X SHO‘BA. KOMPYUTER LINGVISTIKASIDA SUN’IY INTELLEKT

UDK: 808.852

KO‘P TILLI AVTOMATIK NUTQNI TANISH TIZIMLARINI YARATISH MUAMMOLAR VA YECHIMLARI

Ochilov Mannon Musinovich,
Texnika fanlari bo‘yicha falsafa doktori, PhD
ochilov.mannon@mail.ru
TATU

Narzullaev Oybek Otabek o‘g‘li
Magistrant
oybeknarzullaev99@gmail.com
TATU

Annotatsiya. Ushbu maqolada ko‘p tilli avtomatik nutqni tanib olish (Multilanguage ASR) texnologiyasidagi muammolar va ularning yechimlari haqida so‘z yuritiladi. Maqolada shuningdek, kam resursli tillar uchun ma’lumotlar yetishmasligi, tillarning o‘zaro aralashuvi, latentsiya va dialektlar kabi muammolarni hal qilishning turli usullari va yondashuvlarining afzalliklari hamda kamchiliklari ko‘rib chiqiladi.

Abstract. This article discusses the problems and solutions related to multilanguage Automatic Speech Recognition (ASR) technology. The article also examines the advantages and disadvantages of various methods and approaches to address issues such as data scarcity for low-resource languages, code-switching, latency, and dialects.

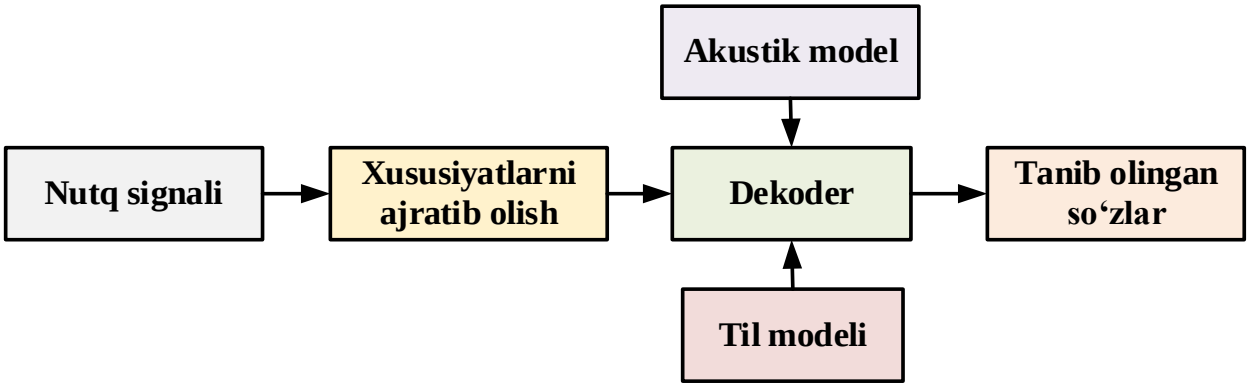
Абстракт. В этой статье рассматриваются проблемы и решения, связанные с многоязычным автоматическим распознаванием речи (multilanguage ASR). Также в статье анализируются преимущества и недостатки различных методов и подходов к решению таких проблем, как нехватка данных для языков с низкими ресурсами, переключение кодов, задержки и диалекты.

Kalit so‘zlar: *Ko‘p tilli ASR, nutqni tanib olish, kam resursli tillar, code-switching, end-to-end modellar, neyron tarmoqlar, SpeechLM, dialektlar.*

So‘nggi o‘n yilliklarda sun’iy intellekt va mashinaviy o‘qitish texnologiyalarining jadal rivojlanishi inson-nutq interfeyslari sohasida katta yutuqlarni keltirib chiqardi. Avtomatik nutqni tanib olish (Automatic Speech Recognition, ASR) tizimlari bugungi kunda virtual yordamchilardan tortib, real vaqtda tarjimaga qadar keng qo‘llanilmoqda[1,2]. Masalan, OpenAI tomonidan

ishlab chiqilgan Whisper modeli yoki Meta kompaniyasining SeamlessM4T tizimi ko‘p tilli real vaqtda tarjima imkoniyatlari orqali global muloqotni osonlashtirmoqda[17,18]. Biroq, global miqyosda ko‘p tilli ANT (multilanguage ASR) tizimlarini joriy etish va ularning samaradorligini oshirish hali ham murakkab va dolzarb muammo sifatida qolmoqda.

Dunyoda 7000 dan ortiq til mavjud bo‘lib, ularning har biri o‘ziga xos fonetik, grammatik va leksik xususiyatlarga ega. Shu bilan birga, zamonaviy jamiyatda tillarning o‘zaro ta’siri, globalizatsiya va ko‘p tilli muhitda muloqot qilish zarurati ko‘p tilli ANT texnologiyalariga bo‘lgan talabni yanada oshirmoqda.



1-rasm. ANT ning tizim arxitekturasi

Ko‘p tilli ANT ning asosiy maqsadi – turli tillarda nutqni aniq tanib olish, uni matnga aylantirish va foydalanuvchilar uchun tabiiy, samarali muloqot imkonini yaratishdir. Bu texnologiya nafaqat kundalik hayotda, balki ta’lim, tibbiyot, biznes va xalqaro aloqalar kabi sohalarda ham muhim ahamiyatga ega. Masalan, ko‘p tilli ANT tizimlari o‘qituvchilarga turli tillarda dars o‘tishda yordam berishi, shifokorlarga bemorlar bilan til to‘siqlarisiz muloqot qilish imkonini berishi yoki xalqaro konferensiyalarda real vaqtda tarjimani ta’minlashi mumkin. Biroq, ushbu imkoniyatlarni to‘liq ro‘yobga chiqarish uchun bir qator jiddiy muammolarni hal qilish zarur.

Hozirgi kunda ANT texnologiyalari asosan yuqori resursli tillar (masalan, ingliz, xitoy, ispan) uchun yaxshi optimallashtirilgan bo‘lib, Google Speech-to-Text, Amazon Transcribe yoki OpenAI Whisper kabi tizimlar ushbu tillarda yuqori aniqlik ko‘rsatadi. Biroq, dunyo tillarining aksariyati kam resursli hisoblanadi, ya’ni ular uchun yetarli audio va matn ma’lumotlari mavjud emas. Bundan tashqari, foydalanuvchilarning bir nechta tilni aralashtirib gapirishi (code-switching), dialektlarning xilma-xilligi, real vaqtda ishlashdagi kechikishlar va nutqning paralingvistik jihatlarini yo‘qotish kabi muammolar ko‘p tilli ANT tizimlari ning keng qo‘llanilishiga to‘sqinlik qilmoqda[8]. Masalan, o‘zbek tilida gapiruvchi

foydalanuvchi ruscha so‘zlarni aralashtirsa yoki xorazm shevasida nutq so‘zlasa, mavjud tizimlar ko‘pincha noto‘g‘ri yoki umuman javob bera olmaydi.

Ushbu muammolar nafaqat texnik jihatdan, balki ijtimoiy va madaniy nuqtai- nazardan ham katta ahamiyatga ega. Agar ko‘p tilli ANT tizimlari faqat bir nechta dominant tillarga xizmat qilsa, bu kam resursli tillarning raqamli dunyoda yo‘qolib borishiga olib kelishi mumkin. Shu sababli, ushbu texnologiyani rivojlantirish nafaqat foydalanuvchi tajribasini yaxshilash, balki lingvistik xilma-xillikni saqlash va global miqyosda teng huquqli muloqotni ta’minlash uchun ham muhimdir.

Mazkur maqola ko‘p tilli ANT texnologiyasidagi asosiy muammolarni chuqur tahlil qilishga va ularga qarshi innovatsion yechimlarni taklif qilishga qaratilgan. Tadqiqotda mavjud tizimlarning cheklovlari o‘rganiladi, ularning samaradorligi sinov ma’lumotlari (masalan, o‘zbek, turk va ingliz tillari aralashmasi) asosida baholanadi va yangi yondashuvlarning afzalliklari isbotlanadi.

Tadqiqotning asosiy maqsadi ko‘p tilli ANT tizimlarini yanada aniq, tezkor va tabiiy qilish yo‘llarini aniqlash hamda kam resursli tillar uchun qamrovni kengaytirishga hissa qo‘shishdir. Ushbu ish nafaqat texnologik jihatdan, balki global muloqotning kelajagini shakllantirish nuqtai-nazaridan ham muhim natijalarga olib kelishi kutilmoqda.

Ko‘p tilli ANTning asosiy muammolari. Ma’lumotlarning yetishmasligi, tillarning aralashuvi, texnik cheklovlar va nutqning tabiiy jihatlarini yo‘qotish kabi omillar foydalanuvchi tajribasini pasaytiradi va tizimlarning samaradorligini cheklaydi bu esa ko‘p tilli ANT tizimlari uchun bir qancha muammolarni keltirib chiqaradi. Quyida ko‘p tilli ANT tizimlarining global miqyosda keng qo‘llanilishiga to‘sqinlik qiladigan asosiy muammolar yoritib beriladi.

1. **Ma’lumotlarning cheklanganligi.** Ko‘p tilli ANT tizimlari yuqori sifatli audio va matn ma’lumotlariga tayanadi. Biroq, dunyodagi 7000 dan ortiq tillarning aksariyati, ayniqsa kam resursli tillar (masalan, o‘zbek, qirg‘iz, suahili) uchun yetarli ma’lumotlar mavjud emas. Bu modellar o‘qitilishi uchun zarur bo‘lgan katta hajmdagi datasetlarning yo‘qligi bilan bog‘liq. Shuning uchun ko‘p tillarda yozma matnlar raqamlashtirilmagan, audio yozuvlar esa professional tarzda to‘planmagan. Masalan, o‘zbek tilida nutqni tanib olish uchun millionlab soatlik audio ma’lumotlar kerak bo‘lsa-da, mavjud resurslar minglab soatlar bilan cheklangan. Ma’lumotlarning kamligi tufayli modellar umumlashtirish qobiliyatini yo‘qotadi va faqat ma’lum bir kontekst yoki dialektda yaxshi ishlaydi. Masalan, toshkent shevasida o‘qitilgan model namangan shevasini tushunmasligi mumkin.

2. **Tillarning o‘zaro aralashuvi (Code-Switching).** Ko‘p tilli muhitda foydalanuvchilar bir gap ichida bir nechta tilni aralashtirishi odatiy holat. Masalan, “*Men bugun meetingga boraman*” (o‘zbek va ingliz tillari aralashmasi). ANT

tizimlari bunday holatlarda qaysi tilga ustunlik berishni bilmay qoladi. Chunki modellar odatda bitta til uchun optimallashtiriladi va til chegaralarini aniqlashda qiynaladi. Bundan tashqari, o‘qitish ma’lumotlari ko‘pincha “sof” tillarga asoslanadi, aralash nutq esa kam uchraydi. Natijada tizim noto‘g‘ri transkripsiya qiladi yoki umuman javob bera olmaydi. Masalan, “*Salom, how are you?*” degan gapni tanib olishda model “*Salom*”ni o‘zbekcha, qolgan qismini inglizcha deb farqlay olmay, xato qiladi. Yana boshqa misol, Hindistonda hind va ingliz tillarini aralashtirib gapiradigan foydalanuvchilar uchun ANT tizimlari ko‘pincha chalkashib qoladi, chunki til o‘tishlari oldindan bashorat qilinmagan.

3. Yuqori latentsiya va xatolarning to‘planishi. An’anaviy ko‘p tilli ANT tizimlari bir nechta bosqichlardan iborat: nutqni tanib olish (ANT), matnni qayta ishlash (Large language model, LLM), va nutqqa aylantirish (Text to speech, TTS). Har bir bosqichda vaqt sarflanadi va xatolar yig‘iladi. Sababi har bir modul alohida ishlaydi va ma’lumotlar uzatilishi jarayonida kechikishlar yuzaga keladi. Masalan, nutqni matnga aylantirishda kichik xato keyingi bosqichda katta muammoga aylanadi. Real vaqtda suhbat qilish imkoni cheklanadi. Masalan, bir foydalanuvchi o‘zbek tilida savol bersa, tizim uni ingliz tilida noto‘g‘ri talqin qilib, javobni boshqa kontekstda berishi mumkin. Holbuki bu latentsiya esa suhbatning tabiiyligini yo‘qotilishiga olib keladi.

4. Paralingvistik ma’lumotlarning yo‘qolishi. Nutqni matnga aylantirish jarayonida ohang, intonatsiya, emotsiya va urg‘u kabi paralingvistik elementlar hisobga olinmaydi. Bu esa ko‘p tilli ANT tizimlarini “sovuq” va tabiiy bo‘lmagan qiladi. Chunki ko‘p tizimlar faqat fonetik transkripsiyaga e’tibor beradi, lekin nutqning hissiy yoki kontekstual jihatlari uchun maxsus o‘qitilmagan bo‘ladi. Masalan, “*Yaxshi*” so‘zi istehzoli yoki samimiy ohangda aytilganini farqlash qiyin. Natijada foydalanuvchi niyatini noto‘g‘ri tushunish yoki javobning noadekvat bo‘lishi kelib chiqadi. Masalan, “*Nima bo‘ldi?*” degan savol g‘azab bilan yoki qiziqish bilan aytilishi mumkin, lekin tizim buni ajrata olmaydi. Misol ingliz tilida “*Sure*” so‘zi turli intonatsiyalar bilan ijobiy yoki salbiy ma’no bildirishi mumkin, ammo ANT tizimlari buni matnda aks ettira olmaydi.

5. Dialektlar va aksentlarning turli-tumanligi. Har bir til ichida dialektlar va aksentlar mavjud bo‘lib, ular nutqni tanib olishni qiyinlashtiradi. Masalan, o‘zbek tilida Xorazm shevasi Toshkent shevasidan sezilarli darajada farq qiladi. Chunki Modellar ko‘pincha standartlashtirilgan til versiyalarida o‘qitiladi, lekin real hayotda odamlar o‘ziga xos talaffuz va so‘z birikmalaridan foydalanadi. Buning natijasida dialekt yoki aksentni tushunolmagan tizim noto‘g‘ri transkripsiya qiladi yoki umuman ishlamaydi. Misol tariqasida O‘zbek tilida “kel” so‘zi Xorazmda esa “gal” deb talaffuz qilinishi mumkin, bu esa tizimni chalkashtiradi.

Ko‘p tilli ANT muammolari uchun yechimlar. Ko‘p tilli modellar, end-to-end yondashuvlar, paralingvistik tahlil va dialektga moslashuv kabi usullar tizimlarning aniqligini oshiradi. Latentsiyani kamaytiradi va foydalanuvchi tajribasini yaxshilaydi. Quyida ko‘p tilli ANTning asosiy muammolarini hal qilishga qaratilgan asosiy yondashuv va yechimlar keltiriladi.

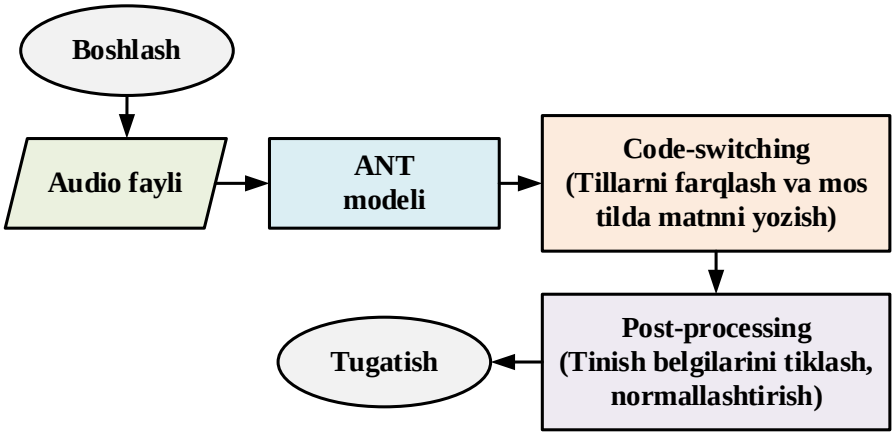
1. Ko‘p tilli modellar va sintetik ma’lumotlar yaratish. Kam resursli tillar uchun ma’lumotlar yetishmasligini bartaraf etish uchun ko‘p tilli modellar birgalikda o‘qitiladi va sintetik ma’lumotlar generatsiyasi qo‘llaniladi. Masalan, J. Devlin va boshqalar o‘z tadqiqotida ko‘p tilli pre-training usulidan foydalanib, kam resursli tillarda matn tahlilining aniqligini 15-20% ga oshirgan[3]. O‘zbek tili uchun ham bundan foydalanilsa, ingliz tilidagi katta ma’lumotlardan transfer learning orqali o‘zbekcha audio tahlilini yaxshilash mumkin. A. Baevski va boshqalar o‘z-o‘zini nazorat qiluvchi o‘qitish usuli bilan oz miqdordagi belgilangan ma’lumotlardan foydalanib, WERni 30% dan 15% ga tushirgan[4]. Bu bilan o‘zbek tili uchun, mavjud 1500 soatlik audio bilan modelni samarali o‘qitish imkoniyati paydo bo‘ladi. G. Hinton va boshqalar sintetik nutq generatsiyasi orqali ma’lumotlar hajmini 50% ga oshirib, nutq sintezini tabiiyligini yaxshilagan[7]. Xuddi shunday o‘zbek tili uchun, sintetik o‘zbekcha audio yaratib, ma’lumotlar bazasini kengaytirish mumkin. Quyida eng mashhur ko‘p tilli modellarga misollar keltirilgan:

1-jadval. Eng keng tarqalgan ko‘p tilli modellar

Model nomi	Ishlab chiqqan kompaniya	Qo‘llab-quvvatlaydigan tillar soni
mBERT (Multilingual BERT)	Google	~104 ta til
XLM-R (XLM-RoBERTa)	Facebook AI (Meta)	100+ til
mT5 (Multilingual T5)	Google	101 ta til
BLOOM	BigScience	46+ til
mGPT (Multilingual GPT)	Sber AI	40+ til
ByT5	Google	Har qanday til
NLLB-200 (No Language Left Behind)	Meta AI	200+ til
SeamlessM4T	Meta AI	Matn va nutq uchun
LaBSE (Language-agnostic BERT Sentence Embedding)	Google	109 til
GPT-4 Turbo (ko‘p tilli imkoniyatlar)	OpenAI	50+ til

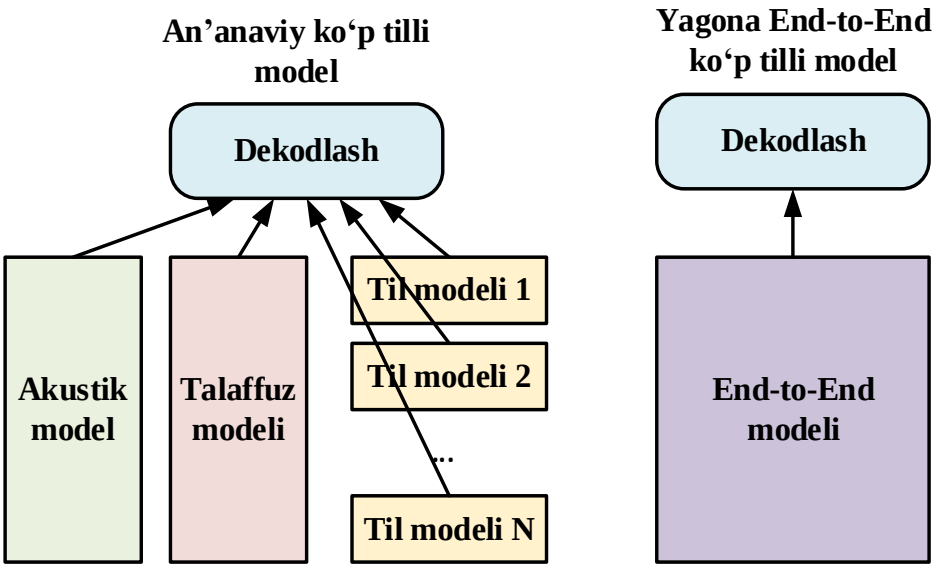
2. Code-Switching uchun maxsus modellar. Tillarni aralashtirib gapirishni aniqlash uchun maxsus ko‘p tilli modellar ishlab chiqiladi. Bunga misol qilib D. Serdyuk va boshqalar end-to-end modellar yordamida hind-ingliz code-switching holatlarida WERni 25% ga kamaytirgan[11]. Bu usuldan foydalanish orqali, “Men meetingga boraman” kabi aralash gaplarni aniq tanib olinishi mumkin. G. Winata va boshqalar ikki tilli so‘z vektorlaridan foydalanib, til o‘tishlarini

aniqlashda 20% aniqlik oshirgan[16]. Bu esa o‘zbek va ingliz tillari aralashmasini tahlil qilishda samaradorlikni oshirish mumkin. T. Schultz va boshqalar ko‘p tilli akustik modellar bilan aralash nutqni tanib olishda 15% yaxshilanishga erishgan[10]. O‘zbek tili uchun bunday foydalansak, o‘zbek-rus yoki o‘zbek-ingliz aralashmalarini yaxshiroq boshqarish imkoni bo‘ladi.



2-rasm. Code-switchingga asoslangan ko‘p tilli ANT ning umumiy arxitekturasi

3. End-to-End Speech Language Models (SpeechLMs). Nutqdan nutqqa to‘g‘ridan-to‘g‘ri o‘tadigan modellar latentsiyani kamaytiradi va xatolarni minimallashtiradi. Chunonchi A. Gulati va boshqalar Conformer modeli yordamida latentsiyani 20-30% ga qisqartirib, WERni 10% ga kamaytirgan[6]. Bu orqali real vaqtda o‘zbekcha suhbatlarda ham kechikishni kamaytirish mumkin. Y. Wang va boshqalar SpeechLM bilan nutqdan nutqqa o‘tishda xatolarni 15% ga qisqartirgan[13]. Bu O‘zbekcha savol-javob tizimlarini tabiiyroq va tezkor qilish imkoni paydo qiladi. S. Watanabe va boshqalar end-to-end ANT bilan ko‘p tilli tizimlarda latentsiyani 25% ga kamaytirgan[14]. O‘zbek tili uchun bundan foydalanilsa, o‘zbek tilidagi tarjimani real vaqtda yaxshilash imkoni yaratiladi.



3–rasm. An’anaviy va End-to-End ko‘p tilli modellarining umumiy tuzilishi.

4. Emotsiya va intonatsiyani tanib oluvchi modellar. Nutqning ohang va emotsiyasini saqlab qolish uchun paralingvistik tahlil qo‘shiladi. F. Weninger o‘z tadqiqotida deep learning yordamida nutqdan emotsiyani aniqlashda 85% aniqlikka erishgan[15]. O‘zbek tili uchun bunday foydalansak, “Yaxshi” so‘zining istehzoli yoki samimiy ohangini farqlash mumkin. Xususan B. McFee o‘z tadqiqotida audio signalning pitch va energy xususiyatlarini tahlil qilib, paralingvistik ma’lumotlarni 90% aniqlikda saqlagan[9]. O‘zbek tili uchun bu nutqning intonatsiyasini matnga o‘tkazish imkoni berishi mumkin. E. Cambria va b. emotsional tahlil bilan nutqni tushunishda 20% yaxshilanishga erishgan[5].

5. Moslashuvchan modellar. Dialekt va aksentlarni tanib olish uchun maxsus datasetlar bilan moslashuvchan modellar o‘qitiladi. Masalan D. Snyder va b. X-Vector texnologiyasi bilan aksentlarni aniqlashda 95% aniqlikka erishgan[12]. Bu yondashuv bizga toshkent va xorazm shevalarini farqlash imkonini berishi mumkin. M. Karafiat va b. bottleneck xususiyatlari bilan dialekt tanib olishda 30% yaxshilanishga erishgan. Bu orqali turli shevalarni tahlil qilishda aniqlik oshadi. deep neural networks bilan aksentlarga moslashishda 25% samaradorlikka erishgan[8].

Yuqoridagi yechimlar yordamida o‘zbek tili uchun ushbu usullarni qo‘llash orqali ko‘p tilli ANT tizimlarini ishlab chiqish va samaradorligini oshirish, kam resursli til sifatidagi cheklovlarni bartaraf etish va foydalanuvchi tajribasini yaxshilash mumkin.

Foydalanilgan adabiyotlar:

1. Mukhamadiyev, A., Mukhiddinov, M., Khujayarov, I., Ochilov, M., & Cho, J. (2023). Development of Language Models for Continuous Uzbek Speech Recognition System. *Sensors*, *23*(3), 1145. https://doi.org/10.3390/ s23031145
2. Musaev, M., Khujayorov, I., Ochilov, M. (2021). Automatic Recognition of Uzbek Speech Based on Integrated Neural Networks. In: Aliev, R.A., Yusupbekov, N.R., Kacprzyk, J., Pedrycz, W., Sadikoglu, F.M. (eds) 11th World Conference “Intelligent System for Industrial Automation” (WCIS-2020). WCIS 2020. Advances in Intelligent Systems and Computing, vol 1323. Springer, Cham. https://doi.org/10.1007/978-3-030-68004-6_28
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv. https://arxiv.org/abs/1810.04805
4. Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. arXiv. https://arxiv.org/abs/2006.11477
5. Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2017). Sentic computing: A common-sense-based framework for concept-level sentiment analysis. IEEE Computational Intelligence Magazine, 12(2), 14–23. https://doi.org/10.1109/MCI.2017.2670544
6. Gulati, A., Qin, J., Chiu, C. C., Parmar, N., Zhang, Y., Yu, J., ... & Pang, R. (2020). Conformer: Convolution-augmented Transformer for speech recognition. arXiv. https://arxiv.org/abs/2005.08100
7. Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 29(6), 82–97.
8. Karafiát, M., Burget, L., Cernocký, J., & Grezl, F. (2016). Multilingual bottleneck features for language and dialect recognition. Computer Speech & Language, 46, 252–267. https://doi.org/10.1016/j.csl.2017.03.002
9. McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in Python. Proceedings of the 14th Python in Science Conference, 8, 18–24. https://doi.org/10.25080/Majora-7b98e3ed-003
10. Schultz, T., Waibel, A., & others. (2006). Multilingual and crosslingual speech recognition. In Handbook of Multilingualism and Multilingual Communication (pp. 281–298). De Gruyter Mouton.
11. Serdyuk, D., Wang, Y., Bengio, Y., & others. (2018). Towards end-to-end spoken language understanding. arXiv. https://arxiv.org/abs/1802.08395

12. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust DNN embeddings for speaker recognition. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5329–5333. https://doi.org/10.1109/ICASSP.2018.8461375

13. Wang, Y., Huang, P. S., Gu, J., & others. (2022). SpeechLM: Enhanced speech pre-training for spoken language understanding. arXiv. https://arxiv.org/abs/2207.04672

14. Watanabe, S., Hori, T., Kim, S., Hershey, J. R., & Hayashi, T. (2018). Hybrid CTC/attention architecture for end-to-end speech recognition. IEEE Journal of Selected Topics in Signal Processing, 11(8), 1240–1253. https://doi.org/10.1109/JSTSP.2017.2763455

15. Weninger, F., Eyben, F., Schuller, B., Mortillaro, M., & Scherer, K. (2013). On the acoustics of emotion in audio: What speech signals reveal about our feelings. Frontiers in Psychology, 4, 738. https://doi.org/10.3389/fpsyg. 2013.00738

16. Winata, G. I., Cahyawijaya, S., Lin, Z., Liu, Z., & Fung, P. (2019). Code-switching language modeling with bilingual word embeddings. arXiv. https://arxiv.org/abs/1909.03460

17. Radford, A., Gao, L., Behrmann, E., Brockman, G., & Sutskever, I. (2022). *Whisper: Robust speech recognition via large-scale weak supervision*. OpenAI. https://openai.com/research/whisper

18. Zhang, X., Fan, A., Siddhant, A., Koehn, P., & Chaudhary, V. (2023). *SeamlessM4T: Massively multilingual & multimodal machine translation*. Meta AI Research. https://arxiv.org/abs/2308.04620