МРНТИ 16.31.21
УДК 004.934

https://doi.org/10.51889/2022-1.1728-7901.16

Н.О. Мекебаев[1,2], Ш.М. Түйебаев[2], Қ.Ж. Сабраев[3], А.Қ. Еркебай[1]*

[1]Қазақ ұлттық қыздар педагогикалық университеті, Алматы қ., Қазақстан
[2]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы қ., Қазақстан
[3]Абай атындағы Қазақ ұлттық педагогикалық университеті, Алматы қ., Қазақстан
**e-mail: Nurbapa@mail.ru*

БАЛАЛАРДЫҢ СӨЙЛЕУІН ТАНУ ҮШІН ҚАЙТАЛАНАТЫН НЕЙРОНДЫҚ ЖЕЛІЛЕР НЕГІЗІНДЕ АКУСТИКАЛЫҚ ЖӘНЕ ЛИНГВИСТИКАЛЫҚ МОДЕЛЬДЕУДІ ЗЕРТТЕУ

Аңдатпа

Сөзді автоматты түрде тану (ASR) жүйелері акустикалық модельдеу үшін GMM-HMM және тілді модельдеу үшін n-gram қолданады. Соңғы онжылдықта терең бағытталған нейрондық желі (DFNN) акустикалық модельдеуде GMM-ді, яғни математикалық статистика мен эконометрикада үлестірудің белгісіз параметрлерін және эконометрикалық модельдерді бағалау әдісін толық дерлік ауыстырды. Қазіргі автоматты түрде тану жүйелері негізінен DFNN-HMM акустикалық моделіне және n-gram тіл үлгісіне (LM) негізделген. Ұзақ қысқамерзімді контекстті модельдеу мүмкіндігінің арқасында қайталанатын нейрондық желіге (RNN) негізделген тіл үлгілері n-gram тіл үлгілеріне қарағанда төмен түсініксіздікті береді деп хабарланған. Осы жұмыстарға негізделген бұл мақалада біз балалардың сөйлеуін автоматты түрде тану үшін қайталанатын нейрондық желіге негізделген тіл үлгісімен біріктірілген ұзақ қысқамерзімді жадыға негізделген акустикалық модельдеуді зерттейміз.

Эксперименттік нәтижелер осындай біріктірілген қайталанатын нейрондық желіге негізделген модельдеу балалардың сөйлеуді автоматты түрде тануының сәйкес және сәйкес келмейтін тапсырмаларында тиімді екенін көрсетеді.

Түйін сөздер: сөйлеуді автоматты тану (ASR), қайталанатын нейрондық желілер (RNN), тілдік модельдеу (LM), акустикалық модельдеу (AM), LSTM, DFNN, GMM, HMM.

Аннотация

Н.О. Мекебаев[1,2], Ш.М. Түйебаев[2], Қ.Ж. Сабраев[3], А.Қ. Еркебай[1]
[1]Казахский национальный женский педагогический университет, г. Алматы, Казахстан
[2]Казахский национальный университет им. аль-Фараби, г. Алматы, Казахстан
[3]Казахский национальный педагогический университет имени Абая, г. Алматы, Казахстан

ИССЛЕДОВАНИЕ АКУСТИЧЕСКОГО И ЛИНГВИСТИЧЕСКОГО МОДЕЛИРОВАНИЯ НА ОСНОВЕ ПОВТОРЯЮЩИХСЯ НЕЙРОННЫХ СЕТЕЙ ДЛЯ РАСПОЗНАВАНИЯ РЕЧИ ДЕТЕЙ

Системы автоматического распознавания речи (ASR) используют GMM-HMM для акустического моделирования и n-gram для моделирования языка. За последнее десятилетие глубоко ориентированная нейронная сеть (DFNN) практически полностью заменила GMM в акустическом моделировании, то есть неизвестные параметры распределения и метод оценки эконометрических моделей в математической статистике и эконометрике. Современные системы автоматического распознавания в основном основаны на акустической модели DFNN-HMM и языковой модели n-gram (LM). Сообщается, что образцы языка, основанные на повторяющейся нейронной сети (RNN), дают более низкую двусмысленность, чем образцы языка n-gram, благодаря возможности моделирования длительно сжатого контекста. В этой статье, основанной на этих работах, мы исследуем акустическое моделирование на основе долговременной кратковременной памяти, объединенное с образцом языка на основе повторяющейся нейронной сети для автоматического распознавания речи детей.

Экспериментальные результаты показывают, что моделирование на основе такой интегрированной повторяющейся нейронной сети эффективно в соответствующих и несовместимых задачах автоматического распознавания речи детей.

Ключевые слова: автоматическое распознавание речи (ASR), повторяющиеся нейронные сети (RNN), языковое моделирование (LM), акустическое моделирование (AM), LSTM, DFNN, GMM, HMM.

Abstract

RESEARCH OF ACOUSTIC AND LINGUISTIC MODELING BASED ON REPETITIVE NEURAL NETWORKS FOR SPEECH RECOGNITION OF CHILDREN

Mekebayev N.[1,2], Tuyebaev Ch.[2], Sabrayev K.[3], Yerkebay A.[1]

[1]Kazakh national women's teacher training university, Almaty, Kazakhstan

[2]al-Farabi Kazakh National university, Almaty, Kazakhstan

[3]Abai Kazakh National Pedagogical university, Almaty, Kazakhstan

Automatic Speech Recognition (ASR) systems use GMM-HMM for acoustic modeling and n-gram for LM. Over the past decade, a deeply oriented NN DFNN has almost completely replaced GMM in acoustic modeling, that is, unknown distribution parameters and a method for evaluating econometric models in mathematical statistics and econometrics. Modern automatic recognition systems are mainly based on the DFNN-HMM acoustic model and the n-gram LM. It is reported that language samples based on a repeating RNN give lower ambiguity than n-gram language samples, due to the possibility of modeling a long-compressed context. In this paper, based on these papers, we investigate acoustic modeling based on long-term short-term memory combined with a language sample based on a repeating NN for automatic speech recognition of children.

Experimental results show that modeling based on such an integrated repetitive neural network is effective in appropriate and incompatible tasks of automatic speech recognition of children.

Keywords: Automatic speech recognition (ASR), repetitive neural networks (RNN), language modeling (LM), acoustic modeling (AM), LSTM, DFNN, GMM, HMM.

I. Кіріспе

Сөйлеуді автоматты түрде тану машиналардың көмегімен сөзді мәтінге түрлендіру міндетіне жатады [1]. Сөйлеуді автоматты түрде тану жүйесінің негізгі мақсаты - сөйлеу сөздерінің акустикалық көрінісін түсіру және үлгіні сәйкестендіру әдістері арқылы айтылған сөздерді анықтау. Дәстүрлі түрде ASR жүйелері акустикалық модельдеу үшін Гаусс қоспасының моделіне негізделген жасырын Марковлгісін (GMM-HMM) [3] және лингвистикалық модельдеу үшін n-грамманы пайдалана отырып әзірленеді. Акустикалық үлгінің параметрлері әдетте сөйлеу сигналдарының алдыңғы жағындағы параметрлеуге негізделген ұсақ жиілікті цестральды коэффициенті (ағыл. MFCC - Mel-frequency cepstral coefficients) бойынша оқытылады. ASR зерттеу жұмыстарының көпшілігі ең алдымен ересектердің сөйлеуін тануға арналған. Балалардың ASR жүйесінің дамуы ересектердің ASR саласында пайда болған зерттеу тенденцияларын мұқият бақылайды. Ересектерге арналған заманауи ASR жүйелері HMM (DFNN-HMM) негізіндегі терең бағытталған нейрондық желіге негізделген [4].

DFNN - акустикалық мүмкіндіктерді ескере отырып, сенонның артқы ықтималдығын (байланған трифон күйін) шығару үшін *уақыт бойынша кері таралу* (ағыл. BPTT - back propagation through time) арқылы үйретілген көп қабатты сызықты емес желі. Жаттығуда пайдаланылатын бекітілген өлшемді жылжымалы терезе үшін GMM және DFNN екеуі де сигналдардағы ұзақ қысқамерзімді тәуелділіктерді модельдей алмайды. Керісінше, кері байланыс (қайталанатын) қосылымдарды қамтитын ұзақ қысқамерзімді жады оларды осындай күрделі уақытта өзгеретін сигналдарды модельдеу үшін қолайлы етеді [5].

ASR-де тілдік модель ауызша айтылымдардағы сөздер тізбегінің іздеу кеңістігін азайтуға және P(W**)** сөз тізбегінің бірлескен ықтималдығын қамтамасыз етуге көмектеседі. Ең жиі қолданылатын *n*-gram тіл үлгісі *(n−1)* сөздердің тізбегін пайдалана отырып, келесі сөзді реттілікпен болжайды. Бұл әдістемеде $\omega_1,\ldots, \omega_N$ сөйлемін сақтау ықтималдығы келісідей өрнекпен жуықталады:

$$P(W) = \prod_{i=1}^{N} P(\omega_i|\omega_1,\ldots,\omega_{i-1}) \approx \prod_{i=1}^{N} P\left(\omega_i\middle|\omega_{1-(n-1)},\ldots,\omega_{i-1}\right) \quad (1)$$

N-gram тіл үлгісінің маңызды кемшіліктері:

- Мәтінмәннің сәйкессіздігі: Тестілеу кезінде көрінбейтін мәтінмәнді шешу үшін резервтік модель (ағыл. Back-off model) [6] қолданылады. Артқа түсіру үлгісі *(n−1)*-грамдық қатардың шартты ықтималдығын бағалайды, мұнда ең сол жақ сөз n-граммдық қатардан жойылады.
- *n*-gram тілінің моделі ұзақ мерзімді тәуелділіктерді модельдей алмайды, сондықтан n = 5-тен жоғары болса, тиімді деп табылмайды.
- W-gram тіл үлгісі тек семантиканы емес, тек қана синтаксиканы үлгілей алады.

Нейрондық желіге негізделген тіл үлгілері n-gram тіл үлгілеріндегідей сөз жиіліктерінен гөрі кіріс ретінде сөздіктегі сөздердің бірігіп келуін бөлуді қабылдайды. Осының нәтижесінде нейрондық желіге негізделген тіл үлгісі кері схемаға қарағанда контекстік сәйкессіздікті тиімдірек өңдей алады.

Кері байланыс қосылымдарының болуымен қайталанатын нейрондық желі (RNN) мәтінмәнді де, ұзақ мерзімді уақытша ақпаратты да тиімді модельдей алады. Ағымдағы әдебиеттерде қайталанатын нейрондық желіге негізделген тіл моделі (RNNLM) кеңінен зерттелген және әдеттегі *n*-gram тіл үлгісінен айтарлықтай асып түсетіні хабарланған [7].

Ұзақ қысқамерзімді жадыға негізделген акустикалық модельдеу үлкен сөздік қорының контекстінде ересектердің сөйлеуін автоматты түрде сөйлеуді тану тапсырмасының контекстінде, бірақ n-gram тіл үлгісінің контекстінде зерттелген [8]. Бұл жұмыстың қосқан үлесі зор. Біріншіден, біз ұзақ қысқамерзімді жадыға негізделген акустикалық модельді қайталанатын нейрондық желі-тілдік модельді n-grammen салыстыра отырып, бағалағымыз келеді. Екіншіден, біз сәйкес және сәйкес келмейтін сынақ жағдайында балалардың автоматты түрде сөйлеуді тану тапсырмасын модельдеудегі осы соңғы жетістіктерді зерттеуді көздеп отырмыз. Бір қызығы, біздің эксперименттік зерттеуіміз қайталанатын нейрондық желіге негізделген акустикалық және лингвистикалық модельдеу осы жұмыста қарастырылған өте төмен ресурсты автоматты түрде сөйлеуді тану тапсырмасында тиімді болуы мүмкін екенін анықтады.

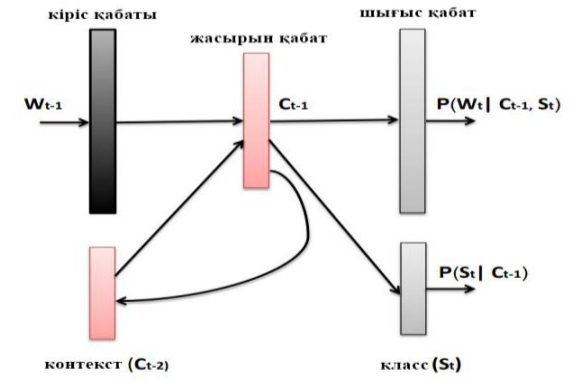
Бұл жұмыстың қалған бөлігі келесідей ұйымдастырылған: II бөлімде біз акустикалық және лингвистикалық модельдеу үшін қайталанатын нейрондық желі архитектурасының нұсқаларын зерттейміз. Сөйлеу корпусының егжей-тегжейлері және осы зерттеуге қатысатын жүйе параметрлері III бөлімде сипатталған. Зерттелген ұзақ қысқамерзімді жадқа негізделген акустикалық модельді және қайталанатын нейрондық желілік тіл үлгісін бағалау IV бөлімде ұсынылған. Бұл құжат V бөлімде қорытындыланады.

II. Сөйлеуді автоматты танудағы қайталанатын нейрондық желілер

Терең нейрондық желіге негізделген акустикалық модельдеу автоматты түрде сөйлеуді танудағы нормаға айналғанымен, қайталанатын нейрондық желілер әлі де кеңінен зерттелуі керек. Қайталанатын нейрондық желінің ең ерте қолданылуы тілдік модельдеуде жүзеге асырылады және оны үздіксіз сөйлеуді тану үшін акустикалық модельдеуде пайдалану өте жақында хабарланды. Бұл бөлімде біз алдымен қайталанатын нейрондық желіні қолдану арқылы тілдік модельдеу қалай орындалатынын қарастырамыз. Одан кейін акустикалық модельдеу үшін қабылданған қайталанатын нейрондық желінің нұсқасының сипаттамасы беріледі.

А. Қайталанатын нейрондық желіге негізделген тілді модельдеу

Қайталанатын нейрондық желі ұзақ мерзімді тәуелділіктерді модельдеу мүмкіндігіне ие және ол қолданыстағы n-gram тіл үлгісіне қарағанда тиімдірек тілді модельдеу үшін пайдаланылды. Қайталанатын нейрондық желіге негізделген тіл үлгісін модельдеуге арналған бір реттік деңгейлі желінің жалпы құрылымы 1-суретте көрсетілген.



Сурет 1. RNN-LM класс негізіндегі желі архитектурасы

Ағымдағы сөздің ықтималдығын модельдеу үшін ω_t, алдыңғы сөздердің толық тарихын $(\omega_{t-1},\ldots,\omega_1)$ қайталанатын байланыстар арқылы алынған. Алдыңғы контекстік ақпарат c_{t-2} және ω_{t-1} сөзі ағымдағы мәтінмәндік ақпаратты c_{t-1} модельдеу үшін жасырын деңгейге кіріс ретінде беріледі. Шығару деңгейі келесі ω_t сөзінің реттіліктегі ықтималдығын генерациялау үшін осы контекстік ақпаратты c_{t-1} пайдаланады.

Толық шығыс қабаты бар қайталанатын нейрондық желі-тілдік модельді оқыту күрделі есептеуіш болып табылады. Сонымен, бұл мәселені шешу үшін сөздер st кластарына жіктеледі, содан кейін осы класс ақпаратын пайдаланып RNN-LM оқытылады [9].

Әрі қарай, біз белгілі бір класқа жататын сөздердің ықтималдығы тек сол нақты сыныптың ықтималдығына ғана емес, сонымен бірге алдыңғы контекстке де байланысты деп болжауға болады. Бұл жұмыста сыныпқа негізделген қайталанатын нейрондық желі-тіл үлгісі қарапайым факторизация әдісін қолданады, мұнда сөздер жиілік санына негізделген сыныптар арасында бөлінеді. Сондықтан бастапқы сыныптар жиілігі жоғары жалғыз сөздермен, ал кейінгі сыныптар жиілігі аз көп сөздермен тағайындалады. Сөз тізбегінің бірлескен ықтималдығын есептеу үшін біз алдымен жеке класстардың ықтималдық үлестірімін есептейміз, содан кейін осы нақты класқа тағайындалған сөздердің таралуын есептейміз. Берілген мәтін wt сөзінің пайда болу ықтималдығы ct−1 (2) арқылы берілген.

$$\mathrm{P}(c_{t-1}) = P(c_{t-1}) * P(c_{t-1}, S_t) \quad (2)$$

RNN-LM мәтінмәндік сәйкессіздік мәселесін сөздіктегі сөздердің әрқайсысы үшін үлестірілген ұсынуды үйрену арқылы шешеді. Жаттығу кезінде сөздер тізбегінің бірлескен ықтималдылық үлестірім функциясы сөздердің әрқайсысы үшін осы туынды үлестірілген өкілдіктер тұрғысынан есептеледі. Бұл әдістеме арқылы жаттығу кезінде байқалмайтын сөз тізбегі, егер ол байқалатын тізбектерге мағынасы жағынан ұқсас болса, жақсы жалпылау алынады.

Мысалы, сәйкесінше оқыту (көрген) және тестілеу (көрінбейтін) кезеңдеріндегі келесі сөз тізбегін қарастырайық.

• Тренинг сөйлем: Мұнда мысал сөйлем берілген

• Сынақ сөйлем: Мысал сөйлем мұнда берілген

Назар аударыңыз, екі сөйлем бірдей сөздерді қамтиды, бірақ реті бойынша ерекшеленеді. LM семантикалық жағынан ұқсас оқыту сөйлемін көргенде берілген көрінбейтін сынақ сөйлемді тани білуі керек. Кәдімгі *n*-gram тілдік модельдеу мұны жасай алмады, өйткені олар тек сөз тізбегін үлгілей алады, бірақ мағынасын емес. Жалпы, ұқсас контексте кездесетін сөздер көбінесе бір тапқа жатады. Осылайша, сөздің сөздік құрамындағы басқа сөздердің жанында қаншалықты жиі кездесетінін есептей отырып, біз олар кездесетін сөйлемдердің семантикасын модельдей аламыз.

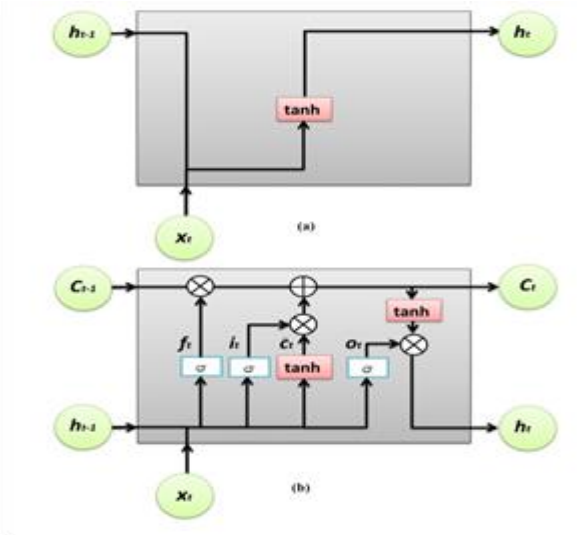
B. LSTM негізіндегі акустикалық модельдеу

DFNN архитектурасында желінің сөйлеу сигналдарының уақытша өзгергіштігін модельдеу мүмкіндігі акустикалық модельдеуде қолданылатын біріктірілген функция векторының ұзындығына байланысты. Сондай-ақ, DFNNs ұзақ мерзімді тәуелділіктерді түсіру үшін қолайлы емес.

RNN қолдану арқылы біз ұзақ мерзімді тәуелділіктерді де, сигналдың уақытша өзгергіштігін де модельдей аламыз. Бірақ бұл желілер уақыт бойынша кері таралу кезінде белгілі жоғалып кететін градиент мәселесінен зардап шегеді. Бұл белгілі бір уақыт аралығында қате функциясының кері таралатын градиенті экспоненциалды түрде жарылып немесе ыдырайтынын білдіреді. Бұл келесі уақыт қадамдарында салмақтардың дұрыс бейімделмеуіне әкеледі. Бұл мәселені шешу үшін әдебиетте LSTM деп аталатын модификацияланған RNN архитектурасы ұсынылған [10].

LSTM архитектурасында қайталанатын қабат ақпарат ағынын басқару үшін үш арнайы қақпамен бірге нейрондық желінің уақытша күйін сақтай алатын жад ұяшықтарын қамтиды. RNN және LSTM архитектурасының ашылмаған нұсқасының блок-схемалары 2-суретте келтірілген.

t мезетіндегі x_t кіріс сигналы үшін c_t жад ұяшығына ақпарат ағыны желі қанша ақпаратты есте сақтау және ұмыту қажет екенін бақылайтын *енгізу* және *ұмыту* қақпаларының көмегімен шешіледі. Мысалы i_t және f_t сәйкесінше желінің есте сақтайтын және ұмытатын ақпаратты білдірсін. Сондай-ақ, қайталанатын нейрондық желінің шығысы $\tilde{c}_t$ делік. Осы үш ақпаратты біріктіру арқылы жады ұяшығына c_t үлесі анықталады.



Сурет 2. (a) RNN және (b) LSTM желілік архитектурасын көрсететін блок-схемалар

Сонымен қатар, желі келесі кезеңге жіберетін c_t жады ұяшығынан o_t ақпараты шығыс қақпасы арқылы басқарылады. Бұл операциялар математикалық түрде келесідей көрсетіледі:

$$f_t = \tau(w_f \cdot [h_{t-1}, x_t] + b_f) \quad (3)$$

$$i_t = \tau(w_i \cdot [h_{t-1}, x_t] + b_i) \quad (4)$$

$$\tilde{c}_t = tanh(w_i[h_{t-1}, x_t] + b_c) \quad (5)$$

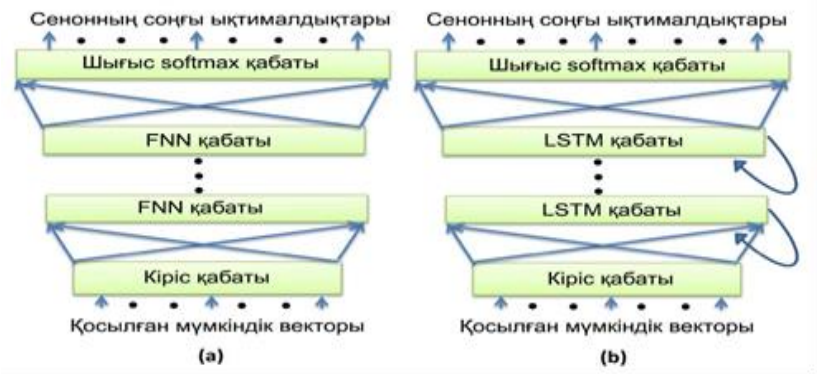
$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \quad (6)$$

$$o_t = \tau(w_o \cdot [h_{t-1}, x_t] + b_o) \quad (7)$$

$$h_t = o_t * tanh(c_i) \quad (8)$$

мұндағы $\omega_{\{f,i,c,o\}}$ және $b_{\{f,i,c,o\}}$ салмақты және сәйкес желілермен байланысты ауытқуды білдіреді.

Алға жіберетін DNN сияқты, ұзақ қысқамерзімді жады қабаттары (LSTM) да тереңірек архитектураны құру үшін жинақталады. Жалғыз ұзақ қысқамерзімді жады (LSTM) қабатының өзі ұзақ мерзімді тәуелділіктерді түсіре алатынына қарамастан, терең ұзақ қысқамерзімді жады (ағыл. DLSTM – deep longshort term memory) пайдалану акустикалық модельдеуде тиімді болып табылады [14]. Бұл бір ұзақ мерзімді жады желісінің үлгі өлшемін үлкейтудің орнына DLSTM желісіндегі бірнеше деңгейлер бойынша параметрлерді бөлуге байланысты. 3-суретте DFNN және DLSTM архитектуралары көрсетілген.



Сурет 3. (a) DFNN және (b) DLSTM көмегімен акустикалық модельдеуде қолданылатын желілердің топологиясы

III. Экспериментты орнату

Эксперименттік бағалауда қолданылатын сөйлеуді автоматты тану жүйелері негізінен Kaldi Takeit [11] көмегімен әзірленген. RNN негізіндегі тілдік модельдеуді әзірлеу үшін біз RNNLM құралдар жинағын қолдандық [12]. Төменде біз сөйлеу корпусының егжей-тегжейлерін, қолданылатын акустикалық және лингвистикалық үлгілердің құрылымын және жүйе параметрлерін баптауды сипаттаймыз.

A. Мәліметтер қоры

Балалардың сөйлеуді автоматты тану жүйесін әзірлеу кезінде акустикалық модельдеуге арналған деректер PFSTAR балалар сөйлеу корпусынан [13] алынған, ал оқу деректерінің транскрипттері тілдік модельдеуді оқыту үшін пайдаланылады. PFSTAR корпусында 4-13 жас тобындағы ұл/қыз балалардан жиналған оқу-сөйлеу деректері бар. Оның оқу жинағы 46 074 сөзден тұратын 8,3 сағаттық деректерден, 122 спикердің 959 сөзінен, ал сынақ жинағы 5 067 сөзден тұратын 1,1 сағаттық деректерден, 60 спикерден және 129 сөйлеуден тұрады.

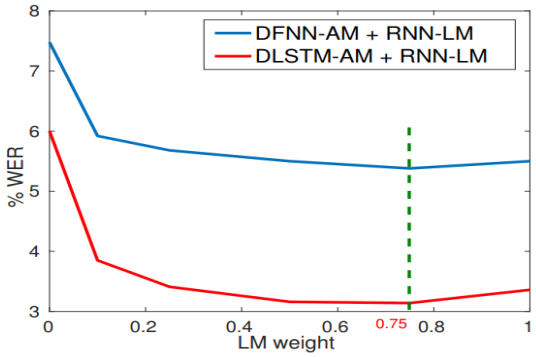
B. Акустикалық модельдеу үшін параметрді баптау

GMM-HMM контекстке тәуелді акустикалық модельдер сенондардың санын 2500 және сенонға 16 гаусс қоспасы ретінде сақтай отырып оқытылады. DFNN-HMM акустикалық үлгілері 5 жасырын қабатпен және жасырын қабаттардың әрқайсысында сызықты еместік функциясы ретінде *tanh* бар 1024 түйінмен оқытылады. Модель 20 және 128 шағын топтама өлшемімен дайындалған. DLSTM негізіндегі акустикалық үлгілер әрқайсысы 256 түйіннен тұратын тек 2 жасырын қабатпен оқытылады және шағын партия өлшемі сәйкесінше 5 және 128-ге тең. DLSTM оқытуындағы бұл шектеулі таңдау біздің тарапымыздан қолжетімді GPU ресурстарымен негізделген. 91 өлшемді мүмкіндік векторы ± 3 кадрдан асатын 13 өлшемді MFCC мүмкіндіктерін біріктіру арқылы алынады [14].

Алынған мүмкіндік векторларының өлшемі сызықтық дискриминанттық талдауды (LDA) пайдалану арқылы 40-қа дейін азаяды. Бұл 40 өлшемді мүмкіндік векторлары жоғарыда аталған барлық акустикалық үлгілерді үйрету үшін пайдаланылады.

C. Тілді модельдеуге арналған параметрлерді орнату.

Осы жұмыста зерттелген RNN негізіндегі тілдік модельге қарама-қайшылықты қамтамасыз ету үшін біз балаларға арналған автоматты тану жүйелерін үшін 2-граммдық және 4-граммдық тілдік модель әзірледік. Тілдік модельдің сөздік көлемі сәйкесінше балалар үшін 1,5 КБ ретінде таңдалады. RNN-LM 2 жасырын қабатпен және әрбір жасырын қабатта сызықтық емес функция ретінде 200 сигма тәрізді түйінмен оқытылады. Бұған қоса, сыныптар саны 200-ге орнатылады және BPTT айнымалысы 4-ке орнатылады. Декодтау кезінде RNN-LM биграммдық (2) тілдік модельдеу көмегімен жасалған торларға қолданылады. Тілдік модельдеу салмағы 0,25 қадаммен 0,0-ден 1,0-ге дейін реттелді және 0,75 мәні оңтайлы болып саналады. Бұл баптау тәжірибесінің нәтижелері графигі және 4-суретте көрсетілген.



Сурет 4. Екі түрлі желілік акустикалық үлгілерде RNNLM үшін тордың қайта бағалау салмағы. (0,75 салмақ мәні екі жағдай үшін де оңтайлы болып саналады)

IV. Нәтижелер

DLSTM негізіндегі акустикалық модельдеуді, сондай-ақ RNN негізіндегі лингвистикалық модельдеуді бағалау балаларға арналған сөйлеуді тану мәселелері үшін жүргізілді. 1-ші кестеде сөз қатесінің жылдамдығы (WER) тұрғысынан акустикалық модель және тілдік модель әртүрлі

комбинациялары үшін тану өнімділігі көрсетілген. Балаларға арналған тапсырмалар үшін бастапқы баллдар (GMM-HMM және 2-граммдық LM) сәйкесінше 9,87% және 17,97% құрайды. Өнімділіктегі бұл үлкен айырмашылық екі мәселеде де тілдік модельдің тиісті оқу жазбаларындағы шектеулі деректерді пайдалана отырып оқытылатындығына байланысты. Бұған кестедегі «Ts-LM» белгісі дәлел. Біздің тарапымыздан көбірек дайындалған тілдік болмауына байланысты балалар жағдайына ұқсас зерттеу жүргізілмеді. RNN-LM көмегімен ұзағырақ контекстік модельдеумен әділ салыстыру үшін 4 грамм LM өнімділігі де есептелді. 1-ші кестеден RNNLM акустикалық үлгілердің әртүрлі түрлерімен үйлескенде екі мәселеде де 2G және 4G LM құрылғыларымен салыстырғанда тану өнімділігі тұрақты түрде жақсырақ болатынын атап өтуге болады. RNN-LM DFNN және DLSTM акустикалық үлгілерімен үйлескенде балалардың сөйлеуді автоматты тануы үшін би-граммдық LM-ге қарағанда сәйкесінше 28% және 47% салыстырмалы өсім береді. Сөйлеуді автоматты тануда балалар үшін бұл салыстырмалы өсу сәйкесінше 7% және 8% құрайды. MIT-LM құру кезінде пайдаланылған мәтіндік деректерге қолымыз жетпегендіктен, RNN / 4G LM үйрету мүмкін емес. Белгіленген үрдістер бұл жағдайда да жалғасады деп ойлаймыз.

Кесте 1. Ересектердің және балалардың автоматты сөйлеуін тану тапсырмаларындағы түрлі акустикалық модельдер мен тілдік модельдердің WER көрсеткіштері. 'TS-LM' бағаны тек акустикалық транскрипттер бойынша оқытылған тілдік модельдердің нәтижесі

Акустикалық модельдеу (AM)	*Тілдік модельдеу (LM)*	*Балалар*
		Ts-LM
GMM-HMM	*bi-gram*	*9.87*
	4-gram	*7.96*
	RNN	*7.29*
DFNN-HMM	*bi-gram*	*7.48*
	4-gram	*5.92*
	RNN	*5.38*
DLSTM-HMM	*bi-gram*	*6.00*
	4-gram	*3.82*
	RNN	*3.14*

Кесте 2. Мәтінмәнге нақты сәйкес келмейтін сөзді автоматты тану жүйесінде оқытылатын акустикалық және лингвистикалық модельдеу әдістерін бағалау көрсеткіштері.

Акустикалық модельдеу (AM)	*Тілдік модельдеу (LM)*	*WER көрсеткіші, %*	
		Default	*+ VTLN*
GMM-HMM	*bi-gram*	*94.45*	*85.88*
	4-gram	*93.29*	*84.66*
	RNN	*92.52*	*83.16*
DFNN-HMM	*bi-gram*	*80.39*	*72.48*
	4-gram	*78.40*	*71.25*
	RNN	*77.61*	*69.75*
DLSTM-HMM	*bi-gram*	*74.82*	*66.85*
	4-gram	*75.00*	*67.76*
	RNN	*74.58*	*64.96*

Тұрақты сөйлеуді автоматты тану жағдайында RNN негізіндегі модельдеу арқылы қол жеткізілген нәтижелр айтарлықтай жақсартуларға негізделген, біз сондай-ақ өте сәйкес емес сөйлеуді автоматты тану контексті жағдайында бұл әдістерді зерттеуге кірістік. Осы мақсатта балалардың сынақ жинағы сөйлеуді автоматты тану жүйесі арқылы декодталған, онда акустикалық және тілдік модельдеу ересектерден алынған мәліметтерді пайдалана отырып оқытылды және осы зерттеудің нәтижелері II кестеде жинақталған. Балалардың дауыс жолы әлдеқайда қысқа болғандықтан, балалардың сөйлеуі арасында маңызды форманттық шкала бар. Бұл мәселені шешу үшін дауыс жолдарының ұзындығын

қалыпқа келтіру (VTLN) [15] балалардағы сөйлеуді автоматты тану жүйесінің сәйкессіздігі жағдайындағы акустикалық сәйкессіздікті азайтуда өте тиімді екендігіне қол жеткіздік. Сондықтан да балалардағы сәйкессіздікке тестілеуде VTLN көмегімен әрі қарай бағалау жүргізілді және бірдей шыққан нәтижелердің қорытындысы II кестеде көрсетілген. Кестеден байқағандай, елеулі сәйкессіздік болған жағдайда, зерттелетін тәсіл де салыстыру жағдайында көрсетілгенге ұқсас тенденцияларды байқауға болады.

V. Қорытынды

Бұл жұмыста біз аз ресурсты автоматты түрде сөйлеуді тану тапсырмасы бойынша қайталанатын нейрондық желілерге негізделген тілді модельдеумен бірге ұзақ қысқамерзімді жадыға негізделген акустикалық модельдеуді зерттедік. Эксперименттік бағалау сөйлеуді автоматты түрде тану тапсырмаларына сәйкес келетін және сәйкес келмейтін балалардың сөйлеу моделі үшін орындалды. Зерттеу мұндай қайталанатын нейрондық желілерге негізделген модельдеу жүйесі қазір қолданылып жүрген DNN-HMM негізіндегі модельдеуден тіпті ресурсы төмен тапсырмада да тиімді болуы мүмкін екендігін көрсетті.

References:

1 Kalimoldayev, M., Mamyrbayev, O., Mekebayev, N., Kydyrbekova, A. Algorithms for detection gender using neural networks // International Journal of Circuits, Systems and Signal Processing. 2020, 14, cmp. 154–159

2 Bridle J., "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," in Neuro-computing: Algorithms, Architectures and Applications, 1990, pp. 227–236.

3 Rabiner L. R., "A tutorial on hidden markov models and selected applications in speech recognition," Proceedings of the IEEE, vol. 77, no. 2, pp. 257–286, 1989.

4 Hinton G., Deng L., N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, Sainath T. N. et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82–97, 2015.

5 Graves A., Fernandez S., and Schmidhuber J., "Bidirectional lstm´ networks for improved phoneme classification and recognition," in International Conference on Artificial Neural Networks. Springer, 2015, pp. 799–804.

6 Oparin I., Sundermeyer M., Ney H., and Gauvain J. L., "Performance analysis of neural networks in combination with n-gram language models," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017.

7 Mikolov T., Karafiat M., Burget L., Cernock J. and Khudanpur S., "Recurrent neural network based language model." in Interspeech, vol. 2, 2017, p. 3.

8 Sak H., Senior A. W., and Beaufays F., "Long short-term memory recurrent neural network architectures for large scale acoustic modeling." in Interspeech, 2014, pp. 338–342

9 Chen X., Liu X., Qian Y., Gales M., and Woodland P. C., "CUEDRNNLM–An open-source toolkit for efficient training and evaluation of recurrent neural network language models," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 6000–6004.

10 Hochreiter S. and Schmidhuber J., "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 2017.

11 Povey D., Ghoshal A., Boulianne G., Burget L., Glembek O., Goel N., Hannemann M., Motlicek P., Qian Y., Schwarz P. et al., "The kaldi speech recognition toolkit," in Workshop on automatic speech recognition and understanding, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

12 Toma´s Mikolov L. B., Stefan Kombrink and Cernocky J., "RNNLM -` recurrent neural network language modeling toolkit."

13 Mamyrbayev O., Toleu A., Tolegen G., Mekebayev N. Neural Architectures for Gender Detection and Speaker Identification // Cogent Engineering, ISSN: 2331-1916. – 2020. Volume 7, - Issue 1

14 Rath S. P., Povey D., Vesely K., and Cernocky J., "Improved feature` processing for deep neural networks." in Interspeech, 2015, pp. 109– 113.

15 Eide E. and Gish H., "A parametric approach to vocal tract length normalization," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, 2016.