СПЕКТРАЛЬНЫ АНАЛІЗ І ВЕРЫФІКАЦЫЯ ІНТАНАЦЫЙНЫХ КОНТУРАЎ ПРЫ АПАРАТНА-ПРАГРАМНАЙ РЭАЛІЗАЦЫІ СІНТЭЗАТАРА МАЎЛЕННЯ

С. Ю. Крышнеў

Установа адукацыі «Гомельскі дзяржаўны ўніверсітэт імя Францыска Скарыны», Рэспубліка Беларусь

І. У. Блыхін

Установа адукацыі «Гомельскі дзяржаўны тэхнічны ўніверсітэт імя П. В. Сухога», Рэспубліка Беларусь

Навуковыя кіраўнікі: Ю. В. Крышнеў, М. У. Буракова

Установа адукацыі «Гомельскі дзяржаўны тэхнічны ўніверсітэт імя П. В. Сухога», Рэспубліка Беларусь

Разгледжаны метады ацэнкі якасці сінтэзаванага маўленчага сігналу на аснове спектральнага аналізу (STFT, мел-спектраграма, метрыка MCD, LPC-спектр). Апісана апаратна-праграмная рэалізацыя інтэрфейсу гукавой карты і прыведзены вынікі верыфікацыі суперсегментнай (прасадычнай) мадуляцыі шляхам параўнання асцылаграм і інтанаграм для розных тыпаў інтанацыйных канструкцый.

Ключавыя словы: сінтэз маўлення, спектральны аналіз, STFT, MCD, LPC-спектр, верыфікацыя, інтанацыйныя канструкцыі, FPGA.

Кантроль якасці сінтэзаванага маўленчага сігналу з'яўляецца неабходным этапам распрацоўкі любой сістэмы «тэкст-у-маўленне»: без аб'ектыўных крытэрыяў ацэнкі немагчыма пацвердзіць, што рэалізаваны алгарытм сапраўды забяспечвае натуральнае гучанне гаворкі і карэктную перадачу яе інтанацыйнага малюнку. У сучасных даследаваннях для гэтага выкарыстоўваюцца як спектральныя метады аналізу сігналу (кароткачасовае пераўтварэнне Фур'е, мел-спектраграмы, метрыка MCD), так і метады візуальнай верыфікацыі прасадычных параметраў – параўнанне асцылаграм і інтанаграм эталонных і сінтэзаваных выказванняў.

У праекце, матэрыялы якога пакладзены ў аснову дадзенага даклада, распрацавана сістэма сінтэзу маўленчых акустычных сігналаў з улікам інтанацыі выказвання; мэтай работы з'яўляецца распрацоўка праграмна-апаратнага комплексу з магчымасцю дынамічнага кіравання інтанацыйным малюнкам. У дакладзе раз-гледжаны этапы спектральнага аналізу сінтэзаванага сігналу, асноўныя рашэнні апаратна-праграмнай рэалізацыі інтэрфейсу гукавой карты, а таксама вынікі верыфікацыі атрыманых інтанацыйных контураў.

Спектральны аналіз сінтэзаванага сігналу. Для верыфікацыі якасці сінтэзу прымяняецца аналіз спектраграмы праз кароткачасовае пераўтварэнне Фур'е (STFT) з параметрамі, узгодненымі з вакодэрам: даўжыня акна $L = 512$ адлікаў; крок $H = 128$; акно Хана:

$$\left|S(l,\,k)\right|^2 = \left|\sum_{n=0}^{L-1} y(lH+n)w(n)e^{-\frac{j2\pi kn}{L}}\right|^2,$$

дзе k – частотны бін $(k = 0, \ldots, L/2)$; l – часавы фрэйм. Спектраграма дазваляе візуальна кантраляваць наяўнасць і размяшчэнне фармант F_1–F_3, правільнасць F_0-гармонік (гарызантальныя палосы з крокам f_0), плыўнасць пераходаў паміж фанемамі і карэктнасць паўз.

Для лікавай ацэнкі якасці сінтэзу выкарыстоўваецца метрыка MCD (Mel Cepstral Distortion), якая вымярае адлегласць паміж мел-кепстральнымі каэфіцыентамі сінтэзаванага і эталоннага сігналаў пры $D = 24$ (стандартнае значэнне). Значэнне MCD < 6 дБ прынята лічыць паказчыкам высокай якасці; для распрацаванай сістэмы дасягнута значэнне 5,2 ± 0,4 дБ на тэставай выбарцы (мал. 1), што пацвярджае блізкасць сінтэзаванага сігналу да эталоннага па спектральных характарыстыках.



Мал. 1. Мел-спектраграмы эталоннага
і сінтэзаванага сігналаў (MCD ≈ 5,2 дБ)

Дадаткова для аналізу фармантнай структуры прымяняецца метад LPC-спектра (Linear Predictive Coding) – гладкай спектральнай ахінальнай, атрыманай метадам лінейнага прадказання, устойлівай да шуму і прыдатнай для дакладнай лакалізацыі фармант. Парадак прадказання абіраецца згодна з правілам Маркеля ($p \approx f_s/1000 + 2 \approx 24$ пры $f_s = 22050$ Гц); фарманты вызначаюцца як карані палі-номнай функцыі A(z), размешчаныя зблізку адзінкавай акружнасці ў z-плоскасці, а іх цэнтральная частата і паласа прапускання вылічваюцца праз аргумент і модуль адпаведнага кораня.

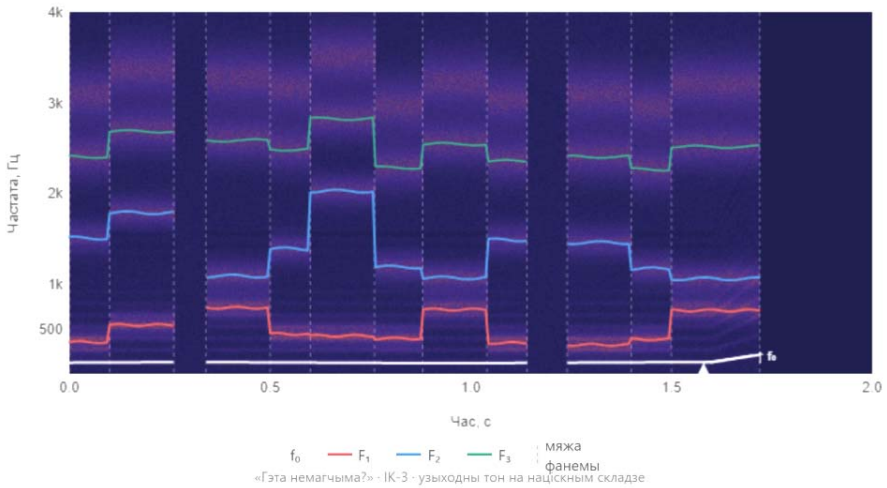
Апаратна-праграмная рэалізацыя інтэрфейсу гукавой карты. Праграмны вакодэр рэалізаваны на мове Python 3.11 з выкарыстаннем бібліятэк NumPy і SciPy для вектарных вылічэнняў і фільтрацыі, а таксама sounddevice – абгортцы над PortAudio для крос-платформавага вываду PCM-даных. Крытычныя па прадукцый-насці фрагменты (OLA-сінтэз, LPC-аналіз) рэалізаваны з выкарыстаннем кампілятара Numba JIT, што забяспечвае паскарэнне вылічэнняў у 20–40 разоў у параўнанні з чыстым Python.

Ключавой метрыкай прадукцыйнасці вакодэра з'яўляецца RTF (Real-Time Factor) – адноснасць часу вылічэнняў да працягласці сінтэзаванага сігналу; значэнне RTF < 1 азначае працу ў рэальным часе. Для поўнага пайплайна (лінгвістычны аналіз + + акустычная мадэль + вакодэр) без JIT-кампіляцыі RTF складае 147 %, у той час як прымяненне Numba JIT зніжае гэты паказчык да 31 %, забяспечваючы амаль трохразовы запас рэальнага часу.

Разгледжана таксама канцэптуальная магчымасць пераносу крытычных блокаў вакодэра на апаратныя платформы тыпу DSP (TI TMS320C6748) або FPGA (Xilinx Artix-7). Ацэнка паказвае, што ўсе блокі – FFT на 512 кропак, узбуджальнік на аснове фазавага акумулятара, крос-фэйд спектральных ахінальных, PSOLA-блок дублявання фрэймаў і інтэрфейс DAC/I2S – тэхнічна маюць быць рэалізаваныя на бюджэтнай FPGA з сумарным выкарыстаннем каля 7000 LUT з 20 800 даступных, што адпавядае прыкладна 35 % рэсурсаў мікрасхемы.

Верыфікацыя інтанацыйных контураў. Для дэманстрацыі ўкладу прасадычнай мадуляцыі параўноўваліся асцылаграмы трох варыянтаў сінтэзаванай фразы: без прасадычнай мадуляцыі ($f_0 = \text{const}, E = \text{const}$), з мадуляцыяй толькі па контуры f_0 і з поўнай прасадычнай мадуляцыяй ($f_0 + E +$ працягласці). Якасны аналіз паказвае, што першы варыянт мае манатонную ахінальную і ўспрымаецца як «рабатызаваная» гаворка, другі змяшчае прыкметны інтанацыйны контур, які ўспрымаецца як апавяданне, а трэці дадаткова вылучае націскны склад падвышанай амплітудай, што адпавядае акцэнтнай структуры слова.

Больш дэталевая верыфікацыя выканана шляхам параўнання інтанаграм для чатырох камунікатыўных тыпаў выказвання (мал. 2), сінтэзаваных на аснове аднаго і таго ж сегментнага матэрыялу і адрозных толькі прасадычным контурам f_0, у адпаведнасці з тыпалогіяй інтанацыйных канструкцый беларускага маўлення [4]. Параўнанне пацвярджае, што менавіта суперсегментны ўзровень цалкам вызначае камунікатыўны тып выказвання пры нязменным сегментным напаўненні, што сведчыць аб карэктнасці абранай архітэктуры з незалежнымі сегментным і прасадычным модулямі, звязанымі праз агульную часавую вось.



Мал. 2. Інтанаграмы розных тыпаў інтанацыйных канструкцый для адной і той жа фанетычнай паслядоўнасці

Дадаткова быў выкананы аналіз спектральнага партрэта прасадычных пераходаў на прыкладзе пытальнай фразы (ІК-3): паказана, што да націскнога склада значэнне f_0 захоўваецца практычна стацыянарным, на націскным складзе рэзка нарастае (узыходзячы тон), а на занацiскны складах спадае, суправаджаючыся зніжэннем інтэнсіўнасці; пры гэтым пераходы фармант на стыках фанем застаюцца плыўнымі, што пацвярджае карэктную працу спектральнага кросфэйда даўжыней 10 мс.

Праведзены спектральны аналіз (STFT, мел-спектраграмы, MCD = 5,2 ± 0,4 дБ) і ацэнка прадукцыйнасці (RTF = 0,31 з Numba JIT) пацвярджаюць, што рэалізаваная сістэма забяспечвае высокую якасць сінтэзу гаворкі ў рэальным часе. Верыфікацыя на аснове параўнання асцылаграм і інтанаграм паказала карэктную перадачу суперсегментных прасадычных характарыстык, а ацэнка апаратнай рэалізацыі на платформе FPGA класа Artix-7 (~35 % рэсурсаў) пацвердзіла тэхнічную магчымасць пераносу вакодэра на ўбудаваныя сістэмы.

Літаратура

1. Kubichek, R. Mel-cepstral distance measure for objective speech quality assessment / R. Kubichek // Proceedings of IEEE Pacific Rim Conference on Communications, Computers and Signal Processing. – 1993. – P. 125–128.
2. Markel, J. D. Linear Prediction of Speech / J. D. Markel, A. H. Gray. – Berlin : Springer-Verlag, 1976. – 288 p.
3. Morise, M. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications / M. Morise, F. Yokomori, K. Ozawa // IEICE Transactions on Information and Systems. – 2016. – Vol. E99-D, N 7. – P. 1877–1884.
4. Выгонная, Л. Ц. Інтанацыя. Націск. Арфаэпія / Л. Ц. Выгонная. – Мінск : Навука і тэхніка, 1991. – 215 с.
5. Xilinx Inc. Artix-7 FPGA Family Data Sheet: DC and AC Switching Characteristics. – San Jose: Xilinx, 2020.
6. Рабинер, Л. Теория и применение цифровой обработки сигналов : пер. с англ. / Л. Рабинер, Б. Голд ; под ред. Ю. Н. Александрова. – М. : Мир, 1978. – 848 с.

ПРАГРАМНАЯ РЭАЛІЗАЦЫЯ ВАКОДЭРА І МІКШЫРАВАННЕ СЕГМЕНТНЫХ І СУПЕРСЕГМЕНТНЫХ КАМПАНЕНТАЎ ПРЫ СІНТЭЗЕ МАЎЛЕННЯ

С. Ю. Крышнеў

Установа адукацыі «Гомельскі дзяржаўны ўніверсітэт імя Францыска Скарыны», Рэспубліка Беларусь

І. У. Блыхін

Установа адукацыі «Гомельскі дзяржаўны тэхнічны ўніверсітэт імя П. В. Сухога», Рэспубліка Беларусь

Навуковыя кіраўнікі: Ю. В. Крышнеў, М. У. Буракова

Установа адукацыі «Гомельскі дзяржаўны тэхнічны ўніверсітэт імя П. В. Сухога», Рэспубліка Беларусь

Разгледжана праграмная рэалізацыя вакодэра сістэмы сінтэзу маўленчых акустычных сігналаў, якая забяспечвае аднаўленне гукавога сігналу па мел-кепстральных каэфіцыентах з улікам кіравання суперсегментнымі (прасадычнымі) параметрамі выказвання – контурам асноўнага тону, інтэнсіўнасцю і працягласцю. Апісаны прынцыпы мікшыравання сегментнага і суперсегментнага патокаў даных пры сінтэзе гаворкі ў рэальным часе.

Ключавыя словы: сінтэз маўлення, вакодэр, мел-кепстральныя каэфіцыенты, прасадычная мадуляцыя, PSOLA, мікшыраванне, суперсегментны ўзровень, WASAPI.

Сучасныя сістэмы сінтэзу маўлення (text-to-speech – TTS) пераўтварыліся з дапаможнага сродку агучвання тэксту ў неад’емны кампанент галасавых памочнікаў, аўтаматычных даведачных сістэм і сродкаў агучвання інфармацыі для людзей з парушэннямі зроку. Пры гэтым якасць успрымання сінтэзаванай гаворкі вызначаецца