OVOZ KONVERTATSIYASIDA NOPARALLEL MA'LUMOTLARNI KENGAYTIRISH: DIQTOR AJRATUVCHI ARXITEKTURA ASOSIDAGI YONDASHUVNING TAHLILI

Subxonqulov Umidjon

Buxoro davlat universiteti

subxonqulovumidjon@gmail.com

Dolzarblik

Ovoz konvertatsiyasi (voice conversion) manba diqtorning nutqini uning lingvistik mazmunini o'zgartirmagan holda maqsadli diqtor ovoziga aylantirish vazifasi bo'lib, sun'iy intellekt asosidagi nutq texnologiyalarining tez rivojlanayotgan yo'nalishlaridan biri hisoblanadi. Amaliyotda ikkala diqtorning bir xil matnni o'qigan parallel audio yozuvlarini yig'ish deyarli imkonsiz bo'lgani sababli, noparallel ma'lumotlar asosida modellashtirish dolzarb ilmiy-amaliy muammoga aylangan. Mavjud yechimlar odatda diqtor va lingvistik mazmunni ajratish (disentanglement) yoki matndan sun'iy parallel audio yaratish (matn-nutq sintezi orqali) strategiyalariga tayanadi, biroq bunday sintez qilingan parallel ma'lumotlarda manba nutqning mahalliy talaffuz uslubi (prosodiya, energiya, fonema davomiyligi) yo'qolib, faqat maqsadli diqtorning uslubi saqlanib qoladi. Natijada konvertatsiya modeli haqiqiy nutq xususiyatlaridan emas, balki sun'iy, uslubiy jihatdan bir xillashtirilgan ma'lumotlardan o'rganishga majbur bo'ladi, bu esa konvertatsiya sifatining pasayishiga olib keladi. Shu sababli, manba nutqning mahalliy uslubini saqlagan holda vaqt bo'yicha to'g'ri tekislangan parallel ma'lumotlarni sun'iy generatsiya qilish usullarini ishlab chiqish muhim ilmiy ahamiyat kasb etadi.

Tadqiqot maqsadi

Tadqiqotning maqsadi noparallel ma'lumotlar asosida ovoz konvertatsiyasi uchun uchma-uch (end-to-end) ma'lumotlarni kengaytirish usulini ishlab chiqishdan iborat bo'lib, unda generatsiya qilingan parallel nutq nafaqat fonema

darajasida vaqt bo'yicha tekislangan, balki manba nutqning mahalliy talaffuz uslubini ham diqtor shaxsidan mustaqil ravishda saqlab qoladi. Bu maqsadga erishish uchun matn-nutq sintezi modeli ichida diqtor shaxsini ajratib oluvchi (speaker disentangler) mustaqil komponent joriy etilishi, shu bilan birga fonema davomiyligi va prosodik xususiyatlarning manba audioga mos kelishini ta'minlaydigan diqqat mexanizmiga asoslangan arxitektura yaratilishi ko'zda tutilgan.

Metodologiya

Taklif etilgan ParaGen tizimi ikki bosqichli arxitekturaga asoslanadi: birinchi bosqichda diqtor shaxsini ajratuvchiga ega matn-nutq sintezi modeli noparallel ma'lumotlarda o'qitiladi va undan foydalanib vaqt bo'yicha tekislangan parallel nutq juftliklari sintez qilinadi; ikkinchi bosqichda esa ushbu sun'iy parallel juftliklar asosida sodda freym-freym spektrogramma konvertatsiya tarmog'i (ikki qatlamli ikki yo'nalishli LSTM) o'qitiladi. Diqtor ajratuvchi ikki o'lchamli konvolyutsion qatlamlar orqali mel-spektrogrammadan freym darajasidagi "eskiz" (sketch) ni ajratib oladi, so'ngra bir o'lchamli konvolyutsion qatlamlar va gradient teskarilash (gradient reversal) texnikasi orqali diskriminativ qarama-qarshi o'qitish (discriminative adversarial training) asosida ushbu vakillashtiruvdan diqtorga oid ma'lumot yo'qotiladi. Alohida ilmiy yangilik sifatida, spektrogrammaning turli chastota zonalari (bandlari) uchun mahalliy joylashuv indeksini kirituvchi "joylashuv konvolyutsiyasi" (location-aware CNN) taklif etiladi bu chastota o'qi bo'yicha rezonans cho'qqilarining nolinye taqsimlanishini hisobga olishga xizmat qiladi va an'anaviy izotropik konvolyutsiyaga nisbatan qo'shimcha uslubiy barqarorlik beradi.

Fonema kontekstlari bilan diqtordan ajratilgan uslubiy vakillashtiruvlar joylashuv-sezgir diqqat mexanizmi orqali tekislanadi va shartli rekurrent dekoder yordamida maqsadli mel-spektrogrammaga dekodlanadi, bunda

kodlash va dekodlash bosqichlaridagi diqtor vektorlari mustaqil, alohida-alohida yangilanadigan parametrlar sifatida ishlatiladi. Bu arxitekturaviy qaror model inferens bosqichida kodlovchi vektorini havola diqtorga, dekodlovchi vektorini esa maqsadli diqtorga almashtirish orqali diqtor identifikatorini almashtirish (identity switching) imkonini yaratadi. Tajribalar CMU ARCTIC korpusidagi to'rtta diqtor (jinsi bo'yicha kesishma konvertatsiyasi) va yordamchi ma'lumot sifatida 108 diqtorli VCTK korpusi asosida o'tkazilgan bo'lib, taklif etilgan ikki variant (odatiy konvolyutsiyali Simp2D va joylashuv-indeksli Loc2D) fonema davomiyligi oldindan tekislangan an'anaviy statistik parametrik nutq sintezi (SPSS) usuli bilan solishtirilgan. Baholash ob'ektiv o'lchovlar (t-SNE vizualizatsiyasi, logF0 tahlili, diqtor klassifikatori aniqligi) va sub'ektiv eshittirish testlari (MUSHRA - tabiiylik, MOS - o'xshashlik) orqali amalga oshirilgan.

Asosiy natijalar

Tajriba natijalari shuni ko'rsatadiki, diskriminativ qarama-qarshi o'qitish diqtor klassifikatori aniqligini 108 diqtorli VCTK to'plamida tasodifiy taxmin darajasiga (≈0.01–0.03) yaqinlashtiradi, bu esa freym darajasidagi vakillashtiruvdan diqtor shaxsining samarali yo'qotilganini tasdiqlaydi. t-SNE vizualizatsiyasi va qatlamlar bo'yicha o'tkazilgan qo'shimcha klassifikatsiya tahlili diqtorga oid ma'lumotning aynan bir o'lchamli konvolyutsion qatlamlar zanjirida bosqichma-bosqich yo'qolishini ko'rsatadi. Kodlash va dekodlash diqtor vektorlarini mustaqil parametr sifatida ajratish natijasida olingan sun'iy nutqning asosiy chastota (pitch) konturi maqsadli diqtorga izchil moslashadi, aksincha, yagona umumiy diqtor vektoridan foydalanilganda konvertatsiyalangan nutqning ohang konturi manba diqtorga yaqinroq qolib ketishi kuzatilgan.

Sub'ektiv baholash natijalariga ko'ra, taklif etilgan ikkala variant (Simp2D va Loc2D) tabiiylik (naturalness) mezoni bo'yicha SPSS bazaviy tizimidan ustun

bo'lgan, ayniqsa ayoldan erkak diqtorga konvertatsiya yo'nalishida farq statistik jihatdan sezilarli darajada katta bo'lgan; diqtorlar o'xshashligi (similarity) mezoni bo'yicha esa taklif etilgan tizimlar bazaviy tizim bilan taqqoslanadigan darajada natija ko'rsatgan. Chastota-indeksli konvolyutsiya (Loc2D) asosiy chastota konversiyasining barqarorligini oshirib, noto'g'ri (ikki baravar oshirilgan yoki kamaytirilgan) pitch konvertatsiyasi holatlarini odatiy konvolyutsiyaga (Simp2D) nisbatan sezilarli darajada kamaytirgan, biroq bu barqarorlik yakuniy freym-freym konvertatsiya modelining umumiy tabiiylik ko'rsatkichiga sezilarli qo'shimcha ustunlik keltirmagan. Bundan tashqari, taklif etilgan usul asosida sintez qilingan parallel ma'lumotlarning vaqt bo'yicha izchilligi shu qadar yuqori bo'lganki, konvertatsiya tarmog'ini bitta yo'nalishli, atigi 64 tugunli LSTM qatlamigacha soddalashtirish ham barqaror pitch konvertatsiyasini ta'minlagan, bu esa SPSS asosida kengaytirilgan ma'lumotlarda kuzatilmagan natijadir.

Xulosa

O'tkazilgan tadqiqot noparallel ma'lumotlar asosida ovoz konvertatsiyasi muammosini matn-nutq sintezi ichiga integratsiyalashgan diqtor ajratuvchi va mahalliy uslub-diqtor ajratish tamoyili orqali yechishning samarali yo'lini ko'rsatadi. Ilmiy yangilik diqtor shaxsi va mahalliy talaffuz uslubini alohida vakillashtiruvchi, gradient teskarilashga asoslangan freym darajasidagi ajratuvchi hamda chastota zonalariga xos joylashuv-indeksli konvolyutsiya mexanizmining birgalikda qo'llanilishida namoyon bo'ladi. Amaliy ahamiyati shundaki, taklif etilgan yondashuv fonema darajasida tekislash zarurligini bartaraf etib, kichik hajmdagi noparallel diqtor ma'lumotlari asosida ham sifatli va nisbatan yengil konvertatsiya modellarini qurish imkonini beradi, bu esa ovozli assistentlar, dublyaj va shaxsiylashtirilgan nutq sintezi tizimlarida amaliy qo'llanilishi mumkin. Kelgusi tadqiqotlar diqtor ajratuvchining yangi, o'qitilmagan diqtorlarga umumlashtirilishini hamda prosodik xatolarni yanada kamaytiruvchi mexanizmlarni ishlab chiqishga qaratilishi maqsadga muvofiqdir.

FOYDALANILGAN ADABIYOTLAR

1. Chen, B., Xu, Z., & Yu, K. (2022). Data augmentation based non-parallel voice conversion with frame-level speaker disentangler. Speech Communication, 136, 14–22. https://doi.org/10.1016/j.specom.2021.10.001

2. Zhang, J.-X., Ling, Z.-H., & Dai, L.-R. (2019). Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 540–552.

3. Kameoka, H., Kaneko, T., Tanaka, K., & Hojo, N. (2018). StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks. Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT), 266–273.

4. Park, S., Kim, D., & Joe, M. (2020). Cotatron: Transcription-guided speech encoder for any-to-many voice conversion without parallel data. Proceedings of Interspeech 2020, 4696–4700.

5. Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., et al. (2018). Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4779–4783.

6. van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.

7. Lorenzo-Trueba, J., Yamagishi, J., Toda, T., Saito, D., Villavicencio, F., Kinnunen, T., & Ling, Z. (2018). The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. Proceedings of Odyssey 2018: The Speaker and Language Recognition Workshop, 195–202.

8. Saito, Y., Ijima, Y., Nishida, K., & Takamichi, S. (2018). Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors. Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5274–5278.

9. Yi, Z., Huang, W.-C., Tian, X., Yamagishi, J., Das, R. K., Kinnunen, T., Ling, Z., & Toda, T. (2020). Voice conversion challenge 2020 — Intra-lingual semi-parallel and cross-lingual voice conversion. Proceedings of the Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020, 80–98.

10. Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., et al. (2017). Tacotron: Towards end-to-end speech synthesis. Proceedings of Interspeech 2017, 4006–4010.