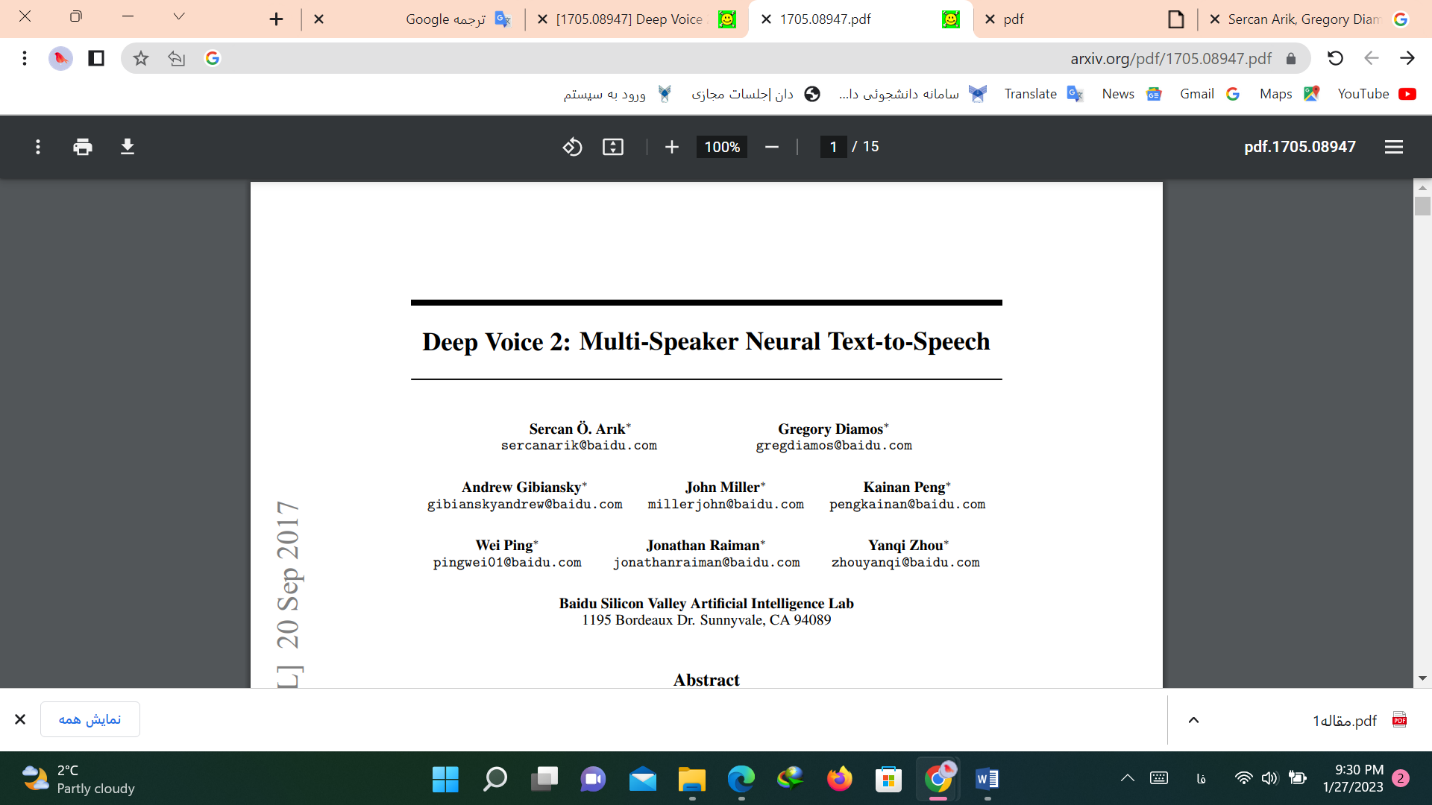


**مقاله 1**

**متن عصبی گوینده به گفتار:**

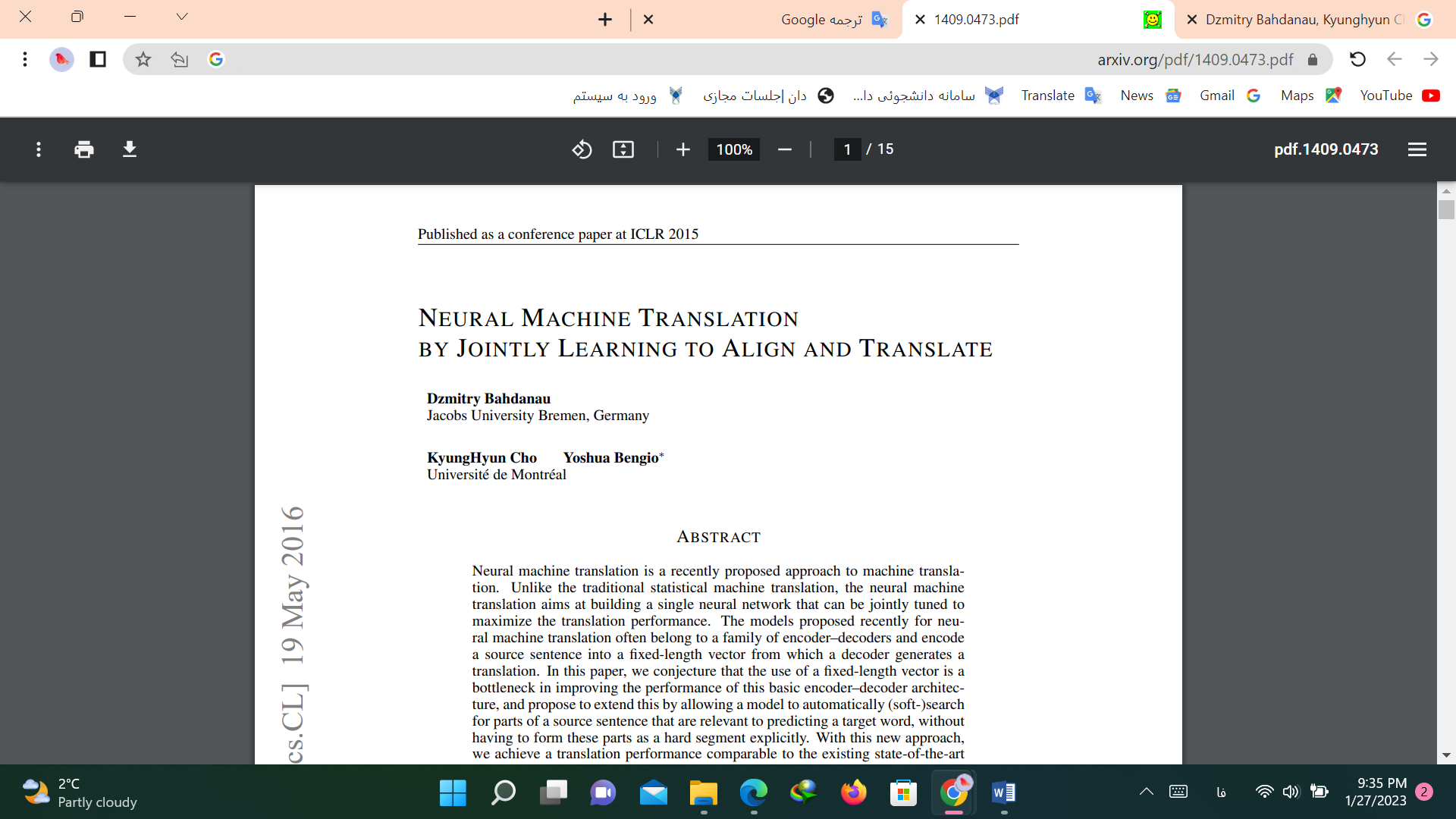
ماDeep Voice 3 را ارائه می‌کنیم، یک سیستم متن به گفتار عصبی مبتنی بر توجه کاملاً کانولوشنی (TTS). Deep Voice 3 با گفتار عصبی پیشرفته مطابقت دارد سیستم های سنتز در طبیعی بودن در حالی که تمرین یک مرتبه قدر سریع تر است. ما Deep Voice 3 را به اندازه‌های مجموعه داده‌ای که برای TTS بی‌سابقه است، تغییر می‌دهیم بیش از هشتصد ساعت صدا از بیش از دو هزار بلندگو. که در علاوه بر این، ما حالت‌های خطای رایج شبکه‌های سنتز گفتار مبتنی بر توجه را شناسایی می‌کنیم، نحوه کاهش آنها را نشان می‌دهیم و چندین روش مختلف ترکیب شکل موج را با هم مقایسه می‌کنیم. ما همچنین نحوه مقیاس بندی استنتاج را توضیح می دهیم ده میلیون درخواست در روز بر روی یک سرور GPU واحد.



**مقاله 2**

**تبدیل متن به گفتار عصبی با چند بلندگو:**

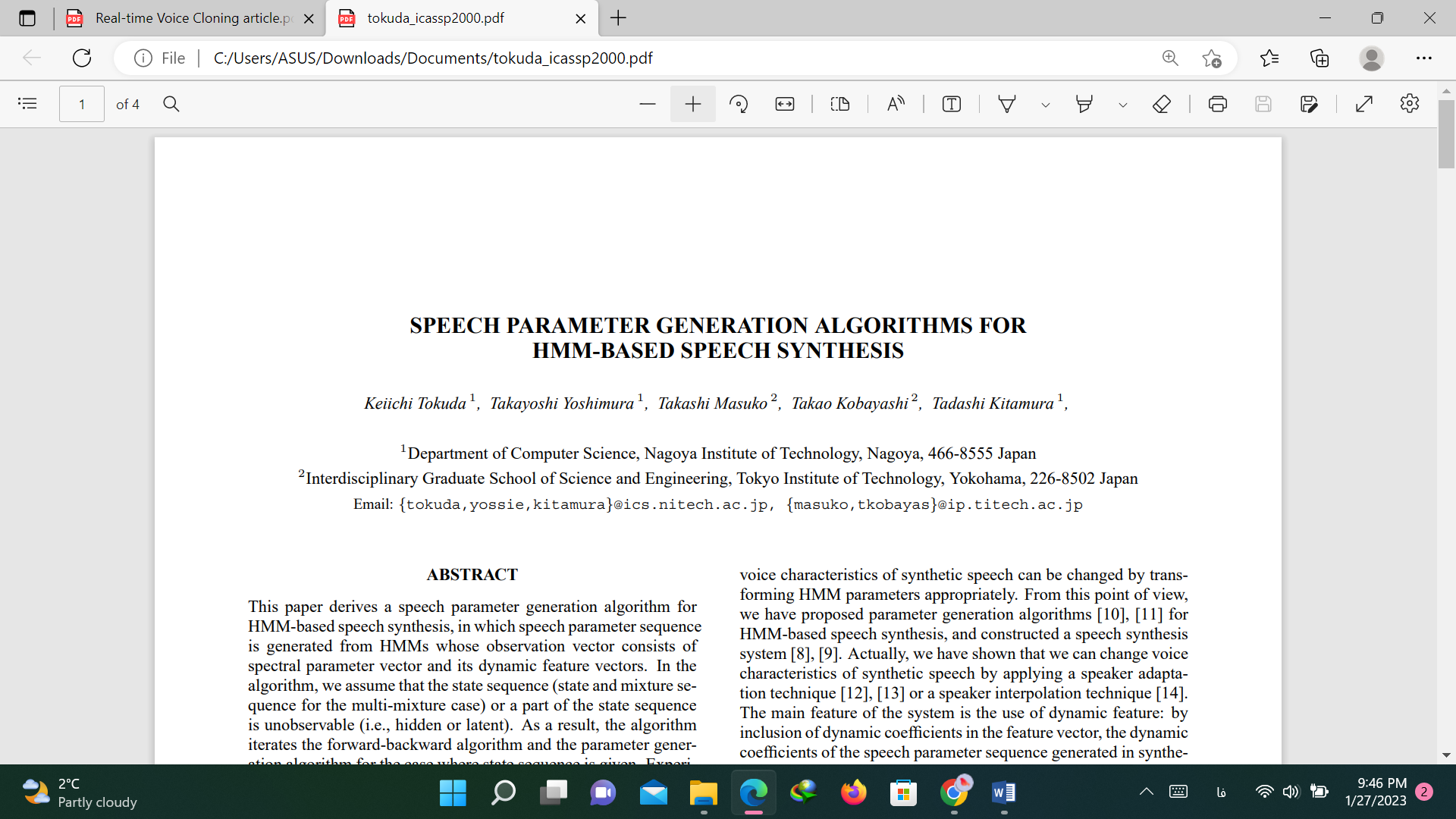
ما تکنیکی را برای تقویت متن به گفتار عصبی (TTS) با تعبیه‌های بلندگوی آموزش پذیر کم‌بعد برای تولید صداهای مختلف از مدل تک به عنوان نقطه شروع، ما بهبودهایی را نسبت به دو رویکرد پیشرفته برای TTS عصبی تک بلندگو نشان می‌دهیم: Deep Voice 1 و Tacotron. Deep Voice 2 را معرفی می کنیم که بر اساس خط لوله مشابهی با Deep ساخته شده است Voice 1، اما ساخته شده با بلوک های ساختمانی با کارایی بالاتر و نشان می دهد بهبود قابل توجه کیفیت صدا نسبت به Deep Voice 1. ما Tacotron را بهبود می بخشیم با معرفی یک کد صوتی عصبی پس از پردازش، و نشان دادن یک بهبود کیفیت صدا سپس تکنیک خود را برای چند بلندگو نشان می دهیم سنتز گفتار برای Deep Voice 2 و Tacotron روی دو TTS چند بلندگو مجموعه داده ها ما نشان می دهیم که یک سیستم عصبی TTS می تواند صدها مورد منحصر به فرد را یاد بگیرد صداها از کمتر از نیم ساعت داده در هر بلندگو، در حالی که به صدای بالا دست می یابند سنتز با کیفیت و حفظ هویت گوینده تقریباً کامل



**مقاله 3**

**ترجمه دستگاه عصبی  
با یادگیری مشترک تراز کردن و ترجمه کردن:**

ترجمه ماشینی عصبی رویکردی است که اخیرا برای ترجمه ماشینی پیشنهاد شده است. برخلاف ترجمه ماشینی آماری سنتی، ماشین عصبی هدف ترجمه ایجاد یک شبکه عصبی واحد است که بتوان به طور مشترک روی آن تنظیم کرد به حداکثر رساندن کارایی ترجمه مدل‌هایی که اخیراً برای ترجمه ماشین عصبی پیشنهاد شده‌اند اغلب به خانواده‌ای از رمزگذار-رمزگشاها و کدگذاری تعلق دارند. یک جمله منبع به یک بردار با طول ثابت که رمزگشا از آن یک را تولید می کند ترجمه در این مقاله، حدس می زنیم که استفاده از یک بردار با طول ثابت a است گلوگاه در بهبود عملکرد این معماری رمزگذار-رمزگشای اساسی است و پیشنهاد می‌کنیم که با اجازه دادن به یک مدل برای جستجوی خودکار (نرم) این را گسترش دهیم. برای بخش‌هایی از جمله مبدا که با پیش‌بینی کلمه مقصد مرتبط هستند، بدون باید این قطعات را به صورت یک بخش سخت به صراحت تشکیل دهند. با این رویکرد جدید، ما به یک عملکرد ترجمه قابل مقایسه با آخرین وضعیت موجود دست می‌یابیمسیستم مبتنی بر عبارت در وظیفه ترجمه انگلیسی به فرانسوی. علاوه بر این، تجزیه و تحلیل کیفی نشان می‌دهد که هم‌ترازی‌های (نرم) یافت شده توسط مدل با هم موافق هستند.



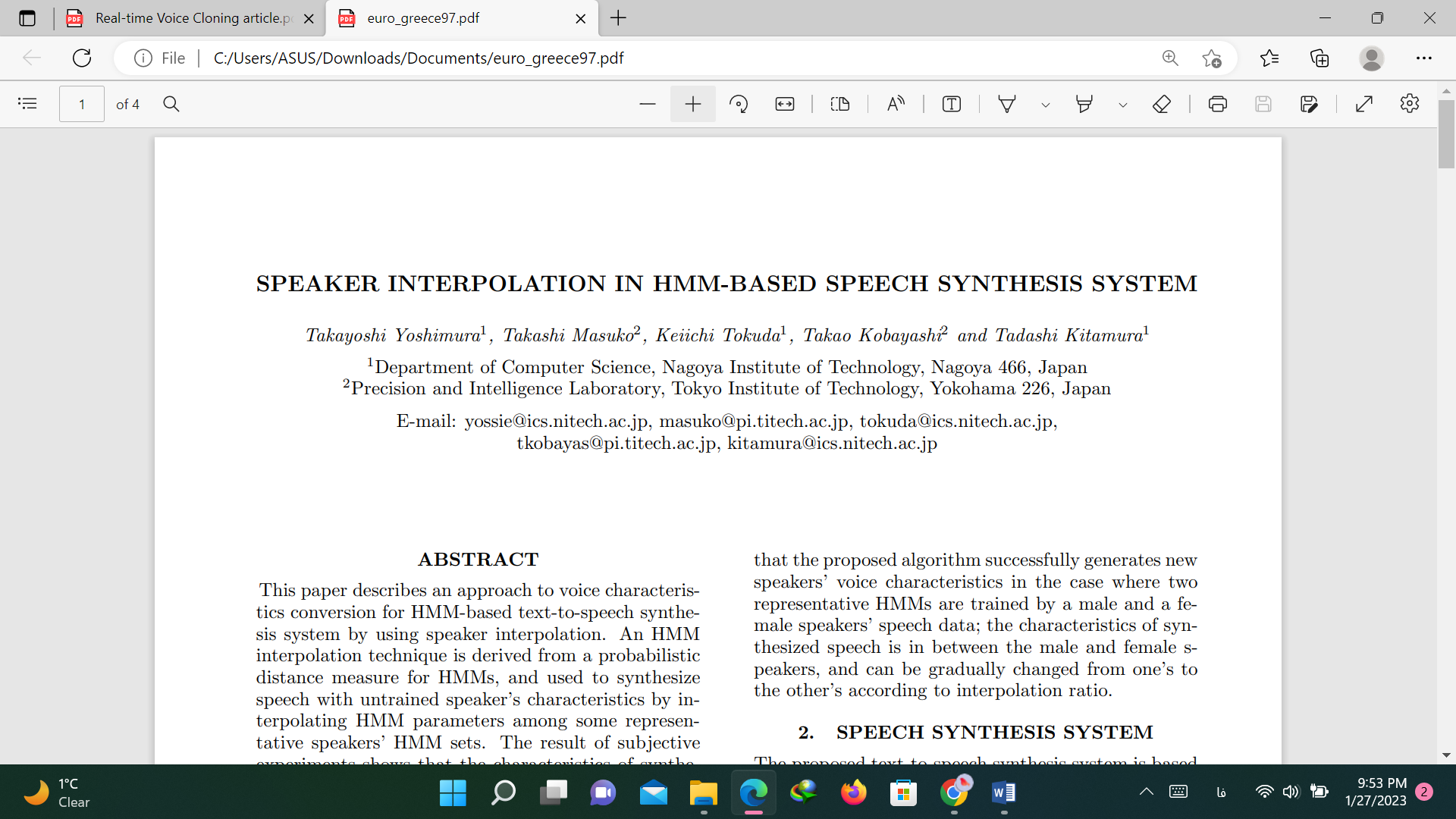
**مقاله4**

**الگوریتم های تولید پارامتر گفتار برای سنتز گفتار مبتنی بر HMM:**

این مقاله یک الگوریتم تولید پارامتر گفتار را برای سنتز گفتار مبتنی بر HMM، که در آن توالی پارامتر گفتار

از HMMهایی که بردار مشاهدات آنها تشکیل شده است تولید می شود بردار پارامتر طیفی و بردارهای ویژگی پویا آن. در

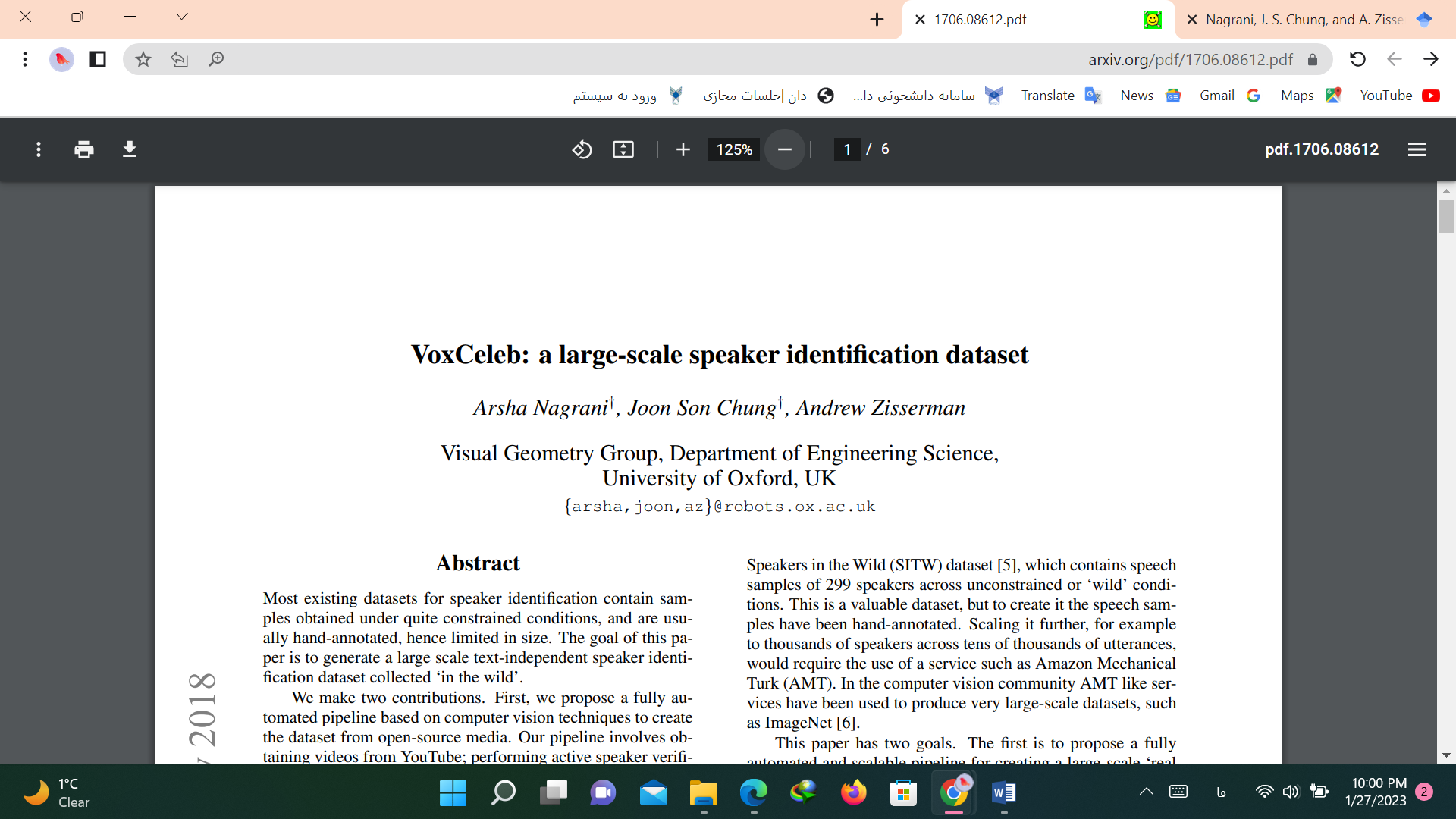
الگوریتم، ما فرض می کنیم که دنباله حالت (توالی حالت و مخلوط برای حالت چند مخلوط) یا بخشی از دنباله حالت غیر قابل مشاهده است (یعنی پنهان یا پنهان). در نتیجه الگوریتم الگوریتم رو به عقب و الگوریتم تولید پارامتر را برای مواردی که توالی حالت داده شده است، تکرار می کند. نتایج تجربی نشان می‌دهد که با استفاده از الگوریتم می‌توانیم تولید مثل کنیم ساختار فرمانت روشن از HMM های چند مخلوط در مقایسه با آن که از HMM های تک مخلوط تولید می شود.



**مقاله 5**

**درون یابی بلندگو در سیستم سنتز گفتار مبتنی بر HMM:**

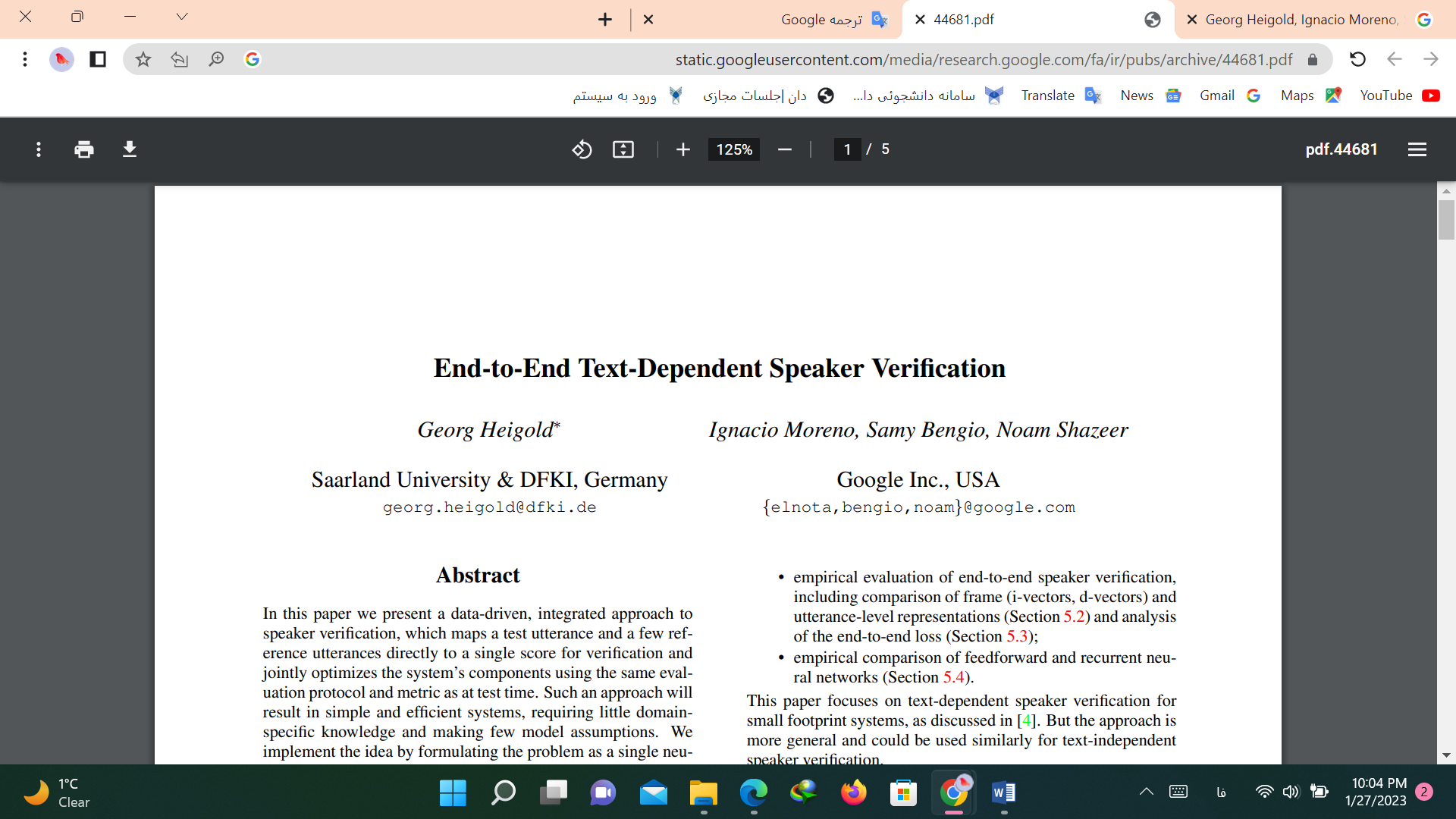
این مقاله رویکردی را برای تبدیل تیک ویژگی‌های صوتی برای سیستم ترکیبی متن به گفتار مبتنی بر HMM با استفاده از درون یابی بلندگو توصیف می‌کند. یک تکنیک درون یابی HMM از یک اندازه گیری فاصله احتمالی برای HMM ها مشتق شده است، و برای ترکیب گفتار با ویژگی های سخنران آموزش ندیده از طریق درون یابی پارامترهای HMM در میان مجموعه های HMM برخی از سخنرانان نماینده استفاده می شود. نتیجه آزمایش‌های ذهنی نشان می‌دهد که ویژگی‌های گفتار با اندازه ترکیبی با تغییر نسبت درونیابی به تدریج از یکی به دیگری تغییر می‌کند.



**مقاله 6**

**یک مجموعه داده شناسایی بلندگو در مقیاس بزرگ:**

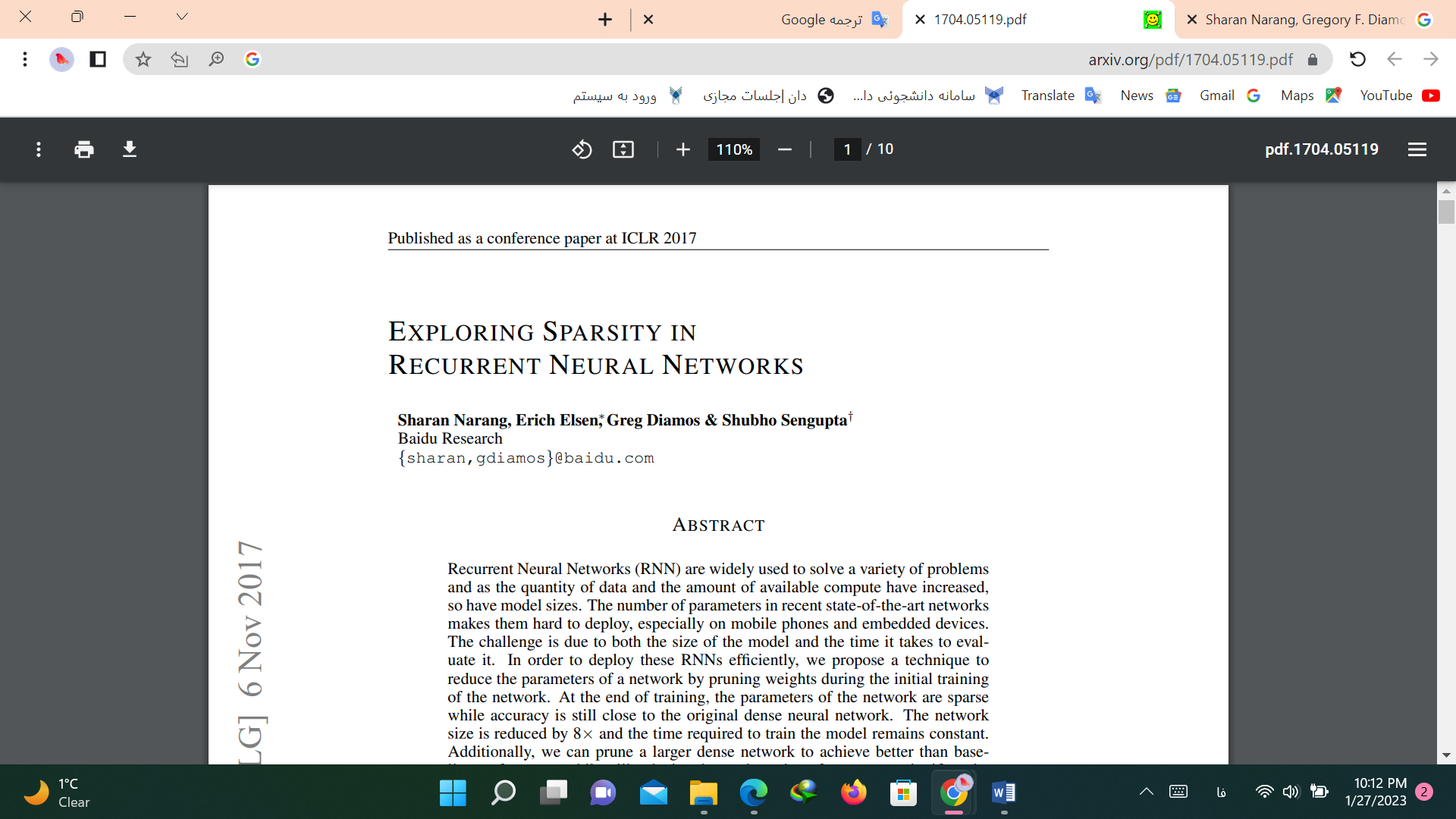
بیشتر مجموعه داده‌های موجود برای شناسایی بلندگو حاوی نمونه‌هایی هستند که تحت شرایط کاملاً محدود به دست آمده‌اند و معمولاً به صورت دستی حاشیه‌نویسی می‌شوند، بنابراین اندازه آنها محدود است. هدف این مقاله تولید یک مجموعه داده شناسایی سخنران مستقل از متن در مقیاس بزرگ است که "در طبیعت" جمع آوری شده است. ما دو کمک می کنیم. ابتدا، ما یک خط لوله کاملاً خودکار مبتنی بر تکنیک‌های بینایی رایانه برای ایجاد مجموعه داده از رسانه منبع باز پیشنهاد می‌کنیم. خط لوله ما شامل به دست آوردن ویدیوها از YouTube است. انجام راستی‌آزمایی بلندگوی فعال با استفاده از یک شبکه عصبی همگام‌سازی دو جریانی (CNN)، و تأیید هویت بلندگو با استفاده از تشخیص چهره مبتنی بر CNN. ما از این خط لوله برای سرپرستی VoxCeleb استفاده می‌کنیم که حاوی صدها هزار گفته «دنیای واقعی» برای بیش از 1000 سلبریتی است. سهم دوم ما اعمال و مقایسه تکنیک‌های مختلف شناسایی بلندگوهای پیشرفته در مجموعه داده‌های ما برای ایجاد عملکرد پایه است. ما نشان می‌دهیم که یک معماری مبتنی بر CNN بهترین عملکرد را هم برای شناسایی و هم برای تأیید به دست می‌آورد.



**مقاله7**

**تأیید صحت بلندگو وابسته به متن:**

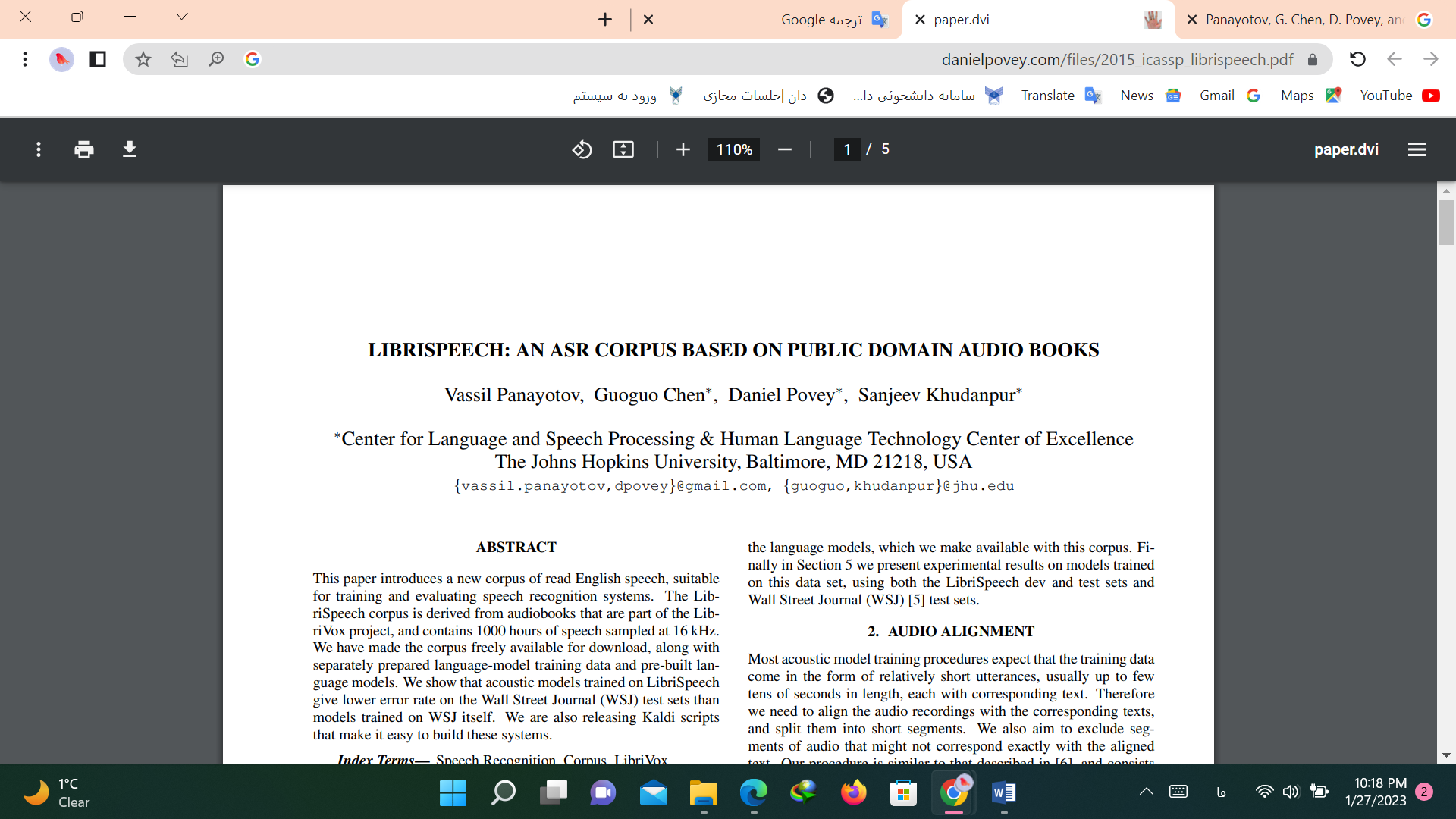
در این مقاله ما یک رویکرد داده محور و یکپارچه ارائه می کنیم تأیید سخنران، که یک گفته آزمایشی و چند گفته مرجع را مستقیماً به یک نمره واحد برای تأیید نگاشت می‌کند و به طور مشترک اجزای سیستم را با استفاده از پروتکل ارزیابی و متریک مشابه در زمان آزمایش بهینه می‌کند. چنین رویکردی منجر به سیستم‌های ساده و کارآمدی می‌شود که به دانش مختص حوزه کمی نیاز دارند و مفروضات مدل کمی دارند. ما این ایده را با فرمول‌بندی مسئله به‌عنوان یک معماری شبکه عصبی منفرد، از جمله تخمین مدل بلندگو تنها بر روی چند گفته، پیاده‌سازی می‌کنیم و آن را در معیار (Ok Google) داخلی خود برای تأیید بلندگوی وابسته به متن ارزیابی می‌کنیم. به نظر می رسد رویکرد پیشنهادی برای برنامه های کاربردی داده های بزرگ مانند ما که به سیستم های بسیار دقیق و آسان برای نگهداری با ردپای کوچک نیاز دارند بسیار مؤثر است.



**مقاله 8**

**کاوش پراکندگی در شبکه های عصبی بازگشتی:**

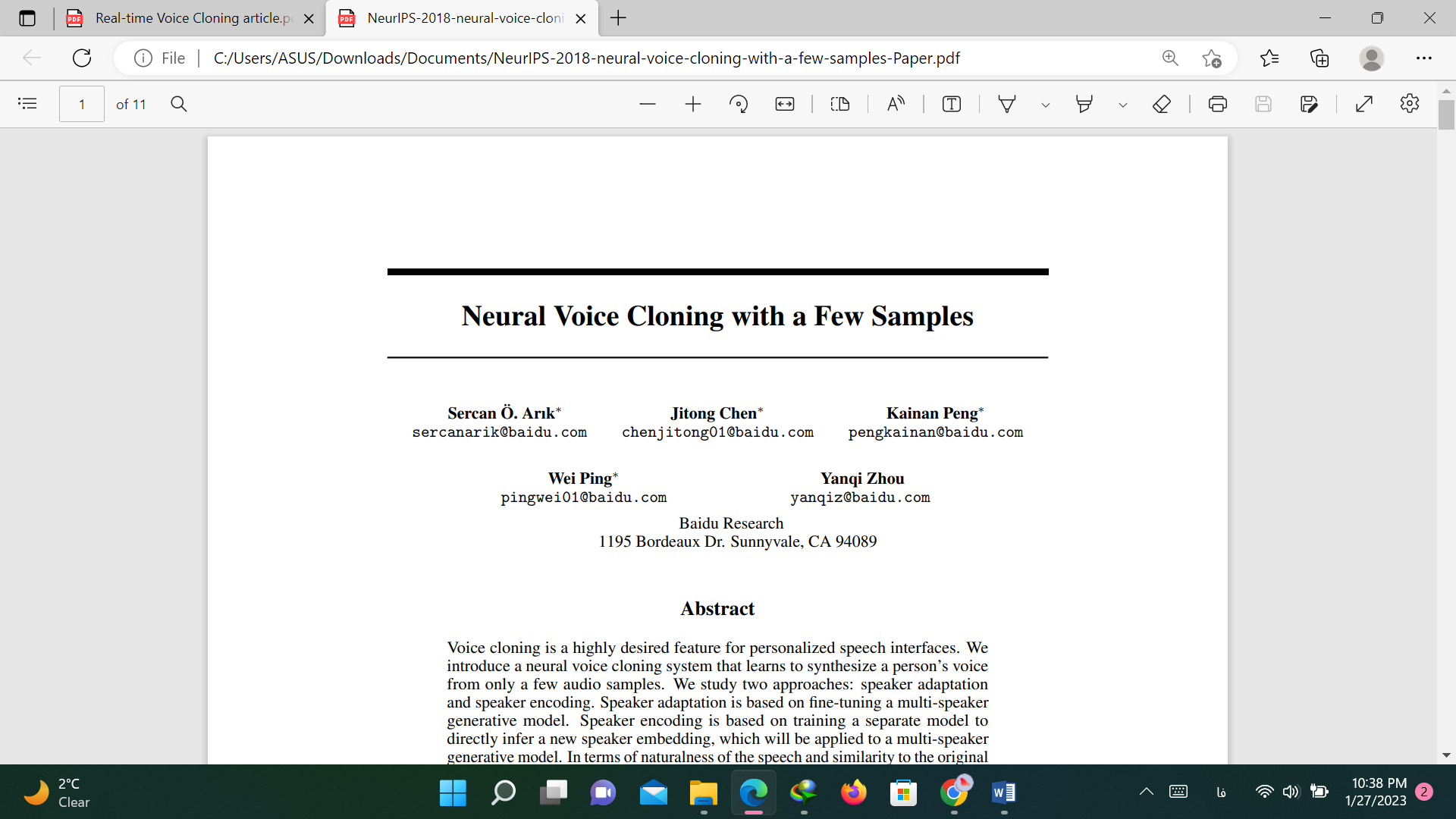
شبکه های عصبی بازگشتی (RNN) به طور گسترده ای برای حل مسائل مختلف مورد استفاده قرار می گیرند و با افزایش کمیت داده ها و مقدار محاسبات موجود، اندازه مدل ها نیز افزایش می یابد. تعداد پارامترها در شبکه های پیشرفته اخیر، استقرار آنها را سخت می کند، به خصوص در تلفن های همراه و دستگاه های تعبیه شده. این چالش هم به دلیل اندازه مدل و هم زمان ارزیابی آن است. به منظور استقرار این RNN ها به طور موثر، ما تکنیکی را برای کاهش پارامترهای یک شبکه با هرس کردن وزن ها در طول آموزش اولیه شبکه پیشنهاد می کنیم. در پایان آموزش، پارامترهای شبکه پراکنده هستند در حالی که دقت هنوز نزدیک به شبکه عصبی متراکم اصلی است. اندازه شبکه 8× کاهش می یابد و زمان مورد نیاز برای آموزش مدل ثابت می ماند. علاوه بر این، ما می‌توانیم یک شبکه متراکم بزرگ‌تر را برای دستیابی به عملکرد بهتر از خط پایه هرس کنیم و در عین حال تعداد کل پارامترها را به میزان قابل توجهی کاهش دهیم. هرس RNN اندازه مدل را کاهش می دهد و همچنین می تواند به افزایش سرعت زمان استنتاج قابل توجه با استفاده از ضرب ماتریس پراکنده کمک کند. معیارها نشان می‌دهند که با استفاده از تکنیک ما، اندازه مدل را می‌توان تا 90 درصد کاهش داد و سرعت آن حدود 2× تا 7× است.



**مقاله9**

**مجموعه ای از ASR بر اساس کتاب های صوتی حوزه عمومی:**

این مقاله مجموعه جدیدی از گفتار خوانده شده انگلیسی را معرفی می کند که برای آموزش و ارزیابی سیستم های تشخیص گفتار مناسب است. مجموعه LibriSpeech از کتاب‌های صوتی که بخشی از پروژه LibriVox هستند مشتق شده است و شامل 1000 ساعت گفتار نمونه‌برداری شده با فرکانس 16 کیلوهرتز است. ما مجموعه را به همراه داده های آموزش مدل زبان و مدل های زبان از پیش ساخته شده به صورت رایگان برای دانلود در دسترس قرار داده ایم. ما نشان می‌دهیم که مدل‌های آکوستیک آموزش‌دیده‌شده در LibriSpeech نسبت به مدل‌های آموزش‌دیده‌شده در خود WSJ نرخ خطای کمتری را در مجموعه‌های آزمایشی وال استریت ژورنال (WSJ) نشان می‌دهند. ما همچنین اسکریپت های Kaldi را منتشر می کنیم که ساخت این سیستم ها را آسان می کند.



**مقاله10**

**شبیه سازی صدای عصبی با چند نمونه:**

شبیه سازی صدا یک ویژگی بسیار مطلوب برای رابط های گفتاری شخصی سازی شده است. ما یک سیستم شبیه سازی صدای عصبی را معرفی می کنیم که می آموزد صدای یک فرد را تنها از چند نمونه صوتی ترکیب کند. ما دو رویکرد را مطالعه می کنیم: سازگاری بلندگو و رمزگذاری بلندگو. انطباق بلندگو بر اساس تنظیم دقیق یک مدل تولیدی چند بلندگو است. رمزگذاری بلندگو بر اساس آموزش یک مدل جداگانه برای استنباط مستقیم تعبیه بلندگوی جدید است که برای یک مدل تولیدی چند بلندگو اعمال خواهد شد. از نظر طبیعی بودن گفتار و شباهت به بلندگوی اصلی، هر دو رویکرد می توانند عملکرد خوبی را حتی با چند فایل صوتی شبیه سازی کنند. 2 در حالی که سازگاری بلندگو می تواند به طبیعی بودن و شباهت کمی بهتر دست یابد، زمان شبیه سازی و حافظه مورد نیاز برای رویکرد رمزگذاری بلندگو به میزان قابل توجهی کمتر است و آن را برای استقرار با منابع کم مطلوب تر می کند.