Shikshan Sanshodhan : Journal of Arts, Humanities and Social Sciences ISSN(o): 2581-6241
Monthly, Peer-Reviewed, Refereed, Indexed Journal Impact Factor: 6.471
Volume - 6, Issue - 12, December - 2023 Publication Date: 31/12/2023

DOIs:10.2018/SS/202312003 --:-- Research Paper / Article / Review

SPEECH RECOGNITION: कम्प्यूटेशनल भाषाविज्ञान

Dr. Rajendra Kumar Mahto[1]
Assistant Professor
Department Of Information Technology,
Dr Shayama Prasad Mukherjee
University, Ranchi
Email - rajendrabit57@gmail.com

Urmila Kumari[2]
Research Scholar
Department Of Khortha (TRL),
DSPMU, Ranchi
Email - urmilarkm94@gmail.com

Abstract: वाक् पहचान (एसआर) कम्प्यूटेशनल भाषाविज्ञान का अंतर-अनुशासनात्मक उप-क्षेत्र है जो कार्यप्रणाली और प्रौद्योगिकियां विकसित करता है जो कंप्यूटर द्वारा पाठ में बोली जाने वाली भाषा की मान्यता और अनुवाद को सक्षम बनाता है। भाषण मान्यता" या सिर्फ "भाषण से पाठ" (एसटीटी) के रूप में भी जाना जाता है। यह भाषा विज्ञान, कंप्यूटर विज्ञान और इलेक्ट्रिकल इंजीनियरिंग क्षेत्रों में ज्ञान और अनुसंधान को शामिल करता है। प्रौद्योगिकी के दृष्टिकोण से, भाषण मान्यता का प्रमुख नवाचारों की कई तरंगों के साथ एक लंबा इतिहास है। वाक् पहचान (या कभी-कभी स्वचालित भाषण मान्यता के रूप में संदर्भित) एक ऐसी प्रक्रिया है जिसके द्वारा कंप्यूटर (या अन्य प्रकार की मशीन) बोले गए शब्दों की पहचान करता है। दोनों ध्वनिक मॉडलिंग और भाषा मॉडलिंग आधुनिक सांख्यिकीय-आधारित भाषण मान्यता एल्गोरिदम के महत्वपूर्ण हिस्से हैं। भाषण मान्यता प्रणालियों को कई अलग-अलग वर्गों में यह वर्णन करके अलग किया जा सकता है कि उनके पास किस प्रकार के उच्चारण को पहचानने की क्षमता है। कुछ एएसआर सिस्टम विशिष्ट उपयोगकर्ताओं की पहचान करने की क्षमता रखते हैं। यह दस्तावेज़ सत्यापन या सुरक्षा प्रणालियों को कवर नहीं करता है।

Key Words: वाक् पहचान, मॉडल, एल्गोरिदम, मार्कोव मॉडल , गतिशील समय, तंत्रिका नेटवर्क , ध्वनिक मॉडलिंग , डिकोडिंग , डीप फीडफोरवर्ड , रिकरंट न्यूरल नेटवर्क्स , ऑटोमैटिक स्पीच रिकॉग्निशन , ऑडियो, भाषा मॉडल, लेक्सिकॉन , व्याकरण , भाषण मान्यता

1. परिचय:

क्या आपने कभी अपने कंप्यूटर से बात की है? यदि आपके पास है, तो आपने एक तकनीक का उपयोग किया है जिसे भाषण मान्यता के रूप में जाना जाता है। वाक् पहचान आपको अपनी आवाज के साथ एक सिस्टम को इनपुट प्रदान करने की अनुमति देती है। जैसे अपने माउस से क्लिक करना, अपने कीबोर्ड पर टाइप करना, या फोन कीपैड पर एक कुंजी दबाने से एप्लिकेशन को इनपुट मिलता है, वाक् पहचान आपको बात करके इनपुट प्रदान करने की अनुमति देती है। डेस्कटॉप की दुनिया में, आपको ऐसा करने में सक्षम होने के लिए एक माइक्रोफोन की आवश्यकता होती है। वाक् पहचान (एसआर) कम्प्यूटेशनल भाषाविज्ञान का अंतर-अनुशासनात्मक उप-क्षेत्र है जो कार्यप्रणाली और प्रौद्योगिकियां विकसित करता है जो कंप्यूटर द्वारा पाठ में बोली जाने वाली भाषा की मान्यता और अनुवाद को सक्षम बनाता है। इसे "स्वचालित भाषण मान्यता" (एएसआर), "कंप्यूटर भाषण मान्यता" या सिर्फ "भाषण से पाठ" (एसटीटी) के रूप में भी जाना जाता है। यह भाषा विज्ञान, कंप्यूटर विज्ञान और इलेक्ट्रिकल इंजीनियरिंग क्षेत्रों में ज्ञान और अनुसंधान को शामिल करता है।

कुछ एसआर सिस्टम "प्रशिक्षण" (" नामांकन " भी कहा जाता है) का उपयोग करते हैं, जहां एक व्यक्तिगत स्पीकर सिस्टम में पाठ या पृथक शब्दावली पढ़ता है। सिस्टम व्यक्ति की विशिष्ट आवाज का विश्लेषण करता है और इसका उपयोग उस व्यक्ति के भाषण की मान्यता को ठीक करने के लिए करता है, जिसके परिणामस्वरूप सटीकता में वृद्धि होती है। वे सिस्टम जो प्रशिक्षण का उपयोग नहीं करते हैं उन्हें "स्पीकर स्वतंत्र" सिस्टम कहा जाता है। प्रशिक्षण का उपयोग करने वाले सिस्टम को "स्पीकर निर्भर" कहा जाता है। वाक् पहचान अनुप्रयोगों में वॉयस डायलिंग (जैसे "कॉल

होम"), कॉल रूटिंग (जैसे "मैं एक कॉल करना चाहूंगा"), डोमेस्टिक उपकरण नियंत्रण, खोज (जैसे एक पॉडकास्ट खोजें जहां विशेष शब्द बोले गए थे) जैसे वॉइस यूजर इंटरफेस शामिल हैं), सरल डेटा प्रविष्टि (जैसे, क्रेडिट कार्ड नंबर दर्ज करना), संरचित दस्तावेजों की तैयारी (जैसे रेडियोलॉजी रिपोर्ट), भाषण-से-पाठ प्रसंस्करण (जैसे, वर्ड प्रोसेसर या ईमेल), और विमान (आमतौर पर डायरेक्ट वॉयस इनपुट कहा जाता है) । स्पीकर को पहचानना उन प्रणालियों में भाषण का अनुवाद करने के कार्य को सरल बना सकता है जो किसी विशिष्ट व्यक्ति की आवाज़ पर प्रशिक्षित किए गए हैं या इसका उपयोग किसी सुरक्षा प्रक्रिया के भाग के रूप में स्पीकर की पहचान को प्रमाणित करने या सत्यापित करने के लिए किया जा सकता है।

2. इतिहास:

प्रौद्योगिकी के दृष्टिकोण से, भाषण मान्यता का प्रमुख नवाचारों की कई तरंगों के साथ एक लंबा इतिहास है। हाल ही में, क्षेत्र को गहन शिक्षा और बड़े डेटा में प्रगति से लाभ हुआ है। अग्रिमों को न केवल क्षेत्र में प्रकाशित अकादमिक पत्रों के उछाल से स्पष्ट किया जाता है, बल्कि दुनिया भर में उद्योग द्वारा भाषण मान्यता प्रणालियों को डिजाइन और तैनात करने में विभिन्न प्रकार के गहन सीखने के तरीकों को अपनाना महत्वपूर्ण है। ये भाषण उद्योग खिलाड़ियों गूगल, माइक्रोसॉफ्ट, आईबीएम, एप्पल, अमेज़न शामिल सीडीएसी जिनमें से कई गहरी सीखने के आधार पर किया जा रहा के रूप में उनके भाषण मान्यता प्रणाली में मूल प्रौद्योगिकी प्रचारित किया है।

1952 में तीन बेल लैब्स के शोधकर्ताओं ने सिंगल-स्पीकर डिजिट रिकग्निशन के लिए एक सिस्टम बनाया। उनके सिस्टम ने प्रत्येक उच्चारण के पावर स्पेक्ट्रम में फ़ॉर्मेंट का पता लगाकर काम किया। 1950 के दशक की तकनीक एकल स्पीकर सिस्टम तक सीमित थी जिसमें लगभग दस शब्दों की शब्दसंग्रह थी।

गुन्नार फंट ने भाषण उत्पादन का स्रोत-फ़िल्टर मॉडल विकसित किया और इसे 1960 में प्रकाशित किया, जो भाषण उत्पादन का एक उपयोगी मॉडल साबित हुआ।

राज रेड्डी 1960 के दशक के अंत में स्टैनफोर्ड विश्वविद्यालय में स्नातक छात्र के रूप में निरंतर भाषण मान्यता लेने वाले पहले व्यक्ति थे। पिछले सिस्टम को प्रत्येक शब्द के बाद उपयोगकर्ताओं को एक विराम देने की आवश्यकता थी। रेड्डी की प्रणाली शतरंज के खेल के लिए बोली जाने वाली कमांड जारी करने के लिए डिज़ाइन की गई थी।

1971 में, DARPA ने अपने भाषण भाषण अनुसंधान कार्यक्रम के माध्यम से पाँच वर्षों के भाषण मान्यता अनुसंधान को महत्वाकांक्षी अंतिम लक्ष्यों के साथ वित्तपोषित किया, जिसमें 1,000 शब्दों का न्यूनतम शब्दावली आकार शामिल है। सरकारी वित्त पोषण ने भाषण मान्यता अनुसंधान को पुनर्जीवित किया जो जॉन पियर्स के पत्र के बाद संयुक्त राज्य अमेरिका में बड़े पैमाने पर छोड़ दिया गया था।

इस तथ्य के बावजूद कि सीएमयू की हार्पी प्रणाली कार्यक्रम के मूल लक्ष्यों को पूरा करती है, कई भविष्यवाणियां प्रचार से ज्यादा कुछ नहीं हैं, जो कि DARPA प्रशासकों को निराश करती हैं। इस निराशा के कारण DARPA ने धन जारी नहीं रखा। इस दौरान कई नवाचार हुए, जैसे कि सीएमयू के हार्पी सिस्टम में उपयोग के लिए बीम की खोज । इस क्षेत्र को अन्य क्षेत्रों में कई एल्गोरिदम की खोज से भी लाभ हुआ जैसे कि रैखिक भविष्य कहनेवाला कोडिंग और सेफस्ट्राल विश्लेषण।

1960 के दशक के उत्तरार्ध के दौरान लियोनार्ड बॉम ने इंस्टीट्यूट फॉर डिफेंस एनालिसिस में मार्कोव श्रृंखला का गणित विकसित किया । सीएमयू में, राज रेड्डी के छात्रों जेम्स बेकर और जेनेट एम। बेकर ने भाषण पहचान के लिए हिडन मार्कोव मॉडल (एचएमएम) का उपयोग शुरू किया। जेम्स बेकर ने अपनी स्नातक शिक्षा के दौरान इंस्टीट्यूट ऑफ डिफेंस एनालिसिस में एक ग्रीष्मकालीन नौकरी से एचएमएम के बारे में सीखा था । HMM के उपयोग ने शोधकर्ताओं को ज्ञान के विभिन्न स्रोतों, जैसे कि ध्वनिकी, भाषा और वाक्यविन्यास को एक एकीकृत संभाव्य मॉडल में संयोजित करने की अनुमति दी।

फ्रेड जेनेल्क की अगुवाई में, आईबीएम ने टंगोरा नामक एक आवाज सक्रिय टाइपराइटर बनाया , जो 1980 के दशक के मध्य तक 20,000 शब्द की शब्दावली को संभाल सकता था । Jelinek के सांख्यिकीय दृष्टिकोण ने मानव मस्तिष्क की प्रक्रियाओं के तरीके का अनुकरण करने पर कम जोर दिया और एचएमएम जैसी सांख्यिकीय मॉडलिंग तकनीकों का उपयोग करने के पक्ष में भाषण को समझा । (जेलाइन के समूह ने स्वतंत्र रूप से भाषण देने के लिए HMMs के एप्लिकेशन की खोज की।) यह भाषाविदों के साथ विवादास्पद था क्योंकि HMMs मानव भाषाओं की कई सामान्य विशेषताओं के लिए बहुत सरल हैं। हालांकि, HMM मॉडलिंग भाषण के लिए एक बहुत

ही उपयोगी तरीका साबित हुआ और 1980 के दशक में प्रमुख भाषण मान्यता एल्गोरिदम बनने के लिए गतिशील समय की जगह ले ली। आईबीएम के पास 1982 में जेम्स एंड जेनेट एम। बेकर द्वारा स्थापित ड्रैगन सिस्टम्स सहित कुछ प्रतिस्पर्धी थे। 1980 के दशक में एन-ग्राम भाषा मॉडल की शुरुआत भी हुई। कैट्ज ने 1987 में बैक-ऑफ मॉडल पेश किया, जिसने भाषा मॉडल को कई लंबाई के एन-ग्राम का उपयोग करने की अनुमति दी। उसी समय, CSELT भी एचएमएम (1980 से डिपाइपोनियों का अध्ययन किया गया था) का उपयोग इतालवी जैसी भाषा को पहचानने के लिए कर रहा था। उसी समय, CSELT ने यूरोपीय परियोजनाओं (एस्प्रिट I, II) की एक श्रृंखला का नेतृत्व किया, और एक पुस्तक में अत्याधुनिक कला को संक्षेप में प्रस्तुत किया, बाद में (2013) पुनर्मुद्रित किया।

क्षेत्र में अधिकांश प्रगति कंप्यूटर की तेजी से बढ़ती क्षमताओं के कारण होती है। 1976 में DARPA कार्यक्रम के अंत में, शोधकर्ताओं के पास सबसे अच्छा कंप्यूटर 4 एमबी रैम वाला पीडीपी -10 था। इन कंप्यूटरों के उपयोग से केवल 30 सेकंड के भाषण को डिकोड करने में 100 मिनट तक का समय लग सकता है। कुछ दशकों बाद, शोधकर्ताओं ने दसियों बार हजारों की संख्या में कंप्यूटिंग शक्ति तक पहुंच हासिल की। जैसे-जैसे तकनीक उन्नत और कंप्यूटर तेज होते गए, शोधकर्ताओं ने कठिन समस्याओं जैसे कि बड़ी वोकैबुलरीज, स्पीकर इंडिपेंडेंस, नॉइज़ वातावरण और वार्तालाप भाषण से निपटना शुरू कर दिया। विशेष रूप से, अधिक कठिन कार्यों के लिए इस स्थानांतरण ने 1980 के दशक से भाषण मान्यता के DARPA फंडिंग की विशेषता है। उदाहरण के लिए, स्पीकर की स्वतंत्रता पर प्रगति पहले वक्ताओं की एक विशाल विविधता पर प्रशिक्षण और फिर बाद में डिकोडिंग के दौरान स्पष्ट स्पीकर अनुकूलन करके की गई थी। शब्द त्रुटि दर में और कमी आई क्योंकि शोधकर्ताओं ने अधिकतम संभावना मॉडल का उपयोग करने के बजाय भेदभावपूर्ण होने के लिए ध्वनिक मॉडल को स्थानांतरित कर दिया।

मध्य-अस्सी के दशक में भाषण पहचान के बारे में कुछ माइक्रोप्रोसेसरों को भी जारी किया गया था: उदाहरण के लिए, 1986 में इसे नीदरलैंड में RIPAC प्रस्तुत किया गया था, एक स्वतंत्र-स्पीकर मान्यता (निरंतर भाषण के लिए) टेलीफोन सेवाओं के लिए चिप।

3. व्यावहारिक भाषण मान्यता:

1990 के दशक ने व्यावसायिक रूप से सफल भाषण मान्यता प्रौद्योगिकियों का पहला परिचय देखा। शुरुआती उत्पादों में से दो ड्रैगन डिक्टेट थे, एक उपभोक्ता उत्पाद जो 1990 में जारी किया गया था और मूल रूप से इसकी कीमत $ 9,000 थी, और कुर्ज़वील एप्लाइड इंटेलिजेंस के एक पहचानकर्ता ने 1987 में जारी किया था। एटीएंडटी ने 1992 में वॉयस रिकॉग्निशन कॉल प्रोसेसिंग सेवा को टेलीफोन कॉल के उपयोग के बिना रूट कॉल के लिए तैनात किया था। एक मानव ऑपरेटर। बेल लैब्स में लॉरेंस राबिनर और अन्य द्वारा तकनीक विकसित की गई थी । इस बिंदु तक, विशिष्ट वाणिज्यिक भाषण मान्यता प्रणाली की शब्दावली औसत मानव शब्दावली से बड़ी थी। राज रेड्डी के पूर्व छात्र, Xuedong हुआंग, ने CMU में स्फिंक्स- II प्रणाली विकसित की। Sphinx-II प्रणाली स्पीकर-इंडिपेंडेंट, बड़ी शब्दावली, निरंतर स्पीच रिकग्निशन करने वाली पहली थी और इसने DARPA के 1992 के मूल्यांकन में सर्वश्रेष्ठ प्रदर्शन किया था। एक बड़ी शब्दावली के साथ निरंतर भाषण को संभालना भाषण मान्यता के इतिहास में एक प्रमुख मील का पत्थर था। हुआंग 1993 में माइक्रोसॉफ्ट में भाषण मान्यता समूह को मिला। राज रेड्डी के छात्र काई-फू ली ने Apple में शामिल हो गए, जहां 1992 में, उन्होंने कैस्पर के रूप में जाना जाने वाले Apple कंप्यूटर के लिए एक भाषण इंटरफ़ेस प्रोटोटाइप विकसित करने में मदद की।

बेल्जियम की एक स्पीच रिकॉग्निशन कंपनी लर्नआउट एंड हॉसपाइ ने 1997 में कुर्ज़वील एप्लाइड इंटेलिजेंस और 2000 में ड्रैगन सिस्टम्स सहित कई अन्य कंपनियों का अधिग्रहण किया। विंडोज एक्सपी ऑपरेटिंग सिस्टम में एल एंड एच भाषण प्रौद्योगिकी का उपयोग किया गया था। L & H एक उद्योग के नेता थे जब तक कि 2001 में एक लेखा घोटाले ने कंपनी को समाप्त नहीं किया। एल एंड एच से भाषण प्रौद्योगिकी को स्कैनसॉफ्ट द्वारा खरीदा गया था जो 2005 में Nuance बन गया। Apple मूल रूप से अपने डिजिटल टीवी सिरी को वाक् पहचान क्षमता प्रदान करने के लिए Nuance से लाइसेंस प्राप्त सॉफ्टवेयर ।

2000 के दशक में DARPA ने दो भाषण मान्यता कार्यक्रमों को प्रायोजित किया: 2002 में प्रभावी सस्ती पुन: प्रयोज्य भाषण-से-पाठ (EARS) और वैश्विक स्वायत्त भाषा शोषण (GALE)। : चार टीमों कान कार्यक्रम में भाग लिया आईबीएम , एक टीम के नेतृत्व में बीबीएन के साथ LIMSI और यूनी। पिट्सबर्ग, कैम्ब्रिज विश्वविद्यालय, और ISCI, SRI और वाशिंगटन विश्वविद्यालय से बना एक दल । ईएआरएस ने 500 से अधिक वक्ताओं से 260 घंटे की रिकॉर्ड की गई बातचीत वाले स्विचबोर्ड टेलीफोन भाषण कॉर्पस के संग्रह को वित्त पोषित किया। गेल कार्यक्रम अरबी और मंदारिन

पर प्रसारित समाचार भाषण पर केंद्रित था। भाषण मान्यता में Google का पहला प्रयास 2007 में Nuance के कुछ शोधकर्ताओं को काम पर रखने के बाद आया। पहला उत्पाद GOOG-411 था, जो टेलीफोन आधारित निर्देशिका सेवा थी। GOOG-411 की रिकॉर्डिंग से मूल्यवान डेटा का उत्पादन हुआ जिसने Google को उनके मान्यता प्रणालियों को बेहतर बनाने में मदद की। Google वॉइस खोज अब 30 से अधिक भाषाओं में समर्थित है।

संयुक्त राज्य अमेरिका में, नेशनल सिक्योरिटी एजेंसी ने कम से कम 2006 से कीवर्ड स्पॉटिंग के लिए एक प्रकार की भाषण मान्यता का उपयोग किया है। यह तकनीक विश्लेषकों को रिकॉर्ड किए गए वार्तालापों की बड़ी मात्रा में खोज करने और कीवर्ड के अलग-अलग उल्लेख करने की अनुमति देती है। रिकॉर्डिंग को अनुक्रमित किया जा सकता है और विश्लेषकों को रुचि के वार्तालाप खोजने के लिए डेटाबेस पर क्वेरीज़ चला सकते हैं। कुछ सरकारी शोध कार्यक्रम वाक् पहचान के खुफिया अनुप्रयोगों, जैसे कि DARPA के EARS के कार्यक्रम और IARPA के बैबल कार्यक्रम पर केंद्रित हैं।

4. वाक् पहचान:

वाक् पहचान (या कभी-कभी स्वचालित भाषण मान्यता के रूप में संदर्भित) एक ऐसी प्रक्रिया है जिसके द्वारा कंप्यूटर (या अन्य प्रकार की मशीन) बोले गए शब्दों की पहचान करता है। मूल रूप से, इसका मतलब है कि कंप्यूटर से बात करना और इसे सही ढंग से समझना कि आप क्या कह रहे हैं। "समझने" से हमारा तात्पर्य है, एप्लिकेशन को उचित रूप से प्रतिक्रिया करने के लिए या इनपुट भाषण को बातचीत के दूसरे माध्यम में परिवर्तित करने के लिए, जो किसी अन्य एप्लिकेशन द्वारा आगे समझ में आता है जो इसे ठीक से संसाधित कर सकता है और उपयोगकर्ता को आवश्यक परिणाम प्रदान कर सकता है। जिन दिनों आपको रखना था कंप्यूटर स्क्रीन को घूरते हुए और चाबी को जोर से दबाएं या कंप्यूटर पर माउस से क्लिक करें ताकि आपकी आज्ञाओं का जवाब जल्द ही अतीत की चीजें हो सकें। Today we बाहर खिंचाव और आराम कर सकते हैं और अपने कंप्यूटर को अपनी बोली लगाने के लिए कह सकते हैं। यह एएसआर (स्वचालित भाषण मान्यता) प्रौद्योगिकी द्वारा संभव किया गया है। भाषण मान्यता कंप्यूटर के साथ बातचीत करने के पारंपरिक तरीकों का एक विकल्प है , जैसे कि कीबोर्ड के माध्यम से पाठीय इनपुट। एक प्रभावी प्रणाली विश्वसनीयता , मानक कीबोर्ड और माउस इनपुट को प्रतिस्थापित या कम कर सकती है । यह विशेष रूप से निम्नलिखित की सहायता कर सकता है:
जो लोग थोड़ा कुंजीपटल कौशल या अनुभव है, जो धीमी गति से टाइपिस्ट हैं, या नहीं है या संसाधनों कुंजीपटल कौशल विकसित करने के।

- डायस्लेक्सिक लोग या अन्य जिन्हें वर्ण या शब्द हेरफेर या पाठ के रूप में उपयोग के साथ समस्या है।
- शारीरिक विकलांगता वाले लोग जो अपने डेटा प्रविष्टि या पढ़ने की क्षमता को प्रभावित करते हैं जो उन्होंने दर्ज किए हैं।

5. आधुनिक प्रणाली:

2000 के दशक की शुरुआत में, स्पीच रिकॉग्निशन में अभी भी पारंपरिक दृष्टिकोणों का वर्चस्व था, जैसे कि हिडन मार्कोव मॉडल फीडफॉर्वर्ड आर्टिफिशियल न्यूरल नेटवर्क के साथ संयुक्त । आज, तथापि, भाषण मान्यता के कई पहलुओं पर एक से ले लिया गया है गहरी सीखने विधि कहा जाता लांग अल्पकालिक स्मृति (LSTM), एक आवर्तक तंत्रिका द्वारा प्रकाशित नेटवर्क सेप Hochreiter_और जुरगेन Schmidhuber_1997 LSTM RNNs गायब हो जाने ढाल समस्या से बचने और में "वेरी डीप लर्निंग" कार्यों को सीख सकते हैं जिनके लिए उन घटनाओं की यादों की आवश्यकता होती है जो हजारों समय पहले असतत समय में हुई थीं, जो भाषण के लिए महत्वपूर्ण है। 2007 के आसपास, कनेक्शनिस्ट टेम्पोरल क्लासिफिकेशन (CTC) द्वारा प्रशिक्षित LSTM ने कुछ अनुप्रयोगों में पारंपरिक भाषण पहचान को बेहतर बनाना शुरू किया। 2015 में, Google की वाक् पहचान ने कथित रूप से CTC- प्रशिक्षित LSTM के माध्यम से 49% की नाटकीय प्रदर्शन छलांग का अनुभव किया, जो अब Google Voice के माध्यम से सभी स्मार्टफोन उपयोगकर्ताओं के लिए उपलब्ध है। ध्वनिक मॉडलिंग के लिए डीप फीड फॉरवर्ड (गैर-आवर्तक) नेटवर्क का उपयोग 2009 के बाद के भाग में जेफ्री हिंटन और उनके छात्रों द्वारा टोरंटो विश्वविद्यालय में और ली डेंग और माइक्रोसॉफ्ट रिसर्च में सहयोगियों, माइक्रोसॉफ्ट के बीच सहयोगात्मक कार्य में शुरू किया गया था। टोरंटो विश्वविद्यालय जिसका बाद में आईबीएम और गूगल को शामिल करने के लिए विस्तार किया गया था (इसलिए "चार शोध समूहों के साझा विचार" उनकी 2012 की

समीक्षा में उपशीर्षक)। एक Microsoft अनुसंधान कार्यकारी ने इस नवाचार को "1979 के बाद से सटीकता में सबसे नाटकीय बदलाव" कहा। पिछले कुछ दशकों के स्थिर वृद्धिशील सुधारों के विपरीत, गहरी शिक्षा के आवेदन में शब्द त्रुटि दर में 30% की कमी आई है। इस नवाचार को जल्दी से पूरे क्षेत्र में अपनाया गया। शोधकर्ताओं ने भाषा मॉडलिंग के लिए भी गहरी सीखने की तकनीक का उपयोग करना शुरू कर दिया है । भाषण मान्यता के लंबे इतिहास में, कृत्रिम तंत्रिका नेटवर्क के उथले रूप और गहरे रूप (जैसे आवर्तक जाल) दोनों को 1980, 1990 और 2000 के कुछ वर्षों के दौरान कई वर्षों तक खोजा गया था। लेकिन इन विधियों ने कभी भी गैर-समान आंतरिक-हस्तनिर्मित गाऊसी मिश्रण मॉडल / हिडन मार्कोव मॉडल (जीएमएम-एचएमएम) तकनीक पर भाषण के सामान्य मॉडल पर आधारित प्रौद्योगिकी को नहीं जीता। 1990 के दशक में कई प्रमुख कठिनाइयों का विश्लेषण किया गया था, जिसमें तंत्रिका संबंधी पूर्वानुमान मॉडल में ढाल कम और कमजोर अस्थायी सहसंबंध संरचना शामिल है। ये सभी कठिनाइयां इन शुरुआती दिनों में बड़े प्रशिक्षण डेटा और बड़ी कंप्यूटिंग शक्ति की कमी के अलावा थीं । अधिकांश भाषण मान्यता शोधकर्ता, जिन्होंने इस तरह की बाधाओं को समझ लिया था, इसलिए बाद में तंत्रिका जाल से दूर चले गए ताकि उदार मॉडलिंग दृष्टिकोण का पीछा किया जा सके, जब तक कि 2009–2010 के आसपास गहन शिक्षण के पुनरुत्थान ने इन सभी कठिनाइयों को पार नहीं कर लिया। हिंटन एट अल। और देंग एट अल। इस हालिया इतिहास का हिस्सा है कि कैसे चार समूहों (टोरंटो विश्वविद्यालय, माइक्रोसॉफ्ट, Google और आईबीएम) के सहयोगियों के साथ एक-दूसरे के साथ उनका सहयोग और भाषण मान्यता के लिए गहरी फीडफोर्वर्ड तंत्रिका नेटवर्क के अनुप्रयोगों के पुनर्जागरण को प्रज्वलित किया ।

6. मॉडल, तरीके और एल्गोरिदम:

दोनों ध्वनिक मॉडलिंग और भाषा मॉडलिंग आधुनिक सांख्यिकीय-आधारित भाषण मान्यता एल्गोरिदम के महत्वपूर्ण हिस्से हैं। कई प्रणालियों में छिपे हुए मार्कोव मॉडल (HMM) का व्यापक रूप से उपयोग किया जाता है। भाषा मॉडलिंग का उपयोग कई अन्य प्राकृतिक भाषा प्रसंस्करण अनुप्रयोगों जैसे दस्तावेज़ वर्गीकरण या सांख्यिकीय मशीन अनुवाद में भी किया जाता है।

a) छिपे हुए मार्कोव मॉडल

आधुनिक सामान्य-उद्देश्य भाषण मान्यता प्रणाली छिपे हुए मार्कोव मॉडल पर आधारित हैं। ये सांख्यिकीय मॉडल हैं जो प्रतीकों या मात्राओं के अनुक्रम का उत्पादन करते हैं। एचएमएम का उपयोग भाषण मान्यता में किया जाता है क्योंकि एक भाषण सिग्नल को एक टुकड़े के स्थिर स्थिर सिग्नल या कम समय के स्थिर सिग्नल के रूप में देखा जा सकता है। थोड़े समय के पैमाने (जैसे, 10 मिलीसेकंड) में, भाषण को एक स्थिर प्रक्रिया के रूप में अनुमानित किया जा सकता है। भाषण को कई स्टोकेस्टिक उद्देश्यों के लिए मार्कोव मॉडल के रूप में सोचा जा सकता है। एक और कारण है कि एचएमएम लोकप्रिय हैं क्योंकि वे स्वचालित रूप से प्रशिक्षित हो सकते हैं और उपयोग करने के लिए सरल और कम्प्यूटेशनल रूप से संभव हैं। भाषण मान्यता में, छिपी हुई मार्कोव मॉडल एन -डायमेंशनल रियल-वैल्यू वैक्टर (एन के साथ एक छोटा पूर्णांक, जैसे कि 10) के एक अनुक्रम का उत्पादन करेगी , इन में से प्रत्येक 10 मिलीसेकंड का उत्पादन। वैक्टर में cepstral coefficients शामिल होते हैं, जो भाषण के थोड़े समय के फूरियर ट्रांसफॉर्म को ले कर प्राप्त होते हैं और एक cosine ट्रांसफॉर्मेशन का उपयोग करके स्पेक्ट्रम को सजाते हैं , फिर पहला (सबसे महत्वपूर्ण) गुणांक लेते हैं। छिपे हुए मार्कोव मॉडल में प्रत्येक राज्य में एक सांख्यिकीय वितरण होता है जो विकर्ण कोवरियन गॉसियंस का मिश्रण होता है, जो प्रत्येक मनाया वेक्टर के लिए संभावना देगा । प्रत्येक शब्द, या (अधिक सामान्य भाषण मान्यता प्रणालियों के लिए), प्रत्येक ध्वनि, एक अलग आउटपुट वितरण होगा; शब्दों या ध्वनियों के अनुक्रम के लिए एक छिपे हुए मार्कोव मॉडल को अलग-अलग शब्दों और स्वरों के लिए अलग-अलग प्रशिक्षित छिपे हुए मार्कोव मॉडल को मिलाकर बनाया गया है। ऊपर वर्णित भाषण पहचान के लिए सबसे आम, HMM- आधारित दृष्टिकोण के मूल तत्व हैं। आधुनिक भाषण मान्यता प्रणाली उपरोक्त वर्णित मूल दृष्टिकोण पर परिणामों को बेहतर बनाने के लिए कई मानक तकनीकों के विभिन्न संयोजनों का उपयोग करती हैं। एक विशिष्ट बड़े-शब्दावली प्रणाली को स्वरों के लिए संदर्भ निर्भरता की आवश्यकता होती है (इसलिए अलग-अलग बाएँ और दाएँ प्रसंग वाले स्वरों को HMM राज्यों के रूप में अलग-अलग अहसास होते हैं); यह विभिन्न स्पीकर और रिकॉर्डिंग स्थितियों को सामान्य करने के लिए cepstral सामान्यीकरण का उपयोग करेगा ; आगे के स्पीकर सामान्यीकरण के लिए यह अधिक सामान्य स्पीकर अनुकूलन के लिए पुरुष-महिला सामान्यीकरण और अधिकतम संभावना रैखिक प्रतिगमन (MLLR) के लिए मुखर ट्रैक्ट लंबाई सामान्यीकरण (VTLN) का उपयोग कर सकता है। सुविधाओं पर कब्जा भाषण गतिशीलता के लिए और इसके अलावा का उपयोग कर सकते

में डेल्टा और डेल्टा-डेल्टा गुणांक तथाकथित होता heteroscedastic रैखिक विभेदक विश्लेषण (HLDA); या डेल्टा और डेल्टा-डेल्टा गुणांक को छोड़ सकता है और splicing और एक LDA- आधारित प्रक्षेपण का उपयोग कर सकता है जिसके बाद शायद विषमलैंगिक रेखीय विभेदक विश्लेषण या एक वैश्विक अर्ध-बंधे सह-विचरण परिवर्तन (जिसे अधिकतम पूर्णता रैखिक परिवर्तन या MLLT भी कहा जाता है)। कई सिस्टम तथाकथित भेदभावपूर्ण प्रशिक्षण तकनीकों का उपयोग करते हैं जो एचएमएम पैरामीटर अनुमान के लिए विशुद्ध रूप से सांख्यिकीय दृष्टिकोण के साथ फैलाव करते हैं और इसके बजाय प्रशिक्षण डेटा के कुछ वर्गीकरण-संबंधित माप का अनुकूलन करते हैं। उदाहरण अधिकतम पारस्परिक जानकारी (MMI), न्यूनतम वर्गीकरण त्रुटि (MCE) और न्यूनतम फ़ोन त्रुटि (MPE) हैं।

भाषण का डिकोडिंग (शब्द तब होता है जब सिस्टम को नए उच्चारण के साथ प्रस्तुत किया जाता है और सबसे संभावित स्रोत वाक्य की गणना करनी चाहिए) शायद सबसे अच्छा रास्ता खोजने के लिए Viterbi एल्गोरिथ्म का उपयोग करेगा , और यहां गतिशील रूप से बनाने के बीच एक विकल्प है छिपे हुए मार्कोव मॉडल का संयोजन, जिसमें ध्वनिक और भाषा मॉडल की जानकारी दोनों शामिल हैं, और इसे पहले से ही संयोजित करना है (परिमित राज्य ट्रांसड्यूसर, या एफएसटी, दृष्टिकोण)। डिकोडिंग के लिए एक संभावित सुधार केवल अच्छे उम्मीदवार रखने के बजाय अच्छे उम्मीदवारों का एक सेट रखने के लिए है, और इन अच्छे उम्मीदवारों को रेट करने के लिए एक बेहतर स्कोरिंग फ़ंक्शन (री स्कोरिंग) का उपयोग करना है ताकि हम इस परिष्कृत स्कोर के अनुसार सर्वश्रेष्ठ चुन सकें। । उम्मीदवारों के सेट को या तो एक सूची (एन-बेस्ट सूची दृष्टिकोण) या मॉडलों के सबसेट (एक जाली) के रूप में रखा जा सकता है । री स्कोरिंग आमतौर पर बेयस रिस्क (या इसके अंदाज़े से पता चलता है) को कम करने की कोशिश करके किया जाता है : अधिकतम संभावित संभावना के साथ स्रोत वाक्य को लेने के बजाय, हम ऐसे वाक्यों को लेने की कोशिश करते हैं जो सभी दिए गए विवरणों के संबंध में किसी दिए गए नुकसान फ़ंक्शन की प्रत्याशा को कम करता है। (यानी, हम उस वाक्य को लेते हैं जो उनकी अनुमानित संभावना से भारित अन्य संभावित वाक्यों के लिए औसत दूरी को कम करता है)। नुकसान का कार्य आमतौर पर लेवेंसहाइट दूरी है, हालांकि यह विशिष्ट कार्यों के लिए अलग-अलग दूरी हो सकता है; निश्चित रूप से, ट्रैक्टेबिलिटी को बनाए रखने के लिए संभावित ट्रांसक्रिप्शन का सेट है। कुशल एल्गोरिदम को फिर से स्कोर किए गए अक्षांशों के लिए तैयार किया गया है, जो कि संपादित दूरी के साथ भारित परिमित राज्य ट्रांसड्यूसर के रूप में प्रतिनिधित्व करते हैं, कुछ मान्यताओं की पुष्टि करने वाले एक परिमित राज्य ट्रांसड्यूसर के रूप में खुद का प्रतिनिधित्व करते हैं।

b) गतिशील समय ताना

डायनेमिक टाइम ताना-बाना एक ऐसा दृष्टिकोण है जो ऐतिहासिक रूप से भाषण मान्यता के लिए उपयोग किया जाता था, लेकिन अब एचएमएम-आधारित दृष्टिकोण से अधिक विस्थापित हो गया है।

डायनेमिक टाइम वारपिंग दो अनुक्रमों के बीच समानता को मापने के लिए एक एल्गोरिथ्म है जो समय या गति में भिन्न हो सकता है। उदाहरण के लिए, चलने के पैटर्न में समानता का पता लगाया जाएगा, भले ही एक वीडियो में व्यक्ति धीरे-धीरे चल रहा हो और अगर दूसरे में वह अधिक तेज़ी से चल रहा हो, या भले ही एक अवलोकन के दौरान तेजी और मंदी हो। DTW को वीडियो, ऑडियो और ग्राफिक्स पर लागू किया गया है - वास्तव में, किसी भी डेटा को रैखिक प्रतिनिधित्व में बदल दिया जा सकता है, जिसका विश्लेषण DTW के साथ किया जा सकता है। अलग-अलग बोलने की गति का सामना करने के लिए एक प्रसिद्ध एप्लिकेशन स्वचालित भाषण मान्यता है। सामान्य तौर पर, यह एक ऐसा तरीका है जो कंप्यूटर को कुछ प्रतिबंधों के साथ दो दिए गए अनुक्रमों (जैसे, समय श्रृंखला) के बीच एक इष्टतम मैच खोजने की अनुमति देता है। यही है, अनुक्रम एक दूसरे से मेल खाने के लिए गैर-रैखिक रूप से "विकृत" हैं। यह अनुक्रम संरेखण विधि अक्सर छिपे हुए मार्कोव मॉडल के संदर्भ में उपयोग की जाती है।

c) तंत्रिका नेटवर्क

कृत्रिम तंत्रिका नेटवर्क

तंत्रिका नेटवर्क 1980 के दशक के अंत में ASR में एक आकर्षक ध्वनिक मॉडलिंग दृष्टिकोण के रूप में उभरा । तब से, तंत्रिका नेटवर्क का उपयोग भाषण मान्यता के कई पहलुओं में किया गया है जैसे कि फ़ोनेमी वर्गीकरण, पृथक शब्द मान्यता, दृश्य-श्रव्य मान्यता, दृश्य-श्रव्य स्पीकर मान्यता और स्पीकर अनुकूलन। एचएमएम के विपरीत, तंत्रिका नेटवर्क फीचर सांख्यिकीय गुणों के बारे में कोई धारणा नहीं बनाते हैं और उनमें कई गुण हैं जो उन्हें भाषण मान्यता के लिए आकर्षक पहचान मॉडल बनाते हैं। जब भाषण सुविधा खंड की संभावनाओं का अनुमान लगाने के लिए उपयोग किया जाता है, तो तंत्रिका नेटवर्क प्राकृतिक और कुशल तरीके से भेदभावपूर्ण प्रशिक्षण की अनुमति देते

हैं। तंत्रिका नेटवर्क के साथ इनपुट सुविधाओं के आंकड़ों पर कुछ धारणाएं बनाई गई हैं। हालांकि, व्यक्तिगत फोन और अलग-थलग शब्दों जैसे लघु-समय इकाइयों को वर्गीकृत करने में उनकी प्रभावशीलता के बावजूद, तंत्रिका नेटवर्क निरंतर मान्यता कार्यों के लिए शायद ही कभी सफल होते हैं, मोटे तौर पर क्योंकि उनकी अस्थायी निर्भरता की क्षमता में कमी है। हालाँकि, हाल ही में LSTM आवर्तक तंत्रिका नेटवर्क (RNN) और समय विलंब तंत्रिका नेटवर्क (TDNN) का उपयोग किया गया है, जो कि अव्यक्त लौकिक निर्भरताओं की पहचान करने और भाषण की मान्यता का कार्य करने के लिए इस जानकारी का उपयोग करने में सक्षम दिखाया गया है। दीप तंत्रिका नेटवर्क और denoising Auto encoders भी एक प्रभावी ढंग से इस समस्या से निपटने के लिए के साथ प्रयोग किया जा रहा था।

लौकिक निर्भरता मॉडल करने के लिए feed forward तंत्रिका नेटवर्क की अक्षमता के कारण , एक वैकल्पिक दृष्टिकोण HMM आधारित मान्यता के लिए तंत्रिका प्रसंस्करण का उपयोग पूर्व-प्रसंस्करण जैसे सुविधा परिवर्तन, आयामीता में कमी के लिए होता है।

d) डीप फीडफोरवर्ड और रिकरंट न्यूरल नेटवर्क्स
ध्यान लगा के पढ़ना या सीखना

एक गहरी फीडफ़र्वर्ड न्यूरल नेटवर्क (डीएनएन) एक कृत्रिम तंत्रिका नेटवर्क है जिसमें इनपुट और आउटपुट परतों के बीच कई छिपी हुई इकाइयों की परतें होती हैं। उथले तंत्रिका नेटवर्क के समान, DNNs जटिल गैर-रैखिक संबंध मॉडल कर सकते हैं। डीएनएन आर्किटेक्चर रचनात्मक मॉडल उत्पन्न करते हैं, जहां अतिरिक्त परतें निचले परतों से सुविधाओं की संरचना को सक्षम करती हैं, जिससे एक विशाल सीखने की क्षमता मिलती है और इस प्रकार भाषण डेटा के जटिल पैटर्न मॉडलिंग की क्षमता होती है । 2010 में औद्योगिक शोधकर्ताओं द्वारा अकादमिक शोधकर्ताओं के सहयोग से बड़े शब्दावली भाषण मान्यता में DNNs की सफलता प्राप्त हुई, जहां निर्णय वृक्षों द्वारा निर्मित संदर्भ आश्रित HMM राज्यों पर आधारित DNN की बड़ी उत्पादन परतें अपनाई गईं। माइक्रोसॉफ्ट रिसर्च की हालिया स्प्रिंगर बुक में अक्टूबर 2014 तक इस विकास और कला की स्थिति की व्यापक समीक्षा देखें। स्वचालित भाषण मान्यता की संबंधित पृष्ठभूमि और हाल के अवलोकन लेखों में विशेष रूप से गहन सीखने सहित विभिन्न मशीन सीखने के प्रतिमानों के प्रभाव को देखें। डीप लर्निंग का एक मूल सिद्धांत है हाथ से तैयार की जाने वाली फीचर इंजीनियरिंग के साथ दूर करना और कच्ची सुविधाओं का उपयोग करना। इस सिद्धांत पहले गहरी की वास्तुकला में सफलतापूर्वक पता लगाया गया था auto encoder "कच्चे" spectrogram या रैखिक फिल्टर बैंक सुविधाओं पर, Mel- पर अपनी श्रेष्ठता दिखा Cepstral विशेषताएं जो तय करने के कुछ चरणों में होते हैं spectrograms से परिवर्तन। भाषण की वास्तविक "कच्ची" विशेषताएं, तरंग, हाल ही में उत्कृष्ट बड़े पैमाने पर भाषण पहचान परिणामों का उत्पादन करने के लिए दिखाई गई हैं।

e) एंड-टू-एंड ऑटोमैटिक स्पीच रिकॉग्निशन

2014 के बाद से, एंड-टू-एंड एएसआर में बहुत अधिक शोध रुचि है। पारंपरिक ध्वन्यात्मक-आधारित (यानी, सभी HMM- आधारित मॉडल) उच्चारण, ध्वनिक और भाषा मॉडल के लिए अलग-अलग घटकों और प्रशिक्षण की आवश्यकता होती है। एंड-टू-एंड मॉडल संयुक्त रूप से भाषण पहचानकर्ता के सभी घटकों को सीखते हैं। यह मूल्यवान है क्योंकि यह प्रशिक्षण प्रक्रिया और परिनियोजन प्रक्रिया को सरल बनाता है। उदाहरण के लिए, सभी एचएमएम-आधारित प्रणालियों के लिए एक एन-ग्राम भाषा मॉडल की आवश्यकता होती है, और एक विशिष्ट एन-ग्राम भाषा मॉडल अक्सर मेमोरी में कई गीगाबाइट लेता है जिससे उन्हें मोबाइल उपकरणों पर तैनात करने के लिए अव्यवहारिक हो जाता है। नतीजतन, Google और Apple (2017 के अनुसार) से आधुनिक वाणिज्यिक ASR सिस्टम क्लाउड पर तैनात किए जाते हैं और स्थानीय रूप से डिवाइस के विपरीत नेटवर्क कनेक्शन की आवश्यकता होती है। एंड-टू-एंड एएसआर का पहला प्रयास 2014 में कनेक्शन के टेम्पोरल क्लासिफिकेशन (CTC) आधारित सिस्टम के साथ था, जो कि Google Deep Mind के एलेक्स ग्रेव्स और टोरंटो विश्वविद्यालय के नवदीप जेटली द्वारा शुरू किया गया था। इस मॉडल में आवर्तक तंत्रिका नेटवर्क और CTC परत शामिल थे। संयुक्त रूप से, आरएनएन-सीटीसी मॉडल उच्चारण और ध्वनिक मॉडल को एक साथ सीखता है, हालांकि यह एचएमएम के समान सशर्त स्वतंत्रता मान्यताओं के कारण भाषा सीखने में असमर्थ है। नतीजतन, सीटीसी मॉडल सीधे अंग्रेजी पात्रों के लिए भाषण ध्वनिकी को मैप करना सीख सकते हैं, लेकिन मॉडल कई सामान्य वर्तनी की गलतियाँ करते हैं और लिपियों को साफ करने के लिए एक अलग भाषा मॉडल पर भरोसा करना चाहिए। बाद में, Baidu ने बहुत बड़े डेटासेट के साथ काम का विस्तार किया और चीनी मंदारिन और अंग्रेजी में

कुछ व्यावसायिक सफलता का प्रदर्शन किया। 2016 में, ऑक्सफ़ोर्ड विश्वविद्यालय ने लिपनेट , एक आरएनएन -सीटीसी वास्तुकला के साथ युग्मित स्पैटियोटेम्पोरल संकल्पों का उपयोग करते हुए पहले अंत-टू-एंड वाक्य स्तर के लेप रीडिंग मॉडल को प्रस्तुत किया, जो एक सीमित व्याकरणिक डेटा में मानव-स्तरीय प्रदर्शन को पार करता है।

हाई स्पीड रिक्रिएशन कैसे काम करता है

एक भाषण मान्यता इंजन (या भाषण पहचानकर्ता) एक ऑडियो स्ट्रीम इनपुट के रूप में लेता है और इसे एक पाठ प्रतिलेखन में बदल देता है। वाक् पहचान प्रक्रिया को एक फ्रंट एंड और बैक एंड के रूप में माना जा सकता है।

कन्वर्ट ऑडियो Input

फ्रंट एंड ऑडियो स्ट्रीम को प्रोसेस करता है, ध्वनि के सेगमेंट को अलग करता है जो संभवतः भाषण हैं और उन्हें सांख्यिक मानों की एक श्रृंखला में परिवर्तित करते हैं जो सिग्नल में मुखर ध्वनियों की विशेषता रखते हैं।

स्पीच मॉडल्स के लिए मैच इनपुट

पिछला छोर एक विशेष खोज इंजन है जो सामने के छोर से उत्पादित आउटपुट लेता है और तीन डेटाबेस में खोज करता है: एक ध्वनिक मॉडल, एक लेक्सिकॉन और एक भाषा मॉडल।

- ध्वनिक मॉडल एक भाषा का ध्वनिक ध्वनियों का प्रतिनिधित्व करता है, और एक विशेष उपयोगकर्ता की भाषण पैटर्न और ध्वनिक वातावरण की विशेषताओं पहचानने के लिए प्रशिक्षित किया जा सकता है।
- शब्दकोश भाषा में शब्द की एक बड़ी संख्या को सूचीबद्ध करता है, और प्रत्येक शब्द उच्चारण कैसे के बारे में जानकारी प्रदान करता है।
- भाषा मॉडल तरीकों से एक भाषा के शब्द जोड़ दिया जाता प्रतिनिधित्व करता है।
- ध्वनि के किसी भी खंड के लिए, कई चीजें हैं जिन्हें स्पीकर संभवतः कह सकता है। एक पहचानकर्ता की गुणवत्ता यह निर्धारित करती है कि वह अपनी खोज को परिष्कृत करने, खराब मैचों को समाप्त करने और अधिक संभावित मैचों का चयन करने में कितना अच्छा है। यह इसकी भाषा और ध्वनिक मॉडल की गुणवत्ता और इसके एल्गोरिदम की प्रभावशीलता पर बड़ा हिस्सा निर्भर करता है, दोनों प्रसंस्करण ध्वनि के लिए और मॉडलों में खोज के लिए।
- व्याकरण
- हालांकि एक पहचानकर्ता की अंतर्निहित भाषा मॉडल एक व्यापक भाषा डोमेन (जैसे कि रोजमर्रा की बोली जाने वाली अंग्रेजी) का प्रतिनिधित्व करने का इरादा है, एक भाषण एप्लिकेशन को अक्सर केवल कुछ कथनों को संसाधित करने की आवश्यकता होती है जिनके पास उस एप्लिकेशन के लिए विशेष अर्थ अर्थ होते हैं। सामान्य प्रयोजन के भाषा मॉडल का उपयोग करने के बजाय , एक एप्लिकेशन को एक व्याकरण का उपयोग करना चाहिए जो पहचानकर्ता को केवल भाषण के लिए सुनने के लिए विवश करता है जो कि आवेदन के लिए सार्थक है। यह निम्नलिखित लाभ प्रदान करता है:
- मान्यता की सटीकता को बढ़ाता है ।
- गारंटी देता है कि सभी मान्यता परिणाम आवेदन के लिए सार्थक हैं ।
- मान्यता प्राप्त पाठ में निहित अर्थ मूल्यों को निर्दिष्ट करने के लिए मान्यता इंजन को सक्षम करता है ।

Microsoft भाषण प्लेटफ़ॉर्म SDK 11 प्रोग्राम व्याकरण को संलेखन के लिए प्रक्रिया प्रदान करता है, और उद्योग-मानक मार्कअप भाषा का उपयोग करते हुए व्याकरण का समर्थन भी करता है ।

7. SPEEC RECOGNITION के प्रकार:

भाषण मान्यता प्रणालियों को कई अलग-अलग वर्गों में यह वर्णन करके अलग किया जा सकता है कि उनके पास किस प्रकार के उच्चारण को पहचानने की क्षमता है। ये कक्षाएं इस तथ्य पर आधारित हैं कि एएसआर की कठिनाइयों में से एक यह निर्धारित करने की क्षमता है कि एक स्पीकर शुरू होता है और एक उच्चारण पूरा करता है। अधिकांश पैकेज एक से अधिक कक्षाओं में फिट हो सकते हैं, जो इस बात पर निर्भर करता है कि वे किस मोड का उपयोग कर रहे हैं।

पृथक शब्द

पृथक शब्द पहचानकर्ताओं को आमतौर पर नमूना विंडो के दोनों किनारों पर शांत (एक ऑडियो सिग्नल की कमी) होने की आवश्यकता होती है। इसका मतलब यह नहीं है कि यह एकल शब्दों को स्वीकार करता है, लेकिन एक बार में एक ही उच्चारण की आवश्यकता होती है। अक्सर, इन प्रणालियों में "सुनो / नॉट-सुनो" राज्य होते हैं, जहां उन्हें बोलने वालों (आमतौर पर ठहराव के दौरान प्रसंस्करण) के बीच प्रतीक्षा करने के लिए स्पीकर की आवश्यकता होती है। पृथक वर्ग इस वर्ग के लिए एक बेहतर नाम हो सकता है।

जुड़े हुए शब्द

कनेक्ट शब्द सिस्टम (या अधिक सही ढंग से 'कनेक्टेड उच्चारण') अलग-अलग शब्दों के समान हैं, लेकिन अलग-अलग उच्चारणों को उनके बीच न्यूनतम ठहराव के साथ 'रन-एक' होने देते हैं।

सतत भाषण

सतत मान्यता अगला कदम है। निरंतर भाषण क्षमताओं वाले पहचानकर्ता बनाने में सबसे कठिन हैं क्योंकि उन्हें उच्चारण सीमाओं को निर्धारित करने के लिए विशेष तरीकों का उपयोग करना चाहिए। निरंतर भाषण पहचानकर्ता उपयोगकर्ताओं को लगभग स्वाभाविक रूप से बोलने की अनुमति देते हैं, जबकि कंप्यूटर सामग्री को निर्धारित करता है। असल में, यह कंप्यूटर डिक्टेशन है।

सहज भाषण

वास्तव में जो सहज भाषण है, उसके लिए कई तरह की परिभाषाएँ प्रतीत होती हैं। एक बुनियादी स्तर पर, यह भाषण के रूप में सोचा जा सकता है जो कि प्राकृतिक लग रहा है और पूर्वाभ्यास नहीं किया गया है। सहज भाषण क्षमता के साथ एक एएसआर प्रणाली विभिन्न प्राकृतिक भाषण सुविधाओं को संभालने में सक्षम होना चाहिए जैसे कि शब्द एक साथ चलाए जा रहे हैं, "ओम" और "आह", और यहां तक कि मामूली स्टेटर।

आवाज सत्यापन / पहचान

कुछ एएसआर सिस्टम विशिष्ट उपयोगकर्ताओं की पहचान करने की क्षमता रखते हैं। यह दस्तावेज़ सत्यापन या सुरक्षा प्रणालियों को कवर नहीं करता है।

बोली जाने वाली भाषा की समझदारी सिग्नल सिग्नल प्रोसेसिंग, ट्रांसक्रिप्शन और उच्च स्तरीय लॉजिक सिस्टम की मांग करती है ताकि हम उस प्रदर्शन को प्राप्त कर सकें, जो हम बातचीत के लिए बोली जाने वाली भाषा बातचीत की कल्पना करते हैं।शब्द सदिश स्थानों में शब्द उच्चारण की योजना बनाने के लिए , यह गहरी वास्तुकला सीखने में आगे के काम के लिए कई अवसरों की आपूर्ति करता है। हम आशा करते हैं कि यह भाषण और भाषा समझ के अत्यंत सक्रिय अनुसंधान क्षेत्र में अब संचार और प्रजनन नेटवर्क में सुधार के लिए संदर्भ बिंदु के रूप में काम करेगा ।

संदर्भ-सूची:

1. Ossama Abdel-Hamid, Abdel rahman Mohamed, Hui Jang, and Gerald Penn. Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In ICASSP, 2012.
2. C. O. Alm, D. Roth, and R. Sproat. Emotions from text: machine learning for text-based emotion prediction. In Proceedings of HLT/EMNLP, pages 579{586, 2005.
3. Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. Emotions from text: Machine learning for text-based emotion prediction. In Proceedings of the Hu-man Language Technology Conference and the Conference on Empirical Meth-ods in Natural Language Processing (HLT/EMNLP), 2005.
4. Andreevskaia and S. Bergler. Mining WordNet for fuzzy sentiment: sentiment tag extraction from WordNet glosses. In Proceedings of the European ACL, pages 209{216, 2006
5. L. B. Bahl, P. de Souza, and R. P. Mercer. Maximum mutual information estimation of hidden markov model parameters for speech recognition. In ICASSP. IEEE, 1986.
6. Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. a neural probabilistic language model. Journal of Machine Learning Research, 3:1137{1155, August 2003.

7. Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is di cult. IEEE Transactions on Neural Networks, 5(2), 1994.
8. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 3:993{1022, May 2003.
9. H. Bourlard and N. Morgan. Connectionist Speech Recognition: A Hybrid Ap-proach. Kluwer Academic Publishers, Norwell, MA, 1993.
10. J. Boyd-Graber and P. Resnik. Holistic sentiment analysis across languages: multilingual supervised latent Dirichlet allocation. In Proceedings of EMNLP, pages 45{55, 2010.
11. Rebecca F. Bruce and Janyce M. Wiebe. Recognizing subjectivity: A case study in manual tagging. Natural Language Engineering, 5(2), 1999.
12. Lu s Cabral and Ali Hortacsu. The dynamics of seller reputation: Theory and evidence from eBay. Working paper, downloaded version revised in March, 2006.