==========

Ғылымның, білімнің және бизнестің интеграциясы
Интеграция науки, образования и бизнеса
Integration of science, education and business

==========

DOI 10.53364/24138614_2022_26_3_49
УДК 004.93

[1]Савостин А.А., [2]Савостина Г.В.
[1,2]Северо-Казахстанский университет им. М. Козыбаева, г. Петропавловск, Республика Казахстан

[1]E-mail: alexey.savostin@gmail.com
[2]E-mail: gvshubina@ku.edu kz

ПОДХОД К АВТОМАТИЧЕСКОМУ РАСПОЗНАВАНИЮ ЭМОЦИЙ ДИКТОРА ПРИ ПОМОЩИ ИНТЕЛЛЕКТУАЛЬНЫХ МЕТОДОВ АНАЛИЗА ИНФОРМАЦИИ

АҚПАРАТТЫ ТАЛДАУДЫҢ ЗИЯТКЕРЛІК ӘДІСТЕРІН ҚОЛДАНА ОТЫРЫП, ДИКТОРДЫҢ ЭМОЦИЯЛАРЫН АВТОМАТТЫ ТҮРДЕ ТАНУҒА КӨЗҚАРАС

AN APPROACH TO AUTOMATIC RECOGNITION OF THE SPEAKER'S EMOTIONS USING INTELLIGENT METHODS OF INFORMATION ANALYSIS

Аңдатпа. Жұмыста диктордың сөйлеуіне сәйкес эмоцияларды автоматты түрде жіктеу мәселесін шешу үшін қажетті әдіснамалық негіздер берілген. Машиналық оқыту әдістеріне негізделген эмоцияларды жіктеу алгоритмін синтездеудің жалпы принципі ұсынылған. Цифрлық сигналдарды өңдеу құралдарын қолдана отырып, сөйлеуден маңызды ақпараттық белгілерді ажыратуға мүмкіндік беретін сөйлеу процесінің моделі ұсынылған. Ықтималдылық тәсіліне негізделген классификатордың математикалық моделін құрудың жалпы процесі сипатталған.

Түйін сөздер: эмоцияны тану, сөйлеу, автоматты жіктеу, машиналық оқыту.

Аннотация. В работе представлены методологические основы, необходимые для решения задачи автоматической классификации эмоций по речи диктора. Предложен общий принцип синтеза алгоритма классификации эмоций на базе методов машинного обучения. Представлена модель процесса речеобразования, позволяющая при помощи инструментов цифровой обработки сигналов выделить из речи значащие информативные признаки. Описан общий процесс построения математической модели классификатора на базе вероятностного подхода.

Ключевые слова: распознавание эмоций, речь, автоматическая классификация, машинное обучение.

Abstract. The paper presents the methodological foundations necessary to solve the problem of automatic classification of emotions by the speaker's speech. The general principle of synthesis of the emotion classification algorithm based on machine learning methods is proposed. A model of the process of speech formation is presented, which allows using digital signal processing tools to isolate significant informative features from speech. The general process of constructing a mathematical model of a classifier based on a probabilistic approach is described.

Keywords: emotion recognition, speech, automatic classification, machine learning.

Введение. В настоящее время наиболее высокую эффективность в задачах автоматического распознавания человеческой речи демонстрируют интеллектуальные методы анализа голосовой информации [1, 2]. В этой связи для задач детектирования эмоций по речи диктора целесообразно также использовать хорошо зарекомендовавшие себя методы анализа на базе технологий машинного обучения (МО) и глубоких нейронных сетей.

Главные преимущества при использовании методов МО заключаются в возможности анализировать большие объемы информации для поиска скрытых закономерностей в данных. Это позволяет сопоставить с объектом исследования некоторые неявные его характеристики для выполнения классификации.

Применение интеллектуальных методов анализа позволяет эффективно автоматизировать задачи, связанные с обработкой больших потоков информации, разработать средства поддержки принятия решений для человеческого персонала в различных отраслях, понизив риски ошибок и снижения внимания. Использование инструментов теории искусственного интеллекта открывает возможности для решения трудных с точки зрения автоматизации проблем, для которых на данный момент был достигнут определенный предел по качеству их функционирования. К таким проблемам, несомненно, относится задача распознавания эмоций человека по его голосу.

Основная часть. Принимая во внимание все вышесказанное, структурную схему, поясняющую процесс синтеза интеллектуального метода автоматического распознавания эмоции по речи диктора, можно представить, как показано на рисунке 1.

В соответствии с рисунком 1 для решения задачи распознавания эмоций необходимо разработать математическую модель, которая будет способна с адекватной точностью выполнять многоклассовую классификацию по семи типам эмоциональных состояний (радость, страх, гнев, печаль, отвращение, удивление и нейтральное состояние). Модель получает на вход признаки классифицируемого объекта, извлекаемые в результате выполнения препроцессинга. На выходе модели результатом классификации является вектор значений вероятностей отнесения исследуемого объекта к одному из семи классов эмоций *Y*.

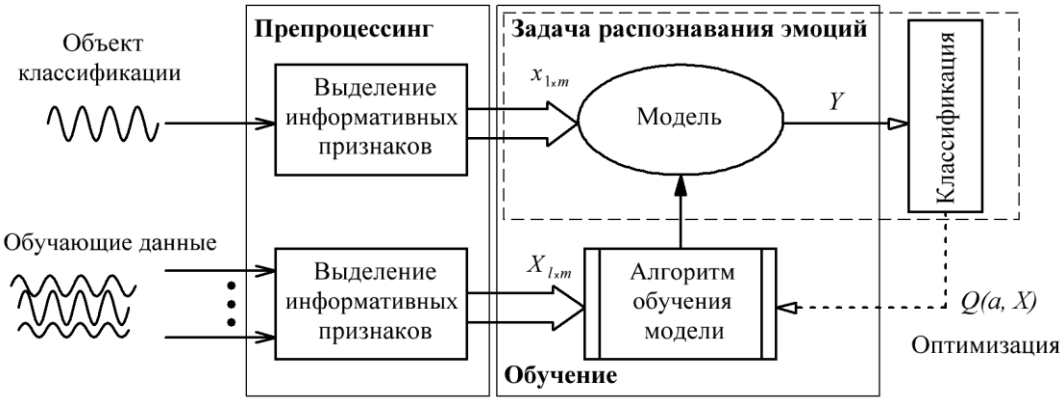

Рисунок 1 – Структура процесса автоматической классификации эмоций интеллектуальными методами анализа информации

Для синтеза требуемой модели предлагается использовать алгоритмические методы из теории МО. В процессе обучения, на вход модели подаются обучающие данные в виде обучающей выборки из аудиозаписей речевых сигналов с различными проявлениями эмоций, из которых извлекаются информативные признаки (препроцессинг). На обучающей выборки для каждого файла заранее известно какому типу эмоции он соответствует.

В результате процесс построения модели представляет собой интеллектуальный метод МО, известный как обучение с учителем или обучение на размеченных данных. Суть данного метода можно сформулировать следующим образом. Для имеющейся обучающей выборки $X = (x_i, y_i)_{i=1}^{l}$ необходимо отыскать такой алгоритм $a \in \mathrm{A}$, для которого будет достигаться минимум функционала ошибки $Q(a, X)$:

$$Q(a, X) \to \min_{a \in \mathrm{A}}. \quad (1)$$

Таким образом, в зависимости от объекта на входе модель формирует вероятность его отнесения к одному из классов эмоций. Полученный ответ модели Y сравнивается с известным правильным ответом. Результат сравнения выражается в виде некоторого принятого функционала ошибки $Q(a, X)$, как показано на рисунке 1. Процесс самообучения алгоритма заключается в стремлении снизить величину функционала ошибки последовательно изменяя значение параметров модели. Алгоритм обучения модели прекращается, когда достигнут глобальный минимум функционала ошибки или один из его локальных минимумов, удовлетворяющих предъявляемым условиям по качеству классификации.

Существование локальных минимумов сильно затрудняет процесс обучения модели. Также большое влияние на качество классификации оказывает правильный выбор минимизируемого функционала ошибки. Для алгоритма обучения модели необходимо подобрать правильные гиперпараметры, определяющие эффективность его работы. При правильной настройке гиперпараметров можно избежать явления переобучения, обеспечив для модели высокую обобщающую способность на новых данных [3].

Таким образом, для получения модели классификатора необходимо прежде всего иметь обучающий набор данных в виде записей человеческой речи с различной эмоциональной окраской. Наличие достаточного количества образцов речевого сигнала имеет основополагающее значение для построения модели классификатора. Однако, для продуктивного поиска решения задачи автоматической классификации эмоций необходимо сформировать четкое понимание о физической природе речевого сигнала и определить информативные признаки объектов.

Для этой цели предлагается использовать эквивалентную модель процесса речеобразования. При этом человеческая речь должна быть представлена набором характеристик, способных выступать в роли информативных признаков для модели классификатора. По этой причине в технической системе речь выгодно интерпретировать, как сигнал, физическим носителем которого является акустическое колебание. Т.е. речь является последовательностью звуков, разделенных паузами различной длительности.

В соответствии с этим, при использовании современных методов цифровой обработки сигналов (ЦОС) в задачах анализа речи необходимо иметь представление о процессах речеобразования в дискретном времени. Для этой цели речевой сигнал представляется в виде отклика нестационарной линейной системы на воздействие шума или квазипериодической последовательности импульсов [4], как показано на рисунке 2.

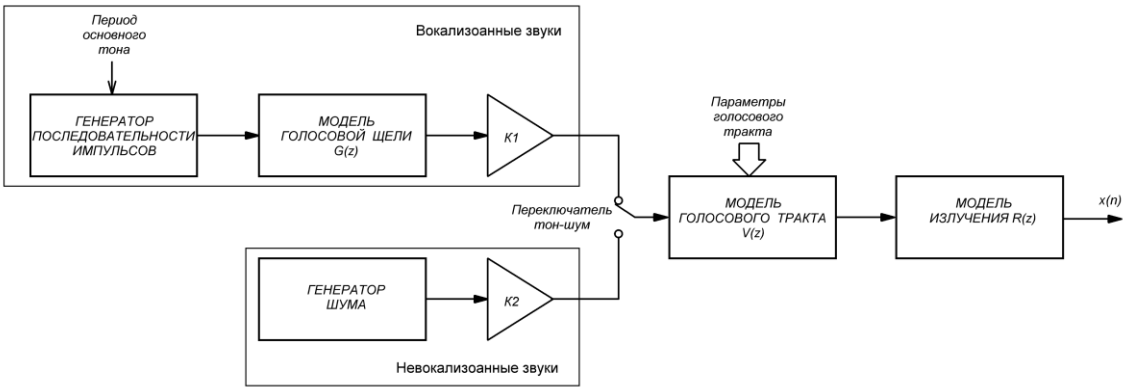

Рисунок 2 – Структурная схема дискретной модели речеобразования

В соответствии с рисунком 2 голосовой тракт может быть представлен эквивалентной моделью в дискретном времени с передаточной функцией вида:

$$V(z) = \frac{G}{1 + \sum_{k=1}^{n} a_k z^{-k}}, \qquad (2)$$

где G, a_k – это коэффициенты, которые изменяются во времени и определяются параметрами голосового тракта: зависимостью площади его поперечного сечения от расстояния вдоль продольной оси.

С позиции цифровой обработки сигналов (ЦОС) коэффициенты фильтра (2) отвечают за положение максимумов на амплитудно-частотной характеристике, которые называются формантными частотами. Формантные частоты (форманты) являются резонансными частотами голосового тракта. Они оказывают непосредственное влияние на формирование индивидуальных звуков речи.

В свою очередь, звуки, используемые в речеобразовании, называются фонемами. Для каждой фонемы огибающая спектра модели (2) приобретает определенную форму в зависимости от положения формантных частот. В процессе произнесения речи фонемы меняются и появляются формантные переходы.

В английском языке существует 42 фонемы, которые подразделяются на гласные, дифтонги, полугласные и согласные [4, с. 45]. Однако, большинство звуков речи можно условно разделить на образующиеся при участии голосовых связок – вокализованные, и образующиеся без использования связок – невокализованные.

Вибрация голосовых связок создает прерывистое движение воздушного потока из легких, которое может считаться периодическим. Соответствующий период повторения импульсов потока воздуха называют периодом основного тона. Как следует из рисунка 2, в модели речеобразования для формирования вокализованных звуков генератор последовательности импульсов формирует единичные импульсы с частотой основного тона F_0. Форма импульсов определяется передаточной функцией $G(z)$ линейной системы, импульсная характеристика которой соответствует форме колебания в голосовой щели. Блок $K1$ определяет интенсивность голосового сигнала при помощи соответствующего коэффициента усиления.

В свою очередь, процесс формирования невокализованных звуков заключается в использовании генератора шума (рисунок 2), мощность которого регулируется коэффициентом усиления $K2$.

В структурной схеме рисунка 2 также учитывается характер изменения звукового давления возле губ в виде модели излучения. Данный эффект можно представить в первом приближении в виде дифференциатора вида [4, с. 102]:

$$R(z) = G'(1 - z^{-1}) \quad (3)$$

Коэффициент усиления *G'* определяет интенсивность голосового возбуждения.

В результате общая передаточная функция дискретной модели речеобразования может быть представлена в виде:

$$H(z) = V(z)G(z)R(z). \quad (4)$$

Необходимо отметить, что представленная модель имеет множество ограничений, связанных с возможностью описания всех фонем языка. Однако, основываясь на структуре рисунка 2 можно сделать ряд важных выводов.

Во-первых: для использования технических средств ЦОС в задачах исследования речи необходимо применять кратковременный анализ сигналов, так как параметры модели будут постоянными лишь на отдельных промежутках времени.

Во-вторых: структура и параметры модели рисунка 2 позволяют предположить, что полезная информация о речевом сигнале будет преимущественно располагаться в частотной области. Т.е. изучение спектрального состава речевого сигнала позволит выявить значимые информативные признаки. Это объясняется тем, что модель речеобразования рисунка 2 представляет собой линейную систему, которая возбуждается периодически или случайно. Поэтому следует ожидать, что спектр выходного сигнала будет отражать свойства и голосового тракта и самого возбуждения.

В-третьих: форма речевого сигнала в основном будет иметь вид квазипериодических колебаний и шума. Причем спектральный состав квазипериодических колебаний будет определяться частотой основного тона F_0 и формантными частотами. Спектр шума распределен по всему диапазону частот, а функция распределения не несет определяющего характера.

Для примера на рисунке 3 представлен вид речевого сигнала при произношении мужчиной слова «two», записанного с частотой дискретизации f_S = 44100 Гц. Невокализованный звук |*t*| проявляется в виде шумового сигнала со времени 2,6 секунды. Примерно с 2,7 секунды на графике можно наблюдать квазипериодический процесс, который относится к вокализованному звуку |*u*:|.

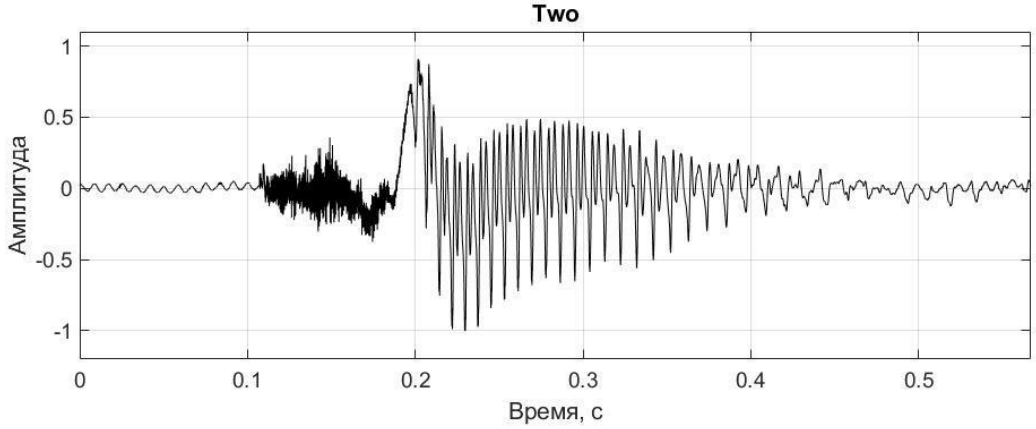

Рисунок 3 – Форма речевого сигнала при произношении слова «two».

На рисунке 4 представлен спектральный состав речевого сигнала при произношении слова «two». Спектры фонем накладываются друг на друга, что затрудняет определение частоты основного тона F_0 и первые форманты.

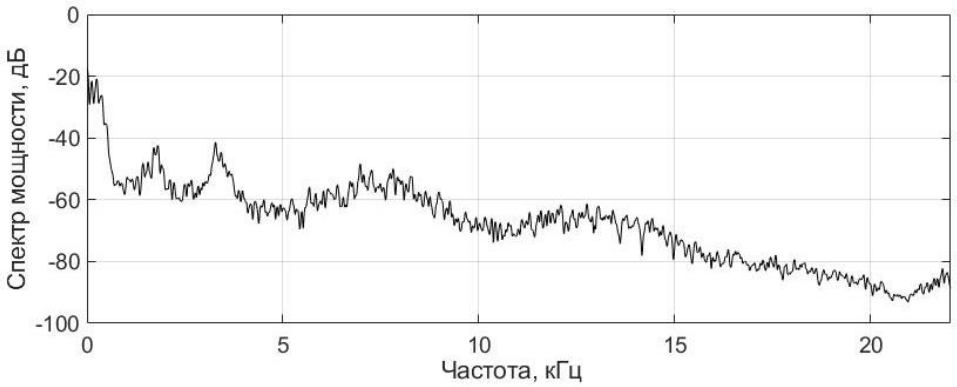

Рисунок 4 – Спектральный состав речевого сигнала при произношении слова «two».

Таким образом, обладая информацией о характеристиках и основных особенностях речевого сигнала, можно разработать структуру процесса предварительной цифровой обработки данных, называемого препроцессингом.

На основании предложенной модели речеобразования, известных особенностей психофизического восприятия звуков человеком, нечеткости и неоднозначности существующих формулировок понятия эмоции, а также неоднозначности и сложности выделения значимых информативных признаков, можно утверждать, что явления, порождающие данные об эмоциональном состоянии человека по речевому сигналу, представляют собой сложный многофакторный процесс. В связи с чем, любая математическая модель, синтезированная для задачи классификации эмоций по речевому сигналу, будет содержать некоторую долю неопределенности, которая не позволяет делать в результате классификации однозначные выводы.

Тогда в процессе обучения при создании математической модели классификатора, в соответствии с рисунком 1, необходимо применить вероятностный подход, т.е. выгодно рассматривать процесс МО с позиции его вероятностной интерпретации.

В основе вероятностных моделей МО лежит теорема Байеса, представленная в следующем виде:

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}, \qquad (5)$$

где X – значения признаков объектов классификации; Y – множество целевых переменных (классы объектов); $P(Y|X)$ – апостериорная вероятность; $P(X|Y)$ – функция правдоподобия; $P(Y)$ – априорная вероятность; $P(X)$ – вероятность наблюдения данных или признаков объектов, полученных на этапе препроцессинга.

Априорная вероятность $P(Y)$ представляет собой вероятность каждого эмоционального класса до наблюдения данных X. Вероятность наблюдения данных $P(X)$ не зависит от Y и может быть устранена путем выполнения нормировки, так как для нее справедливо выражение:

$$P(X) = \sum_{Y} P(X \mid Y)P(Y). \qquad (6)$$

При выполнении классификации математическая модель должна осуществлять решающее правило, согласно которому должен быть выбран класс с максимальной апостериорной вероятностью. Т.е. в соответствии с (5) имеем правило апостериорного максимума (MAP):

$$y_{MAP} = \arg\max_{Y} P(Y \mid X) = \arg\max_{Y} \frac{P(X \mid Y)P(Y)}{P(X)} = \\ = \arg\max_{Y} P(X \mid Y)P(Y). \qquad (7)$$

В соответствии с (7) построение математической модели классификатора будет заключаться в решении задачи оптимизации апостериорного распределения:

$$y_{MAP} = \arg\max_{Y} P(Y)\prod_{x\in X} p(x\,|\,Y) = \\ = \arg\max_{Y}\left(\log P(Y) + \sum_{x\in X}\log P(x\,|\,Y)\right), \tag{8}$$

где $P(X\,|\,Y) = \prod_{x\in X} P(x\,|\,Y)$ – функция маргинального правдоподобия, в соответствии с которой предполагается, что вероятности различных признаков *x* внутри класса независимы.

Если сделать предположение о равномерности априорного распределения $P(Y)$, то можно получить решающее правило максимального правдоподобия (ML):

$$y_{ML} = \arg\max_{Y}\sum_{x\in X}\log P(x\,|\,Y). \tag{9}$$

В выражениях (8), (9) переход к сумме логарифмов делается для упрощения процесса оптимизации, так как логарифм – это монотонная функция и argmax меняться не будет.

Таким образом, искомая модель классификатора эмоций по речевому сигналу будет получена путем решения задачи МО, которая в свою очередь состоит в том, чтобы найти и максимизировать распределение $P(Y|X)$. При этом необходимо выяснить, какие параметры лучше всего соответствуют имеющимся данным, а также существующим априорным представлениям. На практике данная задача реализуется в ходе выполнения оптимизации логарифма правдоподобия модели и регуляризаторов [5].

Выводы. Задача автоматической классификации эмоционального состояния человека по его речи отличается целым рядом трудностей, среди которых главными являются: существующая неоднозначность в формулировке понятия эмоции, сложная структура речевого сигнала и процессов его порождающих, особенности психофизического восприятия звуков человеком, а, следовательно, неопределенность в выборе характеристик речевого сигнала.

В свою очередь, задача автоматической классификации эмоций методами МО требует формирования репрезентативного набора обучающих данных. Выделение информативных признаков целесообразно производить на базе предложенной дискретной системы речеобразования. \при помощи вероятностного подхода к построению модели классификатора определен общий принцип ее обучения, который удовлетворяет как алгоритмам МО, так и методам глубокого обучения.

Список литературы

1. Ekman, P., "Universals and cultural differences in facial expressions of emotion", Nebr. Symp. Motiv. 1971, 207-283, 1972.

2. Uday Kamath, John Liu, James Whitaker Deep Learning for NLP and Speech Recognition. Springer Nature Switzerland AG 2019. P. 621.

3. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / пер. с англ. А. А. Слинкина. - М.: ДМК Пресс, 2015.-400 с.

4. Рабинер Л. Р., Шафер Р. В. Цифровая обработка речевых сигналов: Пер. с англ./Под ред. М. В. Назарова и Ю. Н. Прохорова. – М.: Радио и связь, 1981. — 496 с,

5. Николенко С., Кадурин А., Архангельская Е. Глубокое обучение. — СПб.: Питер, 2018. – 480 с.

References

1. Ekman, P., "Universals and cultural differences in facial expressions of emotion", Nebr. Symp. Motiv. 1971, 207-283, 1972.
2. Uday Kamath, John Liu, James Whitaker Deep Learning for NLP and Speech Recognition. Springer Nature Switzerland AG 2019. P. 621.
3. Flah P. Mashınnoe obýchenıe. Naýka ı ıskýsstvo postroenıa algorıtmov, kotorye ızvlekaıýt znanıa ız dannyh / per. s angl. A. A. Slınkına. - M.: DMK Press, 2015.-400 s.
4. Rabıner L. R., Shafer R. V. Tsıfrovaıa obrabotka rechevyh sıgnalov: Per. s angl./Pod red. M. V. Nazarova ı Iý. N. Prohorova. – M.: Radıo ı svıaz, 1981. — 496 s,
5. Nıkolenko S., Kadýrın A., Arhangelskaıa E. Glýbokoe obýchenıe. — SPb.: Pıter, 2018. – 480 s.

DOI 10.53364/24138614_2022_26_3_56
ӘОЖ 378.147

[1]**Елубай А.М.,** [2]**Тулекова Г.Х.,**[3]**Суранчиева Н.Р.**
[1,2,3]Азаматтық авиация академиясы, Алматы қ., ҚР.

[1]*E-mail: gulnaz.tulekova@mail.ru*
[2]*E-mail: smailova_asem@mail.ru*
[3]*E-mail: nazgul_87@bk.ru*

АВИАЦИЯ САЛАСЫ БОЙЫНША КӘСІБИ МӘТІНДЕРДІ ТЫҢДАЛЫМ АРҚЫЛЫ МЕҢГЕРТУДІҢ ОҢТАЙЛЫ ӘДІСТЕРІ

ОПТИМАЛЬНЫЕ МЕТОДЫ АУДИРОВАНИЯ ПРОФЕССИОНАЛЬНЫХ ТЕКСТОВ ПО АВИАЦИОННОЙ ОТРАСЛИ

OPTIMAL METHODS OF LISTENING TO PROFESSIONAL TEXTS ON THE AVIATION INDUSTRY

Аңдатпа. Мақалада сөйлесім әрекетінің маңызды бір түрі тыңдалым әрекетін меңгертудің маңызы туралы айтылады. Орыс тілді аудиторияда авиация саласы бойынша білім алушыларға тыңдалымды меңгерту арқылы тілдік қарым-қатынасқа түсу мүмкіндігін дамытып, тіл үйренуге деген қызығушылығын, ынтасын, белсенділігін арттыруға болады. Білім алушының тыңдалған ақпаратты қабылдауы, есте сақтауы мәтін мазмұнының ақпараттылығы мен композициялық құрылымына, сондай-ақ мәтін көлемі мен айтылу уақытына байланысты. Аудиомәтінде негізгі ойды білдіретін фактілер мен дәлелдемелерден тұратын ақпараттың болуы, белгілі бір шешімі бар мәселенің болуы, айтылған ойдың логикалық жүйелілігі және соңында мәтіннің мазмұнын ашатын қорытындының берілуі – мәтіннің әдістемелік талапқа сай екенін көрсетеді. Жалпы мәтіннің мазмұны тіл үйренушіге түсінікті болып, тіл үйренуші оны сөйлеу үдерісінде қолдана білу керек. Мәтіндегі таныс емес сөздер 2 пайыздан аспаған жағдайда, тіл үйренуші оларға назар аудармай, мәтінде айтылуға тиісті негізгі ойды меңгеруге тырысады. Тыңдалымды меңгертумен қатар оны бақылау, бағалау қатар жүріп отырғанда ғана жұмысымыздың нәтижесін көре аламыз.

Мақалада авиация саласы бойынша білім алушылардың кәсіби мәтіндерді тыңдалым арқылы меңгертудің оңтайлы әдістері мен тәсілдері қарастырылған. Маманның кәсіби