УДК 621.391.83

Суб’єктивне оцінювання розбірливості зашумленої мови в лекційному приміщенні

Андрійченко О. О.,
e-mail oleksiy.andriichenko@gmail.com
Денисенко О. І.,
e-mail reiden1998@gmail.com
Факультет Електроніки
КПІ ім. Сікорського
Київ, Україна

***Анотація*—При розв’язанні задач проектування споруд та звукоізоляції, кімнату можна розглядати як особливий фільтр, котрий впливає на розбірливість мови двома способами. Перший – пізня реверберація, що виконує функцію шумової завади; другий – ранні відбиття, навпаки, збільшують розбірливість. Тип прослуховування також має значення – бінауральний, на відміну від монаурального, збільшує розбірливість мови, спотвореної шумом та реверберацією. У даній роботі проведено суб'єктивну оцінку впливу характеристик лекційного залу на розбірливість зашумленого мовлення під час бінаурального прослуховування. При цьому однією із задач було спростити експеримент так, щоб учасники могли виконати його самостійно, без використання коштовної техніки.**

Ключові слова — розбірливість мови; суб’єктивна оцінка; бінауральна імпульсна характеристика; білий шум; пізня реверберація; ранні відбиття.

I. Вступ

Суб'єктивна оцінка в режимі бінаурального прослуховування розбірливості (зрозумілості) мови, спотвореної шумом і реверберацією, має значний науковий і практичний інтерес. Така оцінка є важливою для вдосконалення математичних моделей слухової системи людини, калібрування систем об'єктивної оцінки розбірливості мовлення, а також сертифікації каналів і пристроїв зв'язку, приміщень, слухових апаратів, кохлеарних імплантатів тощо [1-14].

Ідея енергетичних співвідношень ранніх та пізніх (early-to-late) відбиттів була розроблена в [2,3], що дозволило врахувати негативні наслідки як пізньої реверберації, так і фонового шуму. У роботі [4,5] було показано, що ця концепція може бути використана для достовірного прогнозування розбірливості мовлення в широкому діапазоні реальних приміщень.

Кількісні оцінки ступеня впливу шуму і реверберації на розбірливість мови в аудиторіях для режиму бінаурального прослуховування наведено в [6–9]. Експериментально показано, що шумові перешкоди є більш небезпечними, ніж реверберація, через близькість джерел шуму у вигляді студентів, що сидять поруч та розмовляють, а також через подібність інтерференційних та сигнальних спектрів. При цьому час реверберації в лекційних кімнатах рідко перевищує 0.9 с, тому пізня реверберація має слабкий ефект маскування через відносно низьку інтенсивність.

Результати автоматизованої суб'єктивної оцінки розбірливості української мови були представлені в роботі [10], де мовні склади типу приголосний-голосний-приголосний (CVC) прослуховувалися двома способами: через навушники (діотичне прослуховування) і через комп'ютерні колонки. Оцінки розбірливості виявилися майже однаковими, хоча для випадку колонок результати оцінювання розбірливості мовлення були трохи вищими (на 1-3%). Було висловлено припущення, що найбільш імовірною причиною виявленого збігу є те, що відстань між комп'ютерними динаміками і слухачем зазвичай не перевищує 0,6–0,8 м, тому ранні відбиття практично не впливають на розбірливість мовлення. У [10] було запропоновано спробувати тестування артикуляції на більшій відстані між слухачами та комп'ютерними динаміками, де ефект ранніх відбиттів є більш помітним. Дійсно, в [11] було виявлено, що ранні відбиття можуть збільшити SNR до 6 дБ, хоча в [12] була поставлена під сумнів можливість такого значного збільшення. Що стосується дії бінаурального прослуховування, то в [12] показано, що воно може збільшити SNR до 2-3 дБ. У той же час можна припустити, що існує небезпека послаблення позитивного ефекту ранніх відбиттів [1].

Метою даної роботи є суб’єктивне оцінювання розбірливості мови, спотвореної спільною дією шуму та реверберації, шляхом бінаурального прослуховування сигналів на різних відстанях між динаміком та слухачем. На відміну від інших подібних експериментів, в даній роботі використано бюджетний варіант апаратно-програмного обладнання, за допомогою якого кожен учасник досліджень міг самостійно вдома за допомогою навушників прослуховувати та фіксувати дані. Це допомогло збільшити кількість учасників експерименту та підвищити точність

отриманих результатів. Зрештою, таке дослідження повинно поліпшити розуміння розглянутої проблеми і, як наслідок, сприяти підвищенню точності прогнозування розбірливості мовлення в приміщеннях.

II. ОРГАНІЗАЦІЯ ЕКСПЕРИМЕНТАЛЬНИХ ДОСЛІДЖЕНЬ

Запис чіткої мови здійснювався в заглушеному приміщенні за участі дев'яти чоловіків та двох жінок віком 20-21 рік, без вад слуху. Засоби запису та устаткування: звукова карта PRESONUS AudioBox USB, мікрофон Superlux ECM 999, програмний пакет Audacity. Запис відбувався з частотою дискретизації 44,1 кГц та глибиною квантування 16 біт. Також використовувалась так звана "несуча фраза" для відтворення тексту диктором. Наприклад, склад "ток" прочитувався як "Запишіть ток тепер".

Синтез аудіофайлів для подальшого аналізу розбірливості відбувався в два етапи. Перший етап – додавання білого шуму до чистої мови (шум зважувався відповідними коефіцієнтами для отримання необхідного значення SNR). Другий етап – фільтрування отриманої адитивної суміші двохканальним цифровим нерекурсивним фільтром з використанням бінауральної імпульсної характеристики приміщення (були використані бінауральні IX для відстаней 2.25-10.2 м від джерела звуку). Отримані сигнали прослуховувались через навушники та фіксувалися слухачами за допомогою клавіатури комп'ютера.

A. Модель сигналу

Модель

$$x(t) = s(t) \otimes h(t) + n(t) \qquad (1)$$

де $\otimes$ є символом згортки, зазвичай використовується для аналітичного опису комбінованого впливу навколишнього шуму $n(t)$ і реверберації на чистий мовний сигнал $s(t)$, $h(t)$ - імпульсна характеристика приміщення.

Модель (1) є досить добре наближеною до реальної ситуації та зручною при розробці алгоритмів придушення шуму та реверберації [15,16]. Однак використання цієї моделі пов'язане із проблемою правильного визначення SNR, оскільки імпульсна характеристика $h(t)$ може бути представлена як

$$h(t) = h_e(t) + h_l(t)\ ,$$

$$h_e(t) = \begin{cases} h(t), t \in 0...50мс; \\ 0,\ \text{t} \notin 0...50\text{мс}, \end{cases}$$

$$h_l(t) = \begin{cases} h(t), t > 50мс; \\ 0,\ \text{t} \le 50\text{мс}, \end{cases}$$

де $h_e(t)$ є початковою частиною $h(t)$, що збільшує прямий сигнал $s(t)$, а $h_l(t)$ є хвостовою частиною $h(t)$, яка діє аналогічно шуму $n(t)$. Щоб обійти цю проблему, в [3] було використано величину "direct speech to noise ratio"

$$SNR = 10\lg D_s / D_n \qquad (2)$$

де D_s і D_n – дисперсії чистого сигналу $s(t)$ та шуму $n(t)$ відповідно. У [10] було запропоновано врахувати підсилення прямого сигналу за рахунок дії ранніх відбиттів: $SNR_e = 10\lg D_{s_e} / D_n$ де D_{s_e} є дисперсією підсиленого сигналу $s_e(t) = s(t) \otimes h_e(t)$.

В даній роботі використано іншу модель сигналу у вигляді згортки бінаурального імпульсного відгуку приміщення із адитивною сумішшю чистого мовлення та шуму:

$$x(t) = [s(t) + n(t)] \otimes h(t) \qquad (3)$$

Модель (3) видається природною, якщо ми хочемо розглядати приміщення як вид шумопоглинаючого фільтра з імпульсною характеристикою $h(t)$. Тому в даній роботі буде використано визначення (2), яке виглядає логічно в рамках моделі (3).

B. Програмне забезпечення для автоматизованого тестування

Під час експерименту слухачу необхідно прослухати файл, ідентифікувати склад та зафіксувати його за допомогою клавіатури. За основу були взяті склади зі списку стандарту ГОСТ Р50840-95 [17], та розроблено дев'ять артикуляційних таблиць української мови. Слухачам пропонувалось прослухати три таблиці складів.

Процес тестування містить п'ять етапів:

- симуляція спотвореної шумом і реверберацією мови;
- озвучення спотворених сигналів;
- прослуховування сигналів;
- фіксація сприйнятого складу;
- розрахунок розбірливості за результатами сприйняття.

Склади відтворювались у випадковому порядку. Також для виключення помилок, пов'язаних із неправильним введенням тексту, було можливим виправлення введеного з клавіатури результату прослуховування.

C. Бінауральні імпульсні характеристики приміщення

При тестуванні було використано шість бінауральних імпульсних характеристик приміщень з бази даних Aachen Impulse Response database [18,19]. Ці імпульсні характеристики належать лекційній аудиторії із розмірами 10.8×10.9×3.15 м, із трьома вікнами, бетонними стінами, паркетною підлогою та дерев'яними столами та стільцями у якості меблів. Гучномовець був розташований на трибуні аудиторії, а мікрофони розташовувались на різних столах на відстані d від гучномовця. Час реверберації RT60 для вимірюваних позицій наведено в табл.1.

Форми хвиль пари бінауральних імпульсних характеристик для дистанцій 2.25 та 10.2 м показані на графіках Рис. 1a та 1b, відповідно.

Таблиця.1.

d, м	2.25	4.00	5.56	7.10	8.68	10.2
RT_{60}, с	0.70	0.72	0.79	0.80	0.081	0.83

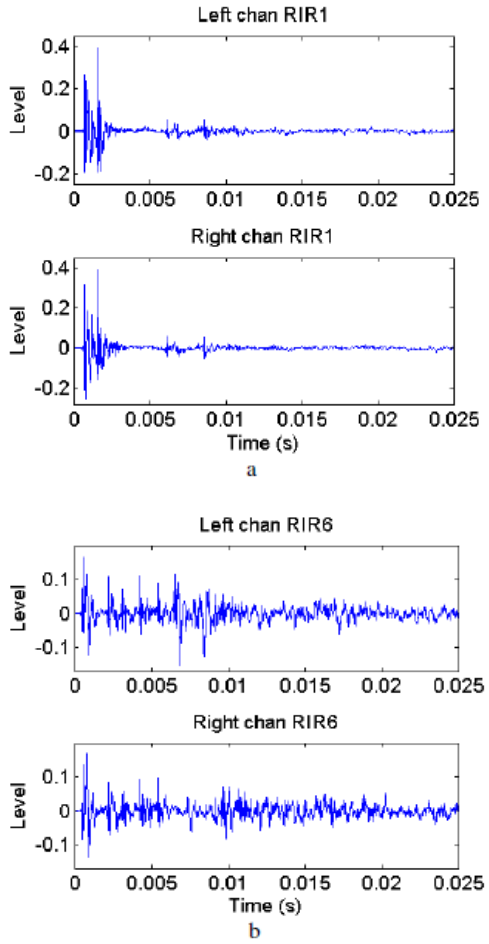

Рис. 1. Форми хвиль бінауральних імпульсних характеристик для d=2.25 м (a) та 10.2 м (b)

Співвідношення енергій ранніх та пізніх відбиттів звуку (C_{50}) для всіх шести імпульсних характеристик наведено на графіку Рис. 2.

Графік Рис. 2 дозволяє зробити висновок, що усі шість використаних бінауральних IX здатні забезпечити гарну розбірливість мови, оскільки відомо, що розбірливість складів є не меншою за 80% (відповідно розбірливість фрази є не нижчою 95%) за умови $C_{50} \geq$ -2 дБ.

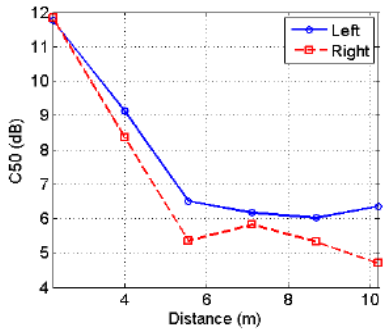

Рис. 2. Значення C_{50} для бінауральних імпульсних характеристик

Крім того, видно, що значення C_{50} мало змінюється для відстаней 5-10 м, тобто дія ранніх відбиттів на розбірливість мови є майже однаковою для цих відстаней.

III. Результати експериментальних досліджень

Усереднені за п'ятнадцятьма слухачами результати оцінювання складової розбірливості мови наведено на графіку Рис. 3a. Оцінка стандартного відхилення цих результатів наведена на графіку Рис. 3b. Видно, що для значень SNR = -10…5 дБ розбірливість мови для відстаней 7-10 м є вищою, ніж для малих відстаней (2.25 м).

Отже, можна сказати, що розбірливість мови покращується при двох факторах - малих значеннях SNR та великій відстані від слухача до джерела, що можна трактувати як результат позитивної дії ранніх відбиттів. Для середніх та великих значень SNR розбірливість є кращою на малих відстанях, що можна пояснити негативною дією пізньої реверберації. До речі, ранні відбиття негативно впливають на спектр мовних сигналів, хоча цей ефект частково компенсується при бінауральному прослуховуванні [1].

IV. Обговорення результатів

Зростання розбірливості зі збільшенням дистанції до диктора можна пояснити поєднанням сприятливого ефекту ранніх відбиттів та бінауральним прослуховуванням. Цікаво порівняти ці результати з оцінкою розбірливості, наведеною в [10], де зашумлені склади прослуховувались через навушники (в діотичному режимі) та комп'ютерний динамік (Рис. 4). При порівнянні графіків Рис. 3 та синіх кривих Рис. 4, можна помітити деяке протиріччя: значення розбірливості у випадку відсутності ранніх відбиттів та бінаурального прослуховування (Рис. 4a) є вищою за таку для випадку присутності обох факторів (Рис. 3a).

Причину цього протиріччя можна пояснити наступним чином. Значення SNR в [10] розраховувалось за формулою $SNR_e = 10\lg D_{s_e} / D_n$, де D_{s_e} – це дисперсія сигналу $s_e(t) = s(t) \otimes h_e(t)$, підсиленого ранніми відбиттями. В окремому випадку при виключній дії шуму, моделі (1) та (3) однакові та $SNR_e = SNR = 10\lg D_s / D_n$. Значення розбірливості, показані на графіку Рис. 4b особливо не відрізняються від значень, показаних на графіку Рис. 4a через малі значення RT_{60} для домашніх помешкань (0.3 с) та малі відстані (0.6-0.8 м), які є близькими до критичної відстані. Таким чином, негативна дія пізньої реверберації може розглядатись як головна причина порівняно низького рівня розбірливості, показаного на Рис. 3a. Тому в майбутніх дослідженнях доцільно кількісно оцінити ступінь впливу ранніх відбиттів на розбірливість.

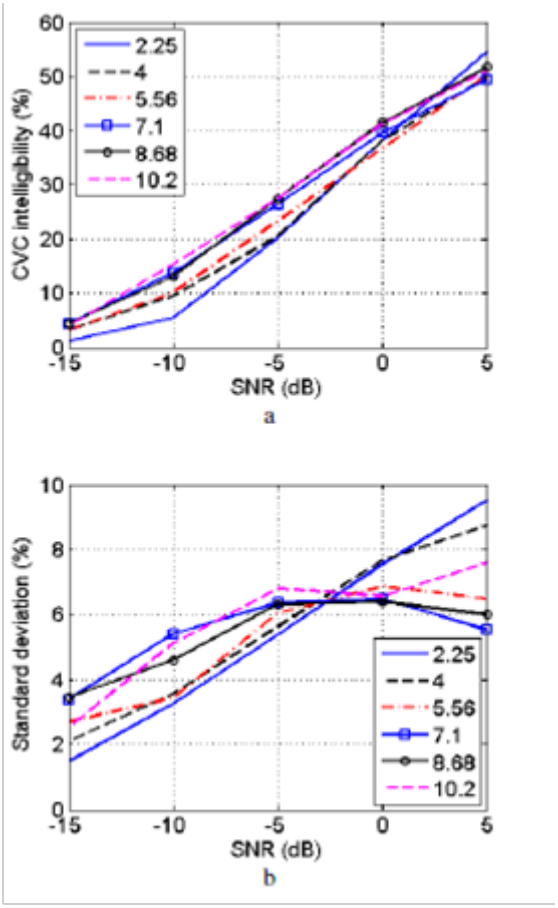

Рис. 3. Середня оцінка розбірливості (a) та нормальне відхилення (b)

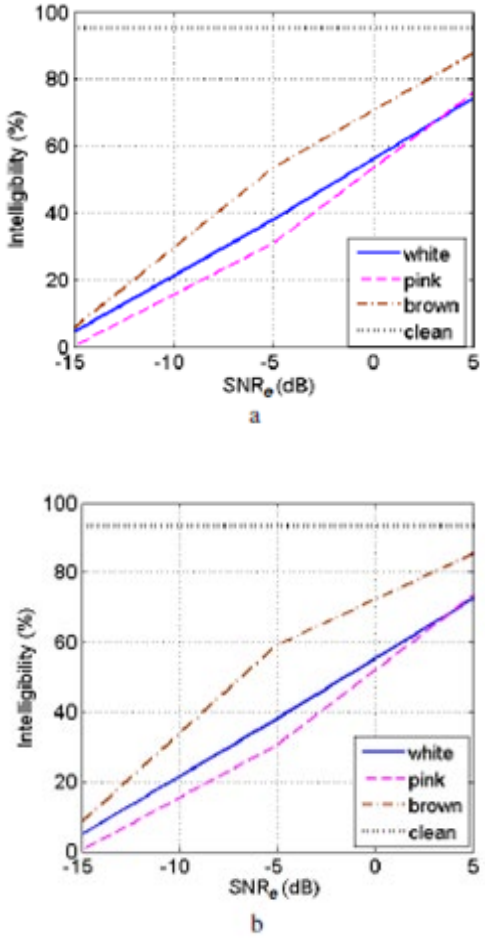

Рис. 4. Оцінка розбірливості для навушників (a) та динаміка комп'ютера (b)

Також необхідно вказати ще на один небажаний фактор, що може призвести до низьких значень розбірливості, наведених на Рис. 3a. Цим фактором є мимовільне запам'ятовування складів слухачами, що зрештою призводить до завищених оцінок розбірливості в пізніших прослуховуваннях. Цей ефект може бути нейтралізованим за допомогою зміни артикуляційних таблиць на протязі експерименту або шляхом випадкової зміни відстані до динаміка. Очевидно, зазначений висновок доцільно врахувати при майбутніх дослідженнях.

Висновки

За допомогою бінаурального прослуховування CVC-складів виконано суб'єктивне оцінювання розбірливості зашумленої мови в приміщенні. Таке оцінювання здійснювався методом комп'ютерного моделювання мови, що дозволяв змоделювати мовний сигнал в середній за розмірами лекційній аудиторії на різних відстанях від диктора.

Показано, що розбірливість мови для великих відстаней (7-10 м) є приблизно на 10% більшою, ніж для малих відстаней (2.25 м) для випадку малого відношення сигнал-шум (SNR = -10…-5 дБ). Значний приріст розбірливості при збільшенні відстані може бути пояснений поєднанням сприятливої дії ранніх відбиттів та бінаурального прослуховування. Розбірливість для середніх та великих відстаней (4-10 м) виявилась меншою, ніж розбірливість на малих відстанях (2.25 м) для випадку великих відношень сигнал-шум (SNR > 3-5 дБ). Цей факт можна пояснити дією пізньої реверберації, що переважає над дією шуму для середніх та великих значень SNR.

Отримані результати мають бути уточненими при майбутніх дослідженнях шляхом нейтралізації ефекту запам'ятовування складів. Крім того, доцільно окремо оцінити вплив ранніх відбиттів на розбірливість мови.

Перелік посилань

[1] J. Blauert Ed., The technology of binaural listening. Springer, Berlin–Heidelberg–New York, 2013.

[2] J. Lochner and J. Burger, "The influence of reflections on auditorium acoustics," Journal of Sound and Vibration, No. 1, pp. 426-454, 1964.

[3] G. Soulodre, N. Popplewell, and J. Bradley, "Combined effects of early reflections and background noise on speech intelligibility," Journal of Sound and Vibration, vol. 135, No.1, pp. 123-133, 1989.

[4] J. Bradley, “Predictors of speech intelligibility in rooms,” J. Acoust. Soc. Am., vol 80, pp. 837–845, 1986.

[5] J. Bradley, “Speech intelligibility studies in classrooms,” J. Acoust. Soc. Am., vol. 80, pp. 846–854, 1986.

[6] J. Bradley, R. Reich, and S. Norcross, “On the combined effects of signal-to-noise ratio and room acoustics on speech intelligibility,” J. Acoust. Soc. Am., vol. 106 (4), Pt. 1, pp. 1820-1828, October 1999.

[7] H Sato and J. Bradley, “Evaluation of acoustical conditions for speech communication in working elementary school classrooms,” J. Acoust. Soc. Am. 106 (4), Pt. 1, pp. 2064–2077, 2004.

[8] J Bradley and H. Sato, “Speech intelligibility test results for grades 1, 3 and 6 children in real classrooms,” Proc. of ICA, Kyoto, 2004.

[9] W. Yang and J. Bradley, “Effects of room acoustics on the intelligibility of speech in classrooms for young children,” J. Acoust. Soc. Am., vol. 125 (2), pp. 922–933, 2009.

[10] A. Prodeus, K. Bukhta, P. Morozko, O. Serhiienko, I. Kotvytskyi, O. Dvornyk, “Automated Subjective Assessment of Speech Intelligibility in Various Listening Modes,” Microsystems, Electronics and Acoustics, vol. 23, no. 3, pp.49-57, 2018.

[11] J. Bradley, H. Sato, and M. Picard, “On the importance of early reflections for speech in rooms,” J. Acoust. Soc. Am., vol. 113, no. 6, pp. 3233-3244, June 2003.

[12] I. Arweiler, J. Buchholz, and T. Dau, “Speech intelligibility enhancement by early reflections,” Proc. of 2nd Int. Symposium on Auditory and Audiological Research (ISAAR 2009), Elsinore, Denmark, August 2009.

[13] S. Naida, O. Pavlenko, “Coupled Circuits Model in Objective Audiometry,” Proc. of the 2018 IEEE 38th International Conference on Electronics and Nanotechnology (ELNANO), pp. 281-286, April 24-26, Kyiv, Ukraine, April 2018.

[14] S. Naida, O. Pavlenko, “Newborn Hearing Screening Based on the Formula for the Middle Ear Norm Parameter,” Proc. of the 2018 IEEE 38th International Conference on Electronics and Nanotechnology (ELNANO), pp. 287-291, April 24-26, Kyiv, Ukraine, April 2018.

[15] J. Benesty, Y. Huang, and J. Chen, Wiener and Adaptive Filters. In Springer Handbook of Speech Processing, J. Benesty, M. Sondhi, and Y. Huang, Eds. Springer-Verlag Berlin Heidelberg, 2008, pp. 103-120.

[16] E. Habets, N. Gaubitch, and P. Naylor, “Temporal selective dereverberation of noisy speech using one microphone,” Proc. of 2008 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 4577-4580, March-April 2008.

[17] A. Prodeus, K. Bukhta, P. Morozko, O. Serhiienko, I. Kotvytskyi, I. Shherbenko, “Automated System for Subjective Evaluation of the Ukrainian Speech Intelligibility,” Proc. of IEEE 38th Int. Conf. on Electronics and Nanotechnology (ELNANO), pp. 533-538, April 24- 26, Kyiv, Ukraine, 2018.

[18] M. Jeub, M. Schäfer, and P. Vary, “A binaural room impulse response database for the evaluation of dereverberation algorithms,” In Int. Conf. Proc. on Digital Signal Processing (DSP), Santorini, Greece, 2009.

[19] Aachen Impulse Response Database. Available on-line: https://www.iks.rwth-aachen.de/en/research/tools-downloads/databases/aachen-impulse-response-database

[20] W. Ahnert, W. Schmidt, Fundamentals to perform acoustical measurements. Appendix to EASERA. Berlin, 2005.

УДК 621.391.83

Субъективное оценивание разборчивости зашумленной речи в лекционном помещении

Андрейченко А. О.,
e-mail oleksiy.andriichenko@gmail.com
Денисенко А. И.,
e-mail: reiden1998@gmail.com
Факультет Електроники
КПИ им. Сикорского,
Киев, Украина

***Аннотация*—При решении задач проектирования сооружений и звукоизоляции, комнату можно рассматривать как некоторый фильтр, который влияет на разборчивость речи двумя способами. Первый – поздняя реверберация, выполняющая роль шумовой помехи; второй – ранние переотражения, наоборот, увеличивают разборчивость. Тип прослушивания также имеет значение – бинауральный, в отличии от монаурального, увеличивает разборчивость речи, искаженной шумом и реверберацией. В данной работе проведено субъективное оценивание влияния характеристик лекционного зала на разборчивость зашумленной речи во время бинаурального прослушивания. При этом одной из целей было упростить эксперимент так, чтобы участники могли выполнить его самостоятельно, без использования дорогостоящего оборудоания.**

Ключевые слова — разборчивость речи; субъективная оценка; бинауральная импульсная характеристика; белый шум; поздняя реверберация; ранние переотражения.

UDC 621.391.83

Subjective Assessment of the Intelligibility of Noised Speech in Lecture Room

O. O. Andriichenko,
e-mail oleksiy.andriichenko@gmail.com
O. I. Denysenko,
e-mail reiden1998@gmail.com
Faculty of Electronics
Igor Sikorsky Kyiv Polytechnic Institute,
Kyiv, Ukraine

***Abstract*—Subjective assessment of the speech intelligibility is of great practical interest, since this parameter can be used in many engineering and natural-mathematical fields, such as bioengineering, design of lecture rooms and concert halls, mathematical modeling and medicine. By this time, several attempts were made to assess the intelligibility of the room. The results of such experiments became conclusions: 1) early reflections improve the intelligibility of distorted noise and reverb speech; 2) using binaural type of listening we will have better intelligibility than with monoural. But to re-implement such experiments, it is necessary to use a lot of expensive equipment, which is not always possible. Therefore, the purpose of this work is to study the influence of the characteristics of the room on the speech intelligibility without the use of a large number of equipment. The idea of the experiment is as follows: distorted by noise and reverb signals must be listened to by the participants in the experiment, the perceived sound is fixed and compared with the undistorted signal. The clear signal was recorded using a microphone, sound card and audio file software. Synthesis of distorted signals took place in two stages: 1) adding white noise; 2) filtering the resulting mix through a non-recursive digital filter using the impulse characteristics of the room for six distances to the source of sound. The mathematical model of the distorted (output) signal was the convolution of an additive mix of clear signal and white noise with a pulse characteristic. To ensure the necessary signal-to-noise ratio, the noise was weighed by the corresponding coefficients at the stage of adding it to the clear signal. Testing was carried out in five stages: 1) simulating of distorted signals; 2) voicing signal to participant; 3) signal listening; 4) fixation of the perceived sound; 5) intelligibility calculating. For testing, six binaural impulse characteristics of the room with known parameters from the Aachen database of impulse characteristics were used. Each of the impulse characteristics corresponds to a certain value of the distance from the sound source to the microphone. After the completion of the experiment, the overall result was obtained by averaging by the number of participants. The overall result has shown that for small values of the signal-to-noise ratio over long distances, the intelligibility is greater than for small distances. For high values, the signal-to-noise ratio is better for small distances. Such results may be explained by the fact that, at long distances, the combination of the effects of early reflections and binaural listening positively affects the intelligibility. The main reason for the low intelligibility of the distorted by noise speech in the room is a late reverberation. As a conclusion we can say that the computer modeling method makes it easy to create a distorted by the noise and influence of the room signal; It is shown that, for small values, the signal-to-noise ratio is more readable than for small distances.**

Keywords – speech intelligibility; subjective assessment; binaural impulse response; white noise; late reverberations; early reflections.