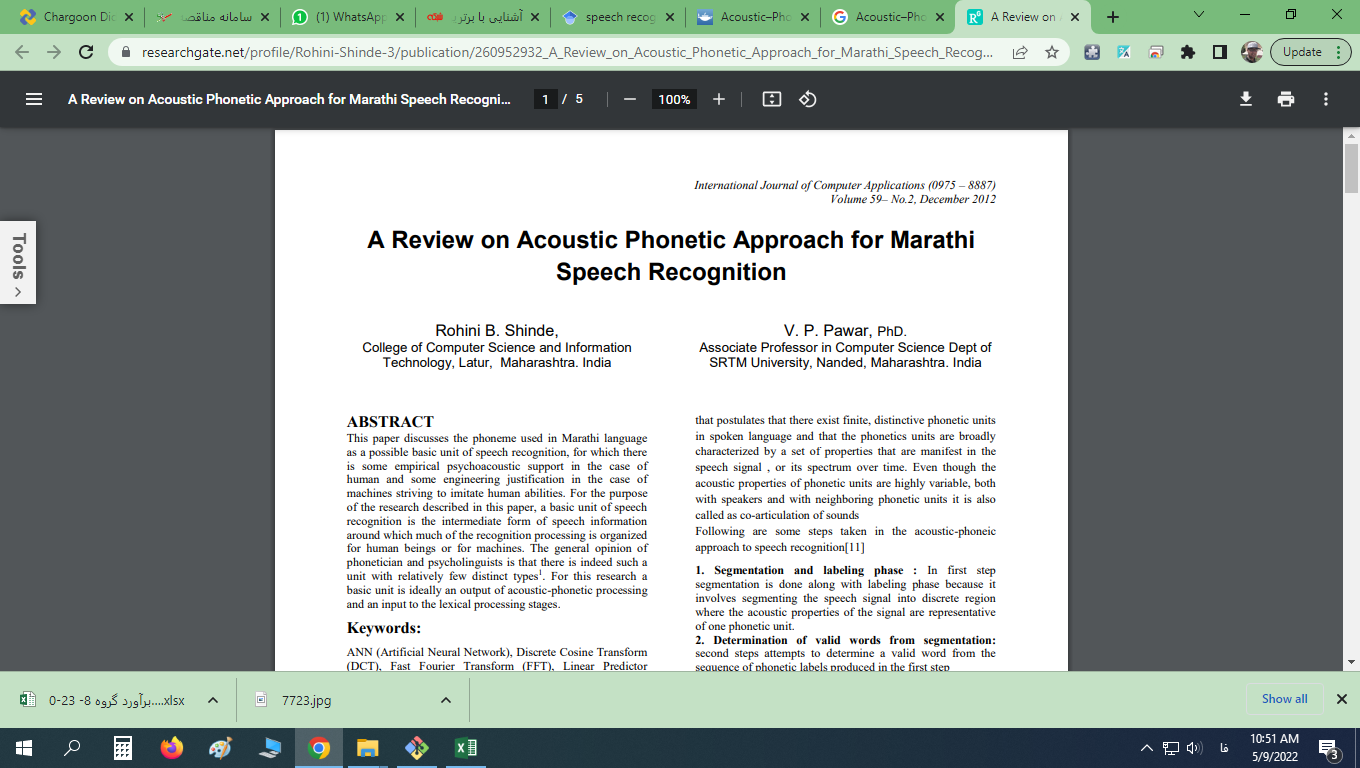
مقدمه:

ساده‌ترین مکانیزمی که انسان‌ها برای برقراری ارتباط با یک‌دیگر از آن استفاده می‌کنند صحبت کردن است. در دنیای فناوری اطلاعات، تشخیص گفتار (Speech Recognition) به توانایی سامانه‌ها در درک محاوره‌های انسانی، پردازش، تفسیر و تبدیل گفتار به متن اشاره دارد. تشخیص گفتار زیرشاخه‌ای از زبان‌شناسی محاسباتی (Computational Linguistic) است. این زیرشاخه در ارتباط با فناوری‌هایی است که داده‌های صوتی را به‌عنوان ورودی دریافت و تجزیه‌وتحلیل می‌کنند. تشخیص گفتار از فناوری‌های مهمی است که شرکت‌های بزرگی در حال کار روی آن هستند. هدف این است که تعامل با کامپیوترها از حالت دیجیتالی و غیر‌منعطف خارج شده و سامانه‌ها بتوانند مفهوم جملات ما را درک کرده و همانند یک انسان به ما پاسخ دهند. امروزه ابزارهای تشخیص احساسات (Scenes Recognition) در کارهای مختلفی مثل نوشتن پیام متنی، پخش موسیقی، دستیارهای مجازی و غیره استفاده می‌شوند.

الگوریتم های فناوری تشخیص گفتار خودکار (ASR) سرنام Automatic Speech Recognition :

روش آکوستیک-آوایی (Acoustic-Phonetic Approach)

روش آکوستیک-آوایی بر پایه آواشناسی صوتی(Acoustic Phonetics) است و بیان می‌کند که زبان گفتاری دارای واحدهای آوایی متناهی و متمایز است و به همین دلیل خواص صوتی (Acoustic Properties) واحدهای آوایی در سیگنال گفتاری یا طیف آن در طول زمان آشکار می‌شود. روش آکوستیک‌-آوایی با تجزیه‌و‌تحلیل طیفی گفتار آغاز می‌شود و در ادامه روی تشخیص و شناسایی آواها متمرکز می‌شود تا خصوصیات طیفی را به ویژگی‌های آوایی منحصربه‌فرد تبدیل کند. بعد از تکمیل این مرحله، نوبت به تفکیک و برچسب‌گذاری می‌رسد. سیگنال گفتار به نواحی آکوستیک پایدار تقسیم می‌شود و به هر ناحیه تقسیم‌شده یک یا چند برچسب آوایی اختصاص داده می‌شود و به این شکل خصوصیات مجموعه‌ای از آواهای مربوط به گفتار مشخص می‌شود. بعد از ایجاد یک توالی از آواهای تقسیم‌بندی و برچسب‌گذاری‌شده، در آخرین فرآیند، ساخت کلمات یا عبارات معنادار انجام می‌شود.



A Review on Acoustic Phonetic Approach for Marathi Speech Recognition Rohini B. Shinde, College of Computer Science and Information Technology, Latur, Maharashtra. India V. P. Pawar, PhD. Associate Professor in Computer Science Dept of SRTM University, Nanded, Maharashtra. India

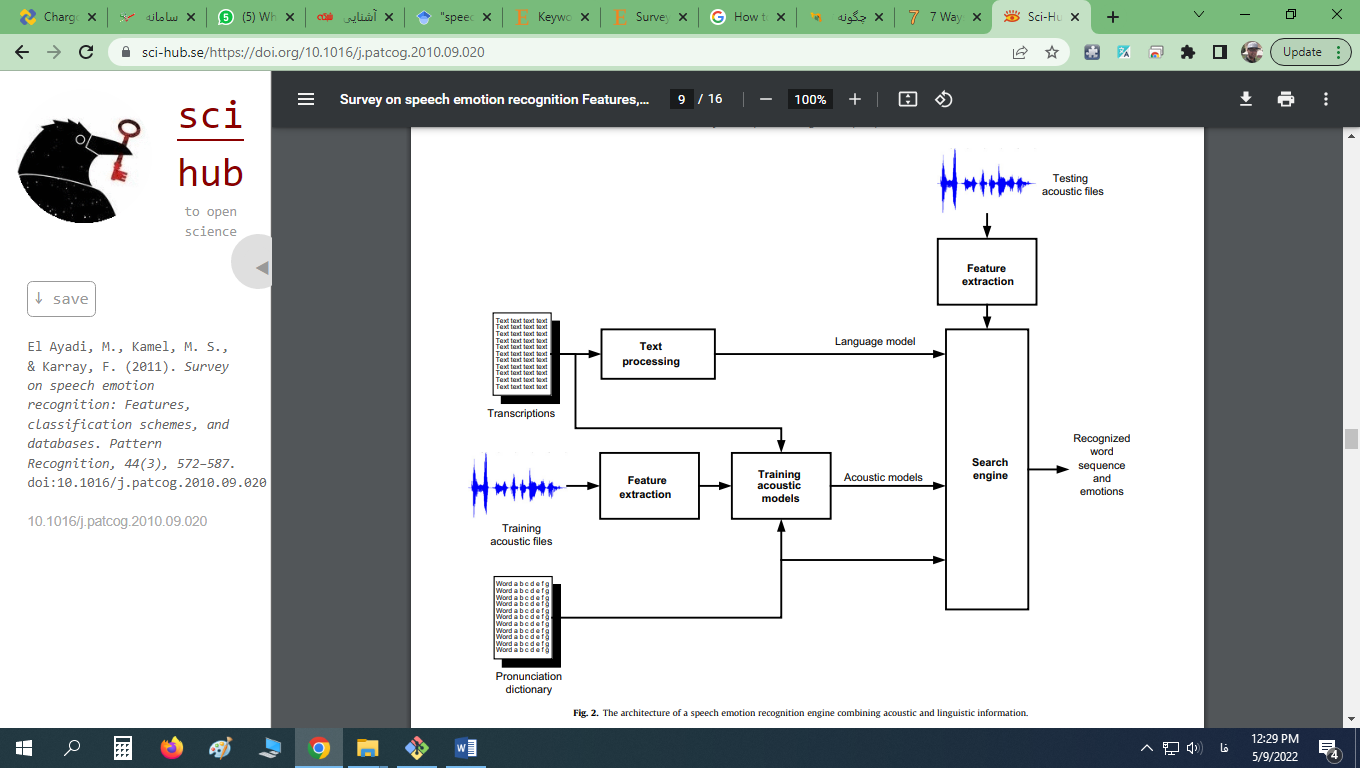
در این بررسی که برای زبان مراتی انجام شده است . هر زبانی درجهان صداهای گفتاری متمایز دارد، یعنی واج. صداهای گفتاری همه زبان ها به حروف صدادار و صامت طبقه بندی می شوند. معمولاً با برخی نمادهای خاص که عموماً آکشاراهای آن زبان نامیده می شوند، نمایش داده می شوند. در اصطلاح آوایی و در اصطلاح زبانی مصوت ها و صامت ها به صورت زیر تعریف می شوند. در اصطلاح آوایی، صدای گفتار به عنوان مصوت تعریف می شود. اگر در تولید آن در حلق و دهان هیچ انسداد و باریک شدن درجه ای وجود نداشته باشد که باعث اصطکاک شنیداری شود، همه اینها مصوت هستند. همه صداهای دیگر به عنوان صامت گرفته می شوند. واحد اصلی سیستم نوشتاری در زبان‌های هندی آکشارا هستند که نمایشی املایی از صداهای گفتاری هستند. یک آکشارا در خط‌های زبان هندی نزدیک به یک هجا است و معمولاً می‌تواند به شکل‌های زیر باشد: C، V، CV، CCV، VC، VV، و CVC که در آن C صامت و V صدادار است. تعداد آواهای گفتاری متمایز از نظر زبانی در یک زبان غالباً موضوع قضاوت است و برای زبان شناسی های مختلف تغییر ناپذیر نیست.

روش تشخیص الگو(Pattern Recognition Approach)

آموزش و مقایسه الگو دو مرحله مهم و حیاتی در فرآیند تطبیق الگوها هستند. در مرحله مقایسه الگو، گفتارهای مبهم و نامشخص ‌به‌طور مستقیم با هر الگوی ‌به‌دست‌آمده در مرحله آموزش مقایسه می‌شوند تا گفتار مبهم بر اساس نزدیک بودن به الگو تشخیص داده شود. روش فوق از یک چارچوب ریاضی یا به عبارت دقیق‌تر مجموعه‌ای از قوانین ریاضی برای ایجاد بازنمایی الگوی گفتاری یکپارچه بر مبنای مجموعه‌ای از نمونه‌های آموزشی برچسب‌گذاری‌شده استفاده می‌کند. هدف این است که فرآیند مقایسه و تطابق الگوها با بالاترین قابلیت اطمینان انجام شود. تشخیص الگو می‌تواند به‌عنوان دسته‌بندی داده‌های ورودی در کلاس‌های شناخته‌شده از طریق استخراج ویژگی‌های مهم یا صفات داده تعریف شود. یک کلاس الگو، یک دسته متمایزشده از طریق برخی صفات و ویژگی‌های مشترک است. ویژگی‌های یک کلاس الگو، صفات نوعی هستند که بین همه الگوهای متعلق به آن کلاس مشترک هستند. ویژگی‌هایی که تفاوت‌های بین کلاس‌های الگو را بیان می‌کنند، اغلب به‌عنوان ویژگی‌های اینترست (Interest Pattern) شناخته می‌شوند. یک الگو توصیفی از یکی از اعضای دسته است که ارائه‌دهنده کلاس الگو است. در بیشتر موارد و برای سهولت، الگوها از طریق یک بردار نمایش داده می‌شوند. در شش دهه گذشته، تطابق الگو مشهورترین روش تشخیص گفتار بوده است.

Survey on speech emotion recognition: Features, classification schemes, and databases Moataz El Ayadi a,, Mohamed S. Kamel b , Fakhri Karray

در این مقاله به برسی ویژگی ها ، دسته بندیهای ویژگی ها ، عملیات پردازش گفتار، یکپارچه سازی ویژگی ها با دیگر منابع ، طبقیه بندی

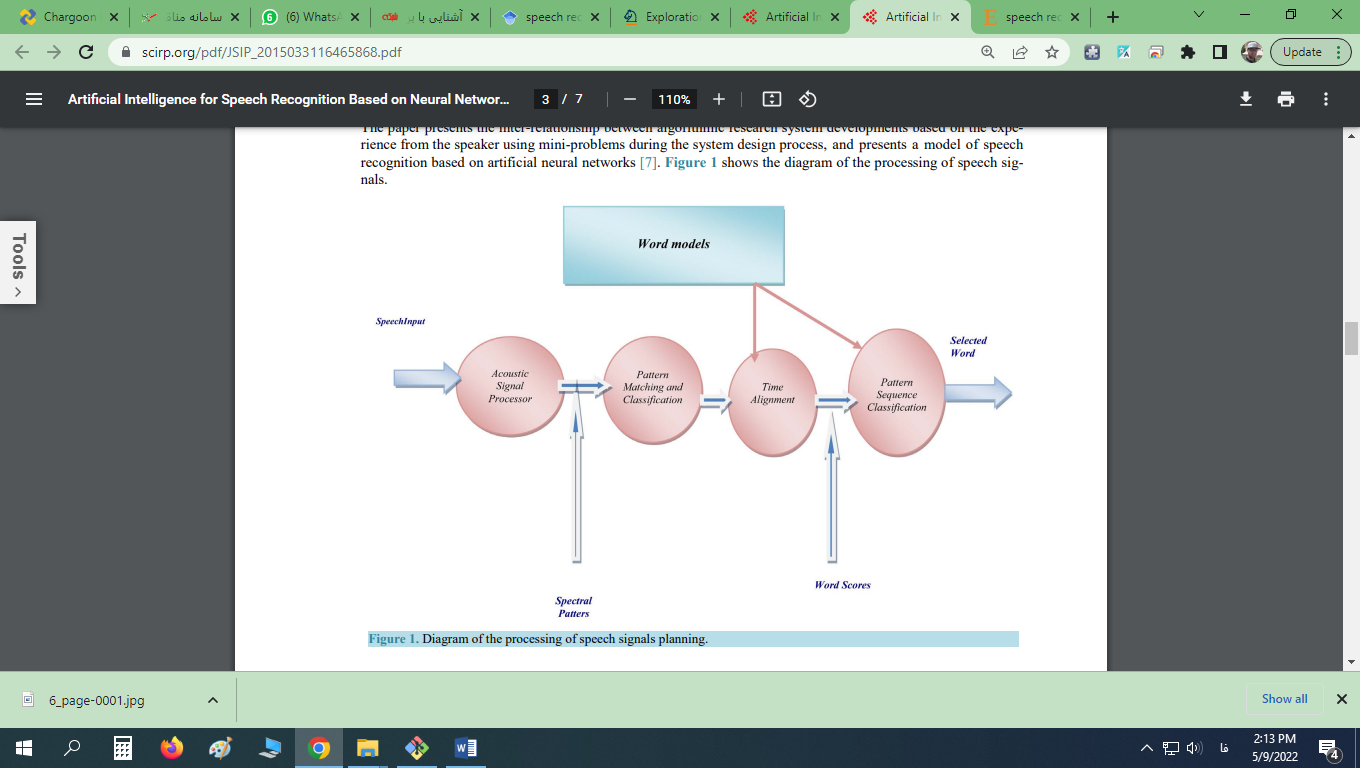


روش هوش مصنوعی (Artificial Intelligence Approach)

در رویکرد هوش مصنوعی ترکیبی از روش‌های آکوستیک-آوایی و تشخیص الگو و مفاهیم مرتبط با دو روش فوق استفاده می‌شود. در تشخیص گفتار خودکار، دو روش اصلی برای تطابق الگو وجود دارد که تطابق الگوی قطعی با استفاده از کش و قوس‌های زمانی پویا (مکانیزم انطباق زمانی پویا) و تطابق الگوی تصادفی با استفاده از مدل‌های پنهان مارکوف (Hidden Markov models) هستند.

Artificial Intelligence for Speech Recognition Based on Neural Networks Takialddin Al Smadi1, Huthaifa A. Al Issa2, Esam Trad3, Khalid A. Al Smadi4 1 Department of Communications and Electronics Engineering, College of Engineering, Jerash University, Jerash, Jordan 2 Department of Electrical and Electronics Engineering, Faculty of Engineering, Al-Balqa Applied University, Al-Huson College University, Al-Huson, Jordan 3 Departments of Communications and Computer Engineering, Jadara University, Irbid, Jordan 4 Jordanian Sudanese Colleges for Science & Technology, Khartoum, Sudan Email: [dsmadi@rambler.ru](mailto:dsmadi@rambler.ru)

در این مقاله درباره نقشه هایی از پیش طراحی شده برای تشخیص خودکار گفتار استفاده شده

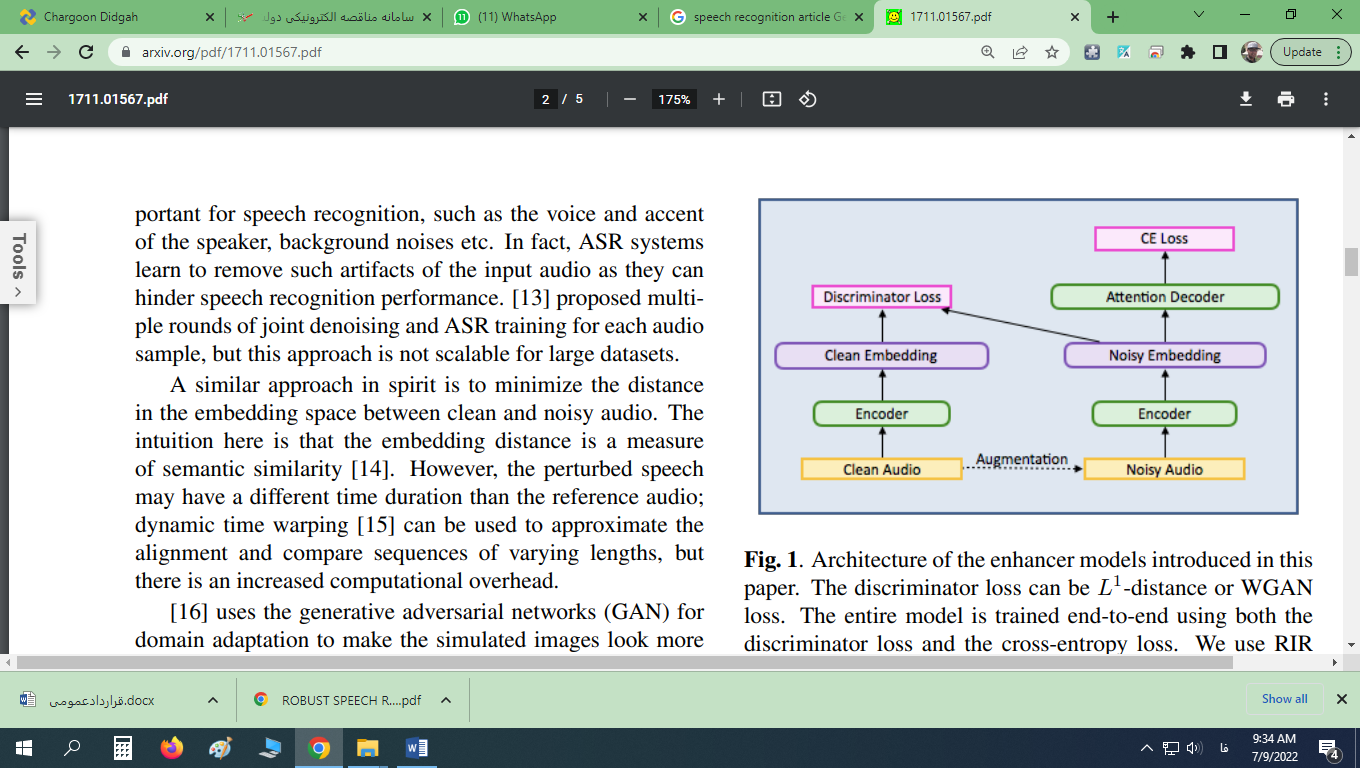


روش یادگیری مولد (Generative Learning Approach)

مدل‌های پنهان مارکوف مبتنی بر مدل‌های آمیخته گوسی (Gaussian Mixture) رایج‌ترین روش یادگیری مولد در سیستم‌های تشخیص گفتار ASR هستند و مدت‌زمان زیادی است که از آن‌ها استفاده می‌شود. مدل‌ آمیخته گوسی از معروف‌ترین الگوریتم‌های خوشه‌بندی است. در الگوریتم خوشه‌بندی آمیخته گوسی، مفروض است هر خوشه از داده‌ها بر مبنای توزیع گوسی (نرمال) ایجاد شوند و داده‌ها نمونه‌ای از توزیع آمیخته گوسی باشند. این مدل با هدف تخمین پارامترهای توزیع هر یک از خوشه‌ها و تعیین برچسب برای مشاهدات استفاده می‌شود. به این ترتیب مشخص می‌شود که هر مشاهده به چه خوشه‌ای اختصاص دارد.

ROBUST SPEECH RECOGNITION USING GENERATIVE ADVERSARIAL NETWORKS Anuroop Sriram∗ , Heewoo Jun∗ , Yashesh Gaur, Sanjeev Satheesh Baidu Research, Sunnyvale, CA, USA

در این مقاله از چارچوب شبکه متخاصم مولد (GAN) برای افزایش استحکام مدل‌های seq-to-seq به شکلی مقیاس‌پذیر و انتها به انتها استفاده کرده اند. انکودر به عنوان جنریتور GAN آموزش میبیند و یک امبدینگ غیر قابل تشخیص بین محیط نویزی و محیط تمیز ایجاد میکند. اگر محدودیتی نباشد حتی بدون همترازی یا استنتاج های پیچیده و یا حتی وقتی آگمنتیشن قابل انجام نیست در تئوری این روش قابل اطمینان است.



discriminator loss به عنوان فاصله L1 یا وزن گن است. کل مدل به صورت اند تو اند هم با ال وان هم با کراس انتروپی مدل شده است از RIR برای شبیه سازی صدا استفاده کرده اند.

یادگیری مبتنی بر تمایز (Discriminative Learning)

استفاده از مدل متمایزگر روی مدل مولد از پارادایم‌های بارز یادگیری متمایزکننده است. در دهه 1990 میلادی، استفاده از شبکه‌های عصبی به‌شکل پرسپترون چندلایه (MLP) با تابع غیرخطی softmax در لایه نهایی مورد توجه بسیاری از کارشناسان قرار گرفت. هنگامی‌که خروجی پرسپترون چندلایه به یک مدل پنهان مارکوف وارد می‌شود، این امکان وجود دارد تا یک مدل توالی متمایز خوب یا ترکیبی MLP-HMM ایجاد کرد، زیرا خروجی را می‌توان به‌عنوان احتمال شرطی تفسیر کرد. پژوهشگران تحقیقات زیادی روی این حوزه انجام دادند تا شبکه عصبی پرسپترون چندلایه بتواند به ساده‌ترین شکل زیرمجموعه‌ای از ویژگی‌ها را در ترکیب با ویژگی‌های قدیمی و سنتی و برای مولد مدل پنهان مارکوف تولید ‌کند. اواخر دهه 80 میلادی، شبکه‌های عصبی آموزش‌داده‌شده با مکانیزم پس‌انتشارخطا (Back Propagation) (الگوریتمی در حوزه یادگیری نظارتی شبکه عصبی با استفاده از گرادیان کاهشی است. در روش مذکور، برای یک شبکه عصبی مصنوعی و تابع خطای مشخص، گرادیان تابع خطا نسبت به وزن‌های شبکه عصبی محاسبه می‌شود.) تبدیل به محبوب‌ترین روش مدل‌سازی آکوستیک به‌منظور تشخیص گفتار شدند. بر عکس مدل پنهان مارکوف، شبکه‌های عصبی هیچ تصوری درباره خصوصیات آماری از ویژگی‌ها ندارند.

Discriminative Learning in Speech Recognition Xiaodong He and Li Deng {xiaohe, deng}@microsoft.com October 2007 Technical Report MSR-TR-2007-129

به عقیده آنها کلید برای درک فرآیند گفتار، توصیف پویای توالی گفتار یا طول متغیر الگوی آن است. دو مشکلی که روبه رو بودن برای یادگیری مبتنی بر تمایز 1) ساخت تابع هدف برای بهینه سازی؛ و 2) تکنیک های بهینه سازی به درد بخور

طبقه‌بندی‌کننده‌ها/تشخیص‌دهنده‌های تمایز مستقیماً از احتمال پسین کلاس (یا تابع متمایز مرتبط) استفاده می‌کنند که با این استدلال که باید مشکل (طبقه بندی/شناخت) را مستقیماً حل کند و همیشه یک مشکل کلی تر را به عنوان یک مرحله میانی حل کند عمل میکند

در "مرحله میانی" تخمین توزیع مشترک اجتناب شده است. برای مثال برخی از کاستی‌های شناخته شده HMM با استفاده از یادگیری تمایزی «مستقیم» حل شده اند، جایگزینی نیاز به یک مدل احتمالی مولد با مجموعه‌ای از «ویژگی‌ها» که به‌طور انعطاف‌پذیر انتخاب شده و همپوشانی دارند، انجام می‌شود.

یادگیری عمیق (Deep learning)

یادگیری عمیق که به‌عنوان یادگیری ویژگی (Feature Learning) نظارت‌نشده یا یادگیری ارائه (Representation Learning) نیز شناخته می‌شود، شاخه نسبتا جدیدی از یادگیری ماشین است. یادگیری عمیق به‌سرعت در حال تبدیل شدن به فناوری استاندارد برای تشخیص گفتار است و با موفقیت جایگزین روش‌هایی مثل آمیخته گوسی برای تشخیص گفتار و کدگذاری ویژگی‌ها در مقیاس بزرگ شده است. معماری‌های مولد عمیق می‌توانند گونه اول خواص همبستگی یا توزیع‌های آماری مشارکتی با داده‌های قابل رؤیت و کلاس‌های مرتبط با آن‌ها را تشخیص دهند. در این‌جا، قانون بیز می‌تواند برای ساخت این نوع از معماری متمایزگر استفاده شود. از کدگذارهای خودکار عمیق، ماشین‌های بولتزمن عمیق، شبکه‌های Sum-Product، شبکه باور عمیق اصلی (Deep Belief Network) و غیره می‌توان برای این منظور استفاده کرد.

Speech Recognition with Deep Learning 1 Lokesh Khurana, 2 Arun Chauhan, 3 Dr. Mohd Naved, 4,\*Prabhishek Singh 1, 2, 4Amity School of Engineering and Technology, Amity University Uttar Pradesh, Noida, India 3 Assistant Professor, Jagannath University 1 lokeshkhurana98@gmail.com 2 arunchauhan414@gmail.com 3 mohdnaved@gmail.com 4 [prabhisheksingh88@gmail.com](mailto:prabhisheksingh88@gmail.com)

یک "الگوریتم یادگیری عمیق" قادر است گفتار/تصویر را به عنوان ورودی بگیرد و سپس وزن ها و سوگیری ها را به اشیاء موجود در گفتار/تصویر اختصاص دهد. شبکه ای است که برای جستجوی الگوها یا شکل های خاص از عملیات فنی کانولوشن استفاده می کند یا به آن نیاز دارد. دارای ساختار چراغ قوه است. هر چراغ قوه نشان دهنده یک نورون در شبکه عصبی کانولوشنال (CNN) است. ورودی، گفتار/تصویر به اولین لایه لایه Convolutional منتقل می شود. خروجی به عنوان نقشه فعال شده از لایه کانولوشن به دست می آید. فیلترهایی در لایه کانولوشن وجود دارند که وظیفه استخراج ویژگی ها از ورودی و عبور آنها را بر عهده دارند. هر فیلتر یک ویژگی جایگزین برای کمک به پیش‌بینی کلاس درست می‌دهد. padding (صفر padding) برای نگه داشتن یا حفظ اندازه واقعی تصویر استفاده می شود. از تکنیک padding قانونی استفاده می شود زیرا به کاهش تعداد ویژگی ها برای عملکرد بهتر مدل کمک می کند. لایه Pooling برای کاهش ابعاد استفاده می شود.شبکه لایه ادغام احتمالاً تنها بر مرتبط ترین الگوهای کشف شده توسط لایه کانولوشن و لایه RELU تمرکز می کند.