALIAN
Khmelnytsky National University

AUDIO SIGNAL SEGMENTATION AND CLUSTERING BASED ON HIDDEN MARKOV MODEL

The audio signal analysis is important task in the condition of continuous growing amount of audio and video information. The article is devoted to the issue of audio signals segmentation and clustering task. Audio signal segmentation means the definitions of borders, on which characteristics are being changed, and homogenous segments allocation. Many existent methods are based on change points detection for segment bounds search. The join approach of simultaneous segmentation and clustering is considered in the article. It is based on a unsupervised learning approach in sequential data models. Hidden Markov models is a successful choice for audio signal modeling because it represents data sequence. The segmentation and clusterization method based on such type of models is investigated. The existent iterative EM algorithm (Expectation-maximization) for hidden Markov models and expanded method for semi-hidden Markov modes and their application to audio signal modeling, segmentation and clustering are presented. Every hidden state of Markov model reflects segments cluster in the proposed method. Initially the model is initialized with notoriously bigger number of clusters than exists by aprioristic estimate with the purpose to reduce the probability of allocation of non-homogeneous segments in one cluster. The segmentation is performed with using of Viterbi algorithm for every cluster. The next step is reducing of the cluster number by the way of joining them. Clusters are joined in order to likelihood-ratio value. A new class represents by the new model, which parameters are formed with EM algorithm. Segmentation is re-estimated with new hidden Markov topology, which contains one less cluster than previous one. The process repeats iteratively to reach the maximum of estimate function. The experiment results of segmentation and clustering are presented in the article, the audio signals have different content type with non-homogenous chunks.

Key words: Markov models, posterior probability, Viterbi algorithm, hidden states sequence.

Вступ. Розпізнавання паттернів в аудіо сигналі є складним завданням із широким спектром застосувань, таких як аналіз відео та аудіо потоків, встановлення меж речень, аналіз діалогів та відокремлення джерел сигналу. Сегментація аудіо сигналу має на меті розбиття аудіо потоку на однорідні сегменти, що є корисним при обробці значних об'ємів інформації. Однією з таких задач є сегментація та кластеризація аудіо сигналу відповідно до спікера, в ідеалі кожний сегмент має містити мовлення тільки одного спікера [1]. Кластеризація аудіо сигналу є випадком класифікації без вчителя сегментів мовлення, що ґрунтується на частотних характеристиках [2]. Серед розроблених методів кластеризації є ієрархічні та агломеративні методи, метод k-середніх та самоорганізуючі карти [3, 4]. Кластеризації спікерів може передувати сегментація аудіо, однак це може призвести до збільшення похибки кластеризації. Сегментація та кластеризація аудіо сигналу може бути оптимізована в одному процесі [5]. Процес сегментації та послідовної чи паралельної кластеризації називають діаризацією. Діаризація – це процес автоматичного розділення аудіо потоку на окремі однорідні сегменти, та встановлення відповідності між джерелами та сегментами мовлення, що дає можливість дати відповідь на питання «хто де говорить». Якщо говорити про сегментацію за джерелами мовлення, то процес діаризації охоплює верифікацію спікерів та ідентифікацію спікерів [6].

Постановка задачі. В багатьох випадках задача сегментації розглядається як задача класифікації з вчителем, коли певні сегменти і точки зміни сегментів розмічають і використовують для подальшого навчання класифікатора. Та це робить алгоритм надзвичайно залежним від наявних розмічених зразків, і може ускладнювати адаптацію до нових аудіо потоків Тому важливо сфокусуватися на підходах, які використовують навчання без вчителя для аудіо сегментації, що дозволяє уникнути потреби в розмічених

даних, і ґрунтуватися на властивостях моделі для коректної сегментації та кластеризації сегментів. Це можливо, якщо не відокремлювати процес пошуку точок зміни спікерів та пошуку схожих сегментів, а натомість об'єднати їх. Природнім в цьому випадку є використання прихованих харківських моделей, які є потужним інструментом моделювання послідовностей. Відомо, що безперервне навчання є ефективним шляхом вдосконалення результатів навчання, особливо для великих потоків аудіо даних.

Важливим для сегментації та кластеризації є представлення аудіо сигналу. Аудіо сигнал $x_a(t)$ часто описується частотним представленням, зокрема віконним перетворенням Фур'є:

Важливим для сегментації та кластеризації є представлення аудіо сигналу. Аудіо сигнал $x_a(t)$ часто описується частотним представленням, зокрема віконним перетворенням Фур'є:

$$x\left(t, e^{i\omega}\right) = \sum_{u=-\infty}^{+\infty} x[u]g[u-t]e^{-i\omega u},$$

де $g[u-t]$ – деяка віконна функція (вікно Гаусса, вікно Хеммінга, вікно Хеннінга), кожне $x\left(t, e^{i\omega}\right)$ для фіксованого t може бути отримане дискретним перетворенням Фур'є сигналу $u \to x[u]g[u-t]$ в часову вікні з центром t, отже є фіксована кількість p коефіцієнтів $x_{t,1}, \ldots, x_{t,p} \in C$.

Надалі розглядаємо аудіо сигнал представлений модулем коефіцієнтів віконного перетворення Фур'є: ,де $x_t \in \Re^p = \left(\left|x_{1,t}\right|, \ldots, \left|x_{1,p}\right|\right)^{\mathrm{T}}$, де індекси $t = 1, \ldots T$ вказують на часові вікна, які мають постійний зсув та можуть мати певне перекриття. Для забезпечення певного рівня інваріантності до гучності звуку, розглядається нормалізований вектор x_t. Для зменшення похибки кластеризації можуть бути використані інші представлення сигналу, такі як кепстральні коефіцієнти або коефіцієнти поглинання.

Для порівняння точок необхідно використати міру подібності $D(x, y)$. В багатьох випадках в якості міри використовують евклідову відстань $D(x, y) = \|x - y\|^2$, однак вибір евклідової відстані передбачає евклідову геометрію. Емпіричні досвід показує, що для аудіо сигналу більш прийнятним є використання альтернативних мір, таких як відстань Кульбака-Лейблера або відстань Ітакури-Сайто, які представляють клас відстаней Брегмана. Відстань Брегмана визначається як $D_\psi(x, y) = \psi(x) - \psi(y) - (x - y, \nabla\psi(y))$,де ψ - строго випукла функція, і з певними припущеннями отримуємо

$$p_\mu(x) = h(x)\exp((x, \theta) - \psi(\theta)) = h_1(x)\exp\left(-D_{\psi^*}(x, \mu)\right),$$

з $h_1(x) = h(x)e^{\psi(x)}$, де ψ^* - спряжена функція ψ, μ – параметр середнього, $\theta = \nabla\psi(\mu)$.

Сегментація та кластеризація аудіо сигналу. Приховані моделі Маркова (ПММ) є прихованими моделями, в яких приховані параметри стану відповідають марківській динаміці, тому є придатними для

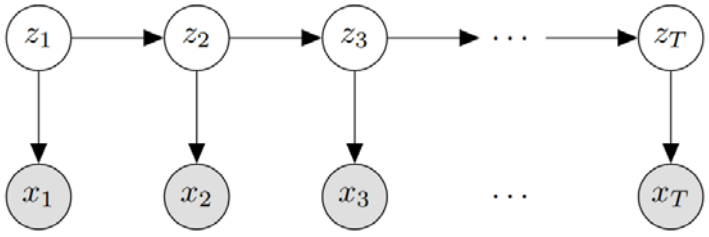


Рис. 1. Графічне представлення моделі ПММ

моделювання даних послідовної структури. $(x_t)_{t=1\ldots T}$ є послідовністю спостережень, $x_t \in \Re^p$, і $(z_t)_{t=1\ldots T}$ є послідовністю прихованих станів, до кожний стан є одним з K станів, так що $z_t \in \{1, \ldots, K\}$. ПММ є генеративною прихованою моделлю, що описується генеративним процесом (рис. 1).

$$z_1 \sim \pi,$$
$$z_t | z_{t-1} = i \sim A_i,\ t = 2, \ldots, T,$$
$$x_t | z_t = i \sim p_{\mu_i},\ t = 1, \ldots, T.$$

де π визначає розподіл z_1, $A \in \Re^{K \times K}$ – матриця переходів $A_{ij} = p(z_t = j | z_{t-1} = i)$, $A\mathbf{1}$=1 і $A_i = (A_{ij})_j$, де $1 = (1, \ldots, 1)^{\mathrm{T}}$. μ_k – параметр k-го розподілу, який асоціюється з дивергенцією Брегмана. Спільна ймовірність послідовності прихованих станів $z_{1:T} = (z_1, \ldots, z_T)$ та спостережень $x_{1:T} = (x_1, \ldots, x_T)$

$$p(x_{1:T}, z_{1:T}; \pi, A, \mu) = p(z_1; \pi)\prod_{t=2}^{T} p(z_t | z_{t-1}; A)\prod_{t=1}^{T} p(x_t | z_t; \mu).$$

Метою ймовірнісного синтезу є генерація прихованих станів з змінних спостереження у випадку

фіксованих параметрів θ моделі. Цей процес передбачає обчислення апостеріорної ймовірності $p(z_Q|x;\theta)$ на множині змінних прихованих станів z_Q. Іншою формою генерації є генерація з використанням оцінки апостеріорної ймовірності, $\mathbf{z}^{MAP} = \arg\max_{\mathbf{z}} p(\mathbf{z}|\mathbf{x};\theta) = \arg\max_{\mathbf{z}} p(\mathbf{x},\mathbf{z};\theta)$.

В процесі генерації ПММ можна виділити наступні завдання:

- Згладжування: обчислення відособленої ймовірності окремого прихованого стану $p(z_t|x_{1:T})$ для $t<T$.
- Фільтрація: обчислення $p(z_t|x_{1:e})$ для процесу генерації онлайн.
- Прогнозування: обчислення $p(z_t|x_{1:T})$ для $t>T$ для прогнозування майбутніх станів.
- Генерація з використанням оцінки апостеріорної ймовірності: обчислення найбільш ймовірної послідовності $z_{1:T}^{MAP} = \arg\max_{z_{1:T}} p(z_{1:T}|x_{1:T})$.

Для обчислення апостеріорних ймовірностей використовується алгоритм прямого-зворотного ходу. За цим алгоритмом рекурсивно обчислюються величини

$$\alpha_t(i) = p(z_t = i, x_1, \dots x_t),$$
$$\beta_t(i) = p(x_{t+1}, \dots, x_T | z_t = i).$$

Якщо встановити $\alpha_1(i) = \pi_i p(x_1|z_1 = i;\mu_i)$, інші α_t обчислюються прямою рекурсією

$$\alpha_{t+1}(j) = p(z_{t+1} = j, x_1, \dots x_{t+1}) =$$
$$= \sum_j p(z_t = i, z_{t+1} = j, x_1, \dots, x_{t+1}) =$$
$$= \sum_j p(z_t = i, x_1, \dots, x_t) p(z_{t+1} = j | z_t = i) p(x_{t+1} | z_{t+1} = j) =$$
$$= \sum_j \alpha_t(i) A_{ij} p(x_{t+1} | z_{t+1} = j;\mu_j) = .$$

Аналогічно, якщо $\beta_T(i) = 1$, β_t обчислюються зворотною рекурсією

$$\beta_t(i) = \sum_j p(x_{t+1}, \dots, x_T | z_t = i, z_{t+1} = j) p(z_{t+1} = j | z_t = i) =$$
$$\beta_t(i) = \sum_j A_{ij} p(x_{t+1} | z_{t+1} = j;\mu_j) \beta_{t+1}(j).$$

Враховуючи обчислені α та β, обчислюються ймовірності

$$p(z_t = i | x_{1:T}) = \frac{p(z_t = i, x_{1:t}) p(x_{t+1:T} | z_t = i)}{p(x_{1:T})} = \frac{1}{Z} \alpha_t(i) \beta_t(i) = ,$$
$$p(z_t = i, z_{t+1} = j | x_{1:T}) = \frac{1}{Z} \alpha(i) A_{ij} p(x_{t+1} | z_{t+1} = j;\mu_j) \beta_{t+1}(j),$$
$$p(z_t = i | x_{1:t}) = \frac{1}{Z} \alpha_t(i), \; p(x_{1:T}) = \sum_i \alpha_T(i).$$

де Z – нормалізована константа рівняння.

Відповідно до оцінки апостеріорної ймовірності $z_{1:T}^{MAP} = \arg\max_{z_{1:T}} p(z_{1:T}|x_{1:T})$ використовують рекурсивну процедуру схожу до прямої рекурсії, в якій Σ_i заміщується $\max_i$ (алгоритм Вітербі). Визначається

$$\gamma_t(i) = \max_{z_1, \dots, z_{t-1}} p(z_1, \dots, z_{t-1}, z_t = i, x_1, \dots, x_t).$$

Якщо встановити $\gamma_1(i) = \pi_i p(x_1|z_1 = i;\mu_i)$, то рекурсія

$$\gamma_{t+1}(j) = \max_i \gamma_t(i) A_{ij} p(x_{t+1} | z_{t+1} = j;\mu_j).$$

Розподіл ймовірностей виходів відповідає наявним частотним компонентам перетворення. Функція ймовірності розглядається у формі $p(x;\mu) = h(x)\exp(-D_{KL}(x\|\mu))$, де $D_{KL}(x\|\mu) = \sum_i x_i \log x_i / y_i$ - відстань Кульбака-Лейблера.

Одним із стандартних підходів, що використовуються в прихованих моделях, є метод максимальної правдоподібності. Для будь-якого розподілу ймовірностей q прихованих величин:

$$\log p(\mathbf{x}:\theta) = \log \sum_{\mathbf{z}} q(\mathbf{z}) \frac{p(\mathbf{x},\mathbf{z};\theta)}{q(\mathbf{z})} \geq \sum_{\mathbf{z}} \log q(\mathbf{z}) \frac{p(\mathbf{x},\mathbf{z};\theta)}{q(\mathbf{z})}.$$

На Е-кроці беруть $q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x};\theta)$, потім на М-кроці визначають максимум нижньої границі відносно θ, що еквівалентно визначенню максимуму $\mathrm{E}_{\mathbf{z}\sim q}[\log p(\mathbf{x},\mathbf{z};\theta)]$. В ПММ на Е-кроці визначається

апостеріорна ймовірність для кожного стана та кожної пари станів в прихованій послідовності, використовуючи метод прямого-зворотного ходу, и ці величини потім використовуються на М-кроці оновлення параметрів. У випадку незалежних спостережень, можна використати факторний розподіл $q(z) = \prod_t q_t(z_t)$.

Відповідно до інкрементного ЕМ-алгориму для ПММ, необхідно інкрементно оновлювати розподіл q з кожним нових спостереженням. Використовуючи умовні залежності ПММ, розподіл визначається як $q(z_{1:T}) = q_1(z_1|x_{1:T})\prod_{t\geq 2} p(z_t|z_{t-1}, x_{1:T})$. Можна обмежити розподіл q до форми $q(z_{1:T}) = q_1(z_1)\prod_{t\geq 2} q(z_t|z_{t-1})$, з $\sum_j q_t(j|i)$ для всіх t та i. В даному випадку q границі по z_t визначаються за виразом $\phi_t(z_t) = \sum_{z_{1:t-1}} q_1(z_1)\ldots q_t(z_t|)$, який може бути обчислений інкрементно з $\phi_1(i) = q_1(i)$ та $\phi_t(j) = \sum_i \phi_{t-1}(i) q_t(j|i)$для $t \geq 2$. Далі границя береться у формі $q(z_{t-1}, z_t) = \phi_{t-1}(z_{t-1}) q(z_t|z_{t-1})$. Тоді отримуємо нижню границю правдоподібності

$$\hat{f}_T(\theta) = \mathrm{E}\left[\log \frac{p_\theta(x_{1:T}, z_{1:T})}{q(z_{1:T})}\right] = \sum_{t=1}^{T}\left[\log \frac{p_\theta(x_t, z_t|)z_{t-1}}{q_t(z_t|z_{t-1})}\right] = \sum_{t=1}^{T}\sum_{z_{t-1}, z_t} \phi_{t-1}(z_{t-1}) q_t(z_t|z_{t-1}) \log \frac{p_\theta(x_t, z_t|)z_{t-1}}{q_t(z_t|z_{t-1})}.$$

Відповідно до розширення до напівприхованої марківської моделі необхідно параметризувати моделі з двома прихованими величинами, станом поточного сегменту, z_t, та лічильником часових кроків з початку сегменту z_t^D. В цьому випадку ймовірності переходів визначають за виразами

$$p\left(z_t = j|z_{t-1} = i, z_t^D = d\right) = \begin{cases} A_{ij}, & d = 1, \\ \delta(i, j), & \textit{інакше}, \end{cases}$$

$$p\left(z_t^D = d'|z_{t-1} = i, z_{t-1}^D = d\right) = \begin{cases} \lambda_i(d), & d' = d+1, \\ 1-\lambda_i(d), & d' = 1, \\ 0 & \textit{інакше}. \end{cases}$$

Якщо $\lambda_i(d) = D_i(d+1)/D_i(d)$, де $D_i(d) := \sum_{d'\geq d} p_i(d')$, тоді апріорна ймовірність отримання сегмента довжини як мінімум d буде рівна $\lambda_i(1)\ldots\lambda_i(d-1) = D_i(d)$, і отримання сегмента з довжиною d буде $\lambda_i(1)\ldots\lambda_i(d-1)(1-\lambda_i(d)) = p_i(d)$.

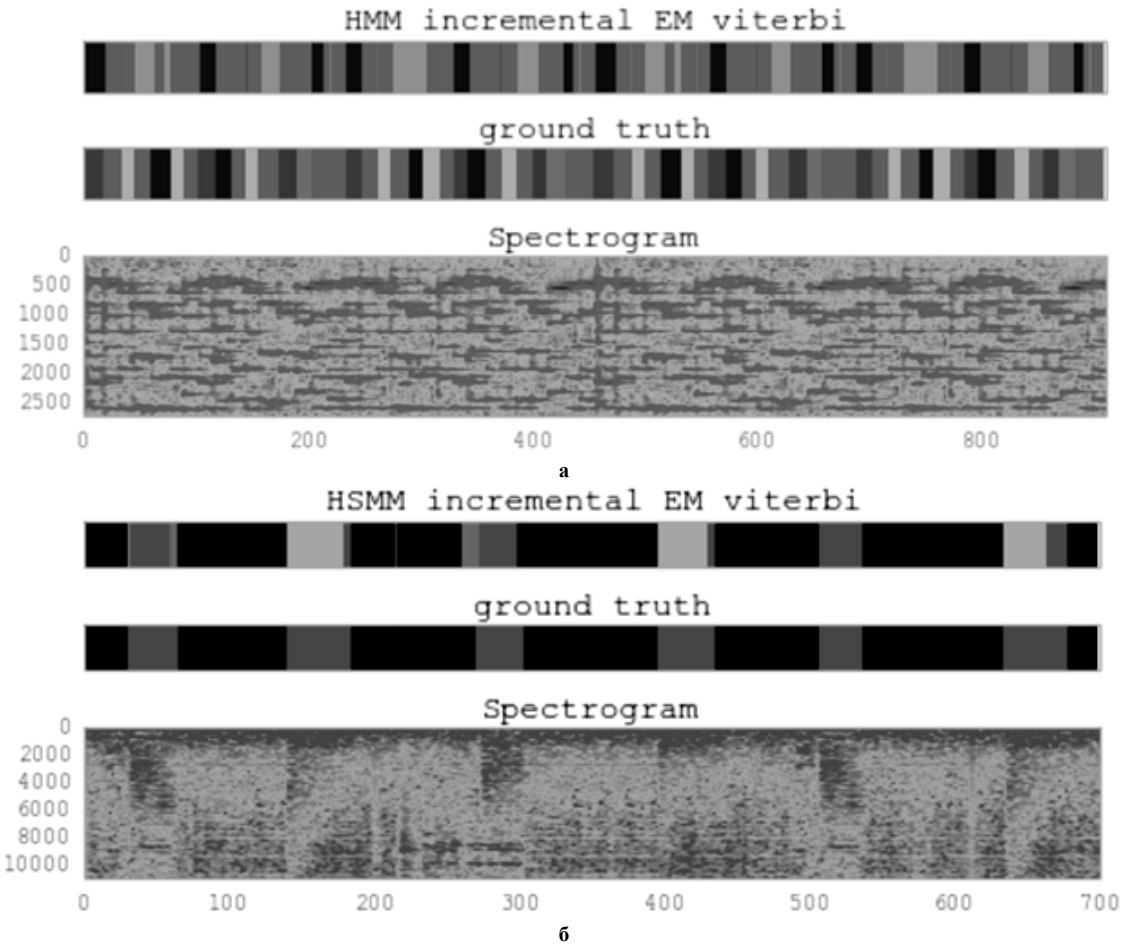
Результати сегментації аудіо сигналу. Розглянутий алгоритм було застосовано для акустичної сегментації аудіо сигналу, який містить музику (рис. 2,а). Було використано віконне перетворення Фур'є, частота дискретизації 22,05 кГц, використано вікно Хаммінга розміру 2048 з зсувом на 256, число частотних компонент 1024. Також сегментацію було застосовано до датасету Office Live Dataset (рис. 2,б), метою було сегментувати різні за характером звуки, такі як звуки падінь, скрипи і тому подібне. Ці звукові сигнали не завжди є однорідні, тому ставилося за мету виділити однорідні сегменти для подальшого пошуку звукових паттернів. Дані по якості сегментації та кластеризації наведені в табл. 1.

Таблиця 1

Дані кластерної належності

Тип аудіо сигналу	Відносна кластерна належність, %	Розрахована кластерна належність, %
Музика	74,43	87,93
Нерівномірний звуковий сигнал	76,02	89,37

Висновки. Кластерний аналіз невідомих аудіо даних є складним завданням, одним з таких, що не має загального рішення. В статті запропоноване рішення на основі прихованих марківської та напів-марківської моделей, використання яких дозоляє об'єднати процеси сегментації та кластеризації. Сходимість навчання моделей потребує подальшого дослідження. Даний метод є цілковито підходом навчання без вчителя, тому дозволяє сегментувати та кластеризувати аудіодані, про які немає попередньої інформації, ґрунтуючись на статистичних характеристиках.



а

б

Рис. 2. Результати сегментації аудіо сигналу, (а) – музика, (б) – звукові сигнали

Література

1. R. Sinha, S.E. Tranter, M.J.F. Gales, P.C. Woodland, The Cambridge University March 2005 speaker diarisation system, in: Proceedings of the European Conference on Speech Communication and Technology, Lisbon, Portugal, September 2005, pp. 2437–2440.

2. W.H. Tsai, S.S. Cheng, H.M. Wang, Speaker clustering of speech utterances using a voice characteristic reference space, in: Proceedings of the International Conference on Spoken Language Processing, Jeju Island, Korea, October 2004.

3. D. Liu, F. Kubala, Online speaker clustering, in: Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, Montreal, Canada, May 2004, pp. 333–336.

4. S.S. Chen, P.S. Gopalakrishnan, Clustering via the Bayesian information criterion with applications in speech recognition, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, Seattle, USA, May 1998, pp. 645–648.

5. S. Meignier, D. Moraru, C. Fredouille, J.F. Bonastre, L. Besacier, Step-by-step and integrated approaches in broadcast news speaker diarization, Comput. Speech Language 20 (2–3) (April–July 2006) 303–330.

6. V. Wan, W.M. Campbell, Support vector machines for speaker verification and identification, in: Proceedings of the Neural Networks for Signal Processing, vol. 10, Sydney, Australia, December 2000, pp. 775–784.

Рецензія/Peer review : 22.10.2018 р. Надрукована/Printed :20.11.2018 р.
Рецензент: д.т.н., проф. Полікаровських О.І.