
بسم الله الرحمن الرحيم

بسمه تعالی

دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده مهندسی برق

پایان نامه دکترای مهندسی برق- مخابرات(سیستم)

عنوان:

افزایش کیفیت و بهبود عملکرد سیستم های تبدیل گفتار فارسی

استاد راهنما:

دکتر ابوالقاسم صیادیان

استاد مشاور:

دکتر حمید شیخ زاده

نگارش:

مهدی اسلامی

۱۳۸۵

بسمه تعالي

تاريخ : ۸۶/۴/۲۳
شماره : .....

فرم اطلاعات پايان نامه
كارشناسي ارشد و دكترا

دانشگاه صنعتي اميركبير
(پلي تكنيک تهران)

معاونت پژوهشي
فرم پروژه تحصيلات تكميلي ۷

مشخصات دانشجو

نام و نام خانوادگي :**مهدی اسلامی** ☐ دانشجوي آزاد ☒ بورسيه ☐ معادل

شماره دانشجويي: **۷۹۲۲۳۹۱۹** دانشكده : **مهندسی برق** رشته تحصيلي: **مخابرات سيستم**

نام و نام خانودگي استاد راهنما : دكتر ابوالقاسم صيادیان

عنوان به فارسي: افزايش كيفيت و بهبود عملكرد سيستم تبديل گفتار فارسی

عنوان به انگليسي: Quality Improvement of Farsi Voice Conversion System.

نوع پروژه دكترا: ☒ **كاربردي** ☐ **بنيادي** ☐ **توسعه اي** ☒ **نظري**

تاريخ شروع: ۸۲/۱/۱۱ **تاريخ خاتمه :** ۸۵/۱۲/۱۴ **تعداد واحد:**

سازمان تامين كننده اعتبار : مركز تحقيقات مخابرات ايران

واژه هاي كليدي به فارسي : تبديل گفتار با كيفيت بالا- مدل مخلوط گوسي GMM- مدل تعميم يافته هارمونيكي GHM– تبديل طيفي

واژه هاي كليدي به انگليسي : High Quality Voice Conversion, GMM Model, GHM Model, Spectral Conversion, Statistical Modelling

---

نظرها و پيشنهادها به منظور بهبود فعاليت هاي پژوهشي دانشگاه:
افزايش امكانات و بودجه تحقيقاتی
استاد راهنما: دكتر ابوالقاسم صيادیان

دانشجو:مهدی اسلامی

امضاء استاد راهنما : **تاريخ:**

نسخه ۱ : معاونت پژوهشي
نسخه ۲ : كتابخانه و به انضمام دو جلد پايان نامه به منظور تسويه حساب با كتابخانه و مركز اسنادومدارک علمي

تقدیم به:

پدر بزرگوارم ...

مادر فداکارم ...

همسر عزیزم ...

و همه شهیدان راه وطن...

تشکر و قدردانی:

با تشکر از استاد محترم راهنما، آقای دکتر صیادیان که در تمامی لحظات انجام پروژه دکتری، راهنمایی ها و دلگرمی های ایشان مشوق اصلی اینجانب در انجام این مهم بوده است و تشکر از پدر گرامی که حمایت ها و تشویق های بی دریغ ایشان سبب ایجاد محیطی آرام و صمیمی در خارج از محیط دانشگاه بود و همچنین تشکر و قدردانی از الطاف و صبر خانواده گرامی خصوصا خواهران و برادر بزرگوار که سبب رفع خستگی ها بود و تشکر از آقای دکتر شیخ زاده که به دفعات از نظرات ارزشمند ایشان در انجام پروژه دکتری بهره مند گردیدم و همچنین تشکر و قدردانی از تمامی اساتید بزرگوار دانشگاه خصوصا: آقایان دکتر فائز، دکتر آقایی نیا، دکتر احدی، دکتر طاهری، دکتر الماس گنج، دکتر مرادی، دکتر عارف، دکتر محامدپور، دکتر احمدیان، دکتر کلانتری، دکتر جبه دار، دکتر جمالی، دکتر احمدی، دکتر مشیری، دکتر فرهنگی، دکتر آرزم و .... و تمامی معلمین بزرگوار در طول دوران تحصیلی، آقایان افتخاری، معصومی، احمدی، تفضلیان، اقبالیون، خادمی، بطحایی، اکبری، امینی، رفیعی، پوستچی، صالحی، فارسی، داوودآبادی و ... که افتخار شاگردی ایشان را داشته ام و تشکر از مدیریت محترم کارخانجات مخابراتی ایران بالاخص آقای مهندس مطلع و مدیریت های قبلی، آقای مهندس واحدی و استاد بزرگوار، آقای مهندس دانشمند که حمایت آنها را همواره در انجام این مهم در کنار خود احساس می نمودم و تشکر و قدردانی از الطاف اساتید بزرگوار: آقایان دکتر عابدی، دکتر مغانی، دکتر محمدی، دکتر کوهساری و دکتر عبدی پور که محیطی صمیمی، پر نشاط و مجهز را در دانشکده مهندسی برق جهت انجام فعالیت های آموزشی/پژوهشی دانشجویان فراهم نموده اند و تشکر از حمایت مرکز تحقیقات مخابرات ایران در انجام این پروژه و در نهایت تشکر از تمام کسانی که اینجانب را در تدوین این رساله یاری نمودند.

چکیده:

در این رساله به مطالعه و پیاده سازی سیستم تبدیل گفتار با کیفیت بالا پرداخته شده و روش های افزایش کیفیت و بهبود عملکرد آنها در زبان فارسی مورد بررسی قرار گرفته است. در سیستم های تبدیل گفتار، گوینده A (مبدا) عباراتی را بیان می کند و هدف از آن عبارت است از تغییر متکلم جملات بیان شده از گوینده A به گوینده B(مقصد). کاربرد سیستم های تبدیل گفتار در ساخت پایگاه دادگان گفتاری جهت کاربرد در سیستم های تبدیل متن به گفتار و بازشناسی گفتار است. همچنین این سیستم قابل استفاده در صداگذاری فیلم ها و ... می باشد. در این قبیل کاربردها، صدای هر گوینده در محیط استودیو ضبط شده است و نیازی به پردازش بلادرنگ ندارد.

روش آماری مبتنی بر GMM بهترین کارایی را در مقایسه با روش های دیگر دارد. این روش بعلت دخالت دادن تمامی خوشه ها در تولید یک بردار برای گوینده جدید، دارای حالت بلـورشدگی(کاهـش وضوح) در صدای بازسازی شده می باشد به گونه ای که بازسازی صدا (با تغییر گوینده) توسط روشهای مذکور نسبت به حالت کاملاً طبیعی، فاصله زیادی دارد. در روش GMM(2) از مدل های متعدد GMM برای مدلسازی هر واج استفاده نموده ایم. همچنین در مرحله متناظرکردن خوشه های هر حالت، قبل از اعمال الگوریتم DTW از یک تبدیل LMR برای انطباق بیشتر پارامترهای دو حالت متناظر از دو گوینده استفاده می شود. در روش GMM(3) به منظور ارائه تخمـین دقیـق تر سیگـنال گفـتاری و کیفـیت بالاتـر سیگنال تـبدیل شده، از مـدل GHM استفـاده می شودکه از کارکردی بهتر نسبت به روش قبل برخوردار است. در الگوریتمGMM(4) ویژگی های گفتار بطـور پیوسته با استفاده از همبستـگی میان ویژگی های گفتـار گوینده مبدا و مقصد، تغییر می نمایند. به منظور غلبه بر هموار شدگی طیفی ناشی از متوسط گیری آماری، از واحدهای آوایی نیمه هجا، به عنوان کوچکترین واحدهای آوایی شامل اطلاعات عروضی گفتار استفاده می شود. همچنین با توجه به مزایای GHM، از آن به عنوان آنالیز و سنتز کننده استفاده می شود.

در این رساله به ارائه روشی می پردازیم که علاوه بر ویژگی های درون قابی، از ویژگی های برون قابی (دینامیکی) برای یافتن بردار تبدیل یافته از گفتار گوینده A به گفتار گوینده B استفاده می کند. این روش مبتنی بر VQ بوده و در آن از یک ساختار شبکه برای یافتن یکی از بهترین مسیرها جهت نگاشت دنباله ای از قاب های گفتاری مربوط به کتاب کد گوینده A به کتاب کد گوینده B استفاده می شود. معیار بهینگی در یافتن مسیر عبارت است از: حفظ ویژگی های دینامیکی گفتار گوینده بعلاوه پیوستگی طیفی در گفتار تبدیل یافته.نوآوری دیگر ارائه شده، استفاده از نیمه هجا به عنوان کوچکترین واحد گفتاری در برگیرنده اطلاعات عروضی از گفتار گوینده است که متناسب با ساختار زبان فارسی می باشد. در نهایت به کمک اصلاحات مناسب دیگری که در روش یادگیری و طراحی تبدیلهای خطی مورد نیاز انجام شده است، به عملکرد بسیار مناسبی در تبدیل گفتار در مقایسه با روشهای رایج نائل شده ایم.

علائم اختصاری:

ABS Analysis By Synthesis**.**
ANN Artificial Neural Networks.
C Consonant.
CQ Communication Quality.
DFW Dynamic Frequency Warping.
DTW Dynamic Time Warping.
EM Expectation Maximazation.
GHM Generalized Harmonic Model.
GMM Gaussian Mixture Model.
HMM Hidden Markov Model.
HNM Harmonic plus Noise Model.
LAR Log Area Ratio.
LMR Linear Multivariate Regression.
LPC Linear Prediction Coding.
LP-PSOLA Linear Prediction Pitch Synchronous Overlap and Add.
LSF Line Spectral Frequency.
MAP Maximum a Posteriori.
MFCC Mel Frequency Cepstral Coefficient.
ML Maximum Liklihood.
MOS Mean Opinion Score.
MVF Maximum Voiced Frequency.
NCC Normalized Cross Correlation.
NUU Non Uniform Units.
OLA Overlap and Add.
PM Periodicity measure.
RMS Root Mean Square.
SD Spectral Distortion.
SM Sinusoidal Model.
SSM Stochastic Segment Modeling.
TD-PSOLA Time Domain Pitch Synchronous Overlap and Add.
TSVQ Trellis-Structured Vector Quantization.
TTS Text-to-Speech.
V Vowel**.**
VC Voice Conversion.
VQ Vector Quantization.
VT Voice Transformation.

فهرست عناوین:

فصل ۱.

معرفی و تشریح کاربرد سیستم های تبدیل گفتار و اهداف رساله

فصل ۲.

طراحی مجموعه گفتاری جهت آموزش و ارزیابی

فصل ۳.

طراحی یک مدل آنالیز/سنتز گفتار با کیفیت بالا

فصل ٤.

طراحی مجموعه توابع تبدیل اطلاعات طیفی وابسته به متن

فصل ۵.

سیستم تبدیل گفتار بر پایه TSVQ

فصل ٦. تایج نهایی، نوآوری ها و کارهای آینده

فهرست شکل ها

شماره صفحه

فهرست جداول

شماره صفحه

فصل ۱

معرفی و تشریح کاربرد سیستم های تبدیل گفتار و

اهداف رساله

۱-۱- پیش گفتار

گفتار، یکی از ساده ترین راه های برقراری ارتباط میان مردم می باشد. از زمان آغاز تحقیقات بر روی روش های ایجاد ارتباط میان انسان و ماشین، گفتار به عنوان یکی از مطلوب ترین راه های تعامل با کامپیوتر بوده است. در نتیجه بازشناسی گفتار و مبدل متن به گفتار به عنوان موضوعات مهم در ایجاد ارتباط موثر میان ماشین و انسان مورد مطالعه قرار گرفته است.

به منظور طبیعی تر شدن محاوره میان انسان و ماشین، بایستی تمامی جنبه های موثر در کیفیت گفتار ماشینی درنظر گرفته شوند. لذا گفتار حاصله علاوه بر داشتن مفهوم و تمامی اطلاعات، بایستی اطلاعاتی در مورد احساسات و هویت گوینده را منتقل سازد.

هویت گوینده یا به عبارت دیگر، صدای فرد، به عنوان یک ویژگی کلیدی در محاورات شفاهی می باشد. هویت گوینده امکان تمایز میان گویندگان را در یک کنفرانس تلفنی، برنامه رادیویی و ... برای ما فراهم می سازد. از طرفی هنگام استفاده از آن در سیستم های محاوره کامپیوتری اغلب افراد خواستار محاوره با کامپیوتر با صدای یک فرد خاص هستند. تا جایی که گاهی این سیستم ها بدلیل مورد پذیرش قرار نگرفتن صدای مورد استفاده برای کامپیوتر مورد رغبت دیگران قرار نمی گیرد. در سالهای اخیر از فن آوری تبدیل گفتار به منظور روشی برای کنترل هویت صوتی[1] صدای طبیعی و یا گفتار سنتز شده استفاده می شود.

در این رساله به مطالعه و طراحی سیستم تبدیل گفتار با کیفیت بالا برای زبان فارسی می پردازیم. در بخش مقدمه، مفهوم تبدیل گفتار تعریف می شود. در فصل ۱-۲ کاربردهای این تکنولوژی مورد بررسی قرار خواهد گرفت. در فصل ۱-۳ اجزای اصلی سیستم بررسی می گردد و در ادامه اهداف مدنظر رساله تشریح می گردد. در بخش آخر نیز مروری بر فصول رساله صورت می گیرد.

---

[1] - Voice Identity

۱-۲- تعریف سیستم های تبدیل گفتار

سیستم های تبدیل گفتار، صدای یک گوینده(گوینده مبدا) را به گونه ای تغییر می دهند که گویی یک گوینده دیگر(گوینده هدف) آن را بیان کرده است. بنابراین با فرض داشتن دو گوینده، هدف از بکارگیری یک سیستم تبدیل گفتار عبارت است از: تعیین یک تابع تبدیل به قسمی که گفتار گوینده مبدا چنان به نظر برسد که توسط گوینده مقصد گفته شده باشد. سیگنال گفتار حاوی انواع مختلفی از اطلاعات از قبیل: محتوای زبانی[۱]، هویت گوینده و نویز محیطی[۲] است. زمینه های متعددی از فن آوری گفتار بر روی هر کدام از این انواع اطلاعات متمرکز می باشند. تمرکز سیستم های تبدیل گفتار بر روی هویت گوینده است.

۱-۳- کاربردهای تبدیل گفتار

VC[۳] اساسا به عنوان روشی برای ایجاد گویندگان جدید در سیستم تبدیل متن به گفتار مورد استفاده قرار گرفته است. کاربردهای دیگر سیستم VC در حوزه های دیگری از قبیل ترجمه خودکار گفتار به گفتار، آموزش، مددرسانی پزشکی و سرگرمی مشاهده شده است.

۱-۳-۱-کاربرد در سیستم های TTS[۴]:

کیفیت سنتز کننده متن به گفتار با بهره گیری از مجموعه گفتاری بزرگ و همچنین فن آوری انتخاب واحد[۵] افزایش می یابد. این دسته از سیستم ها غالبا با نام TTSبر پایه گفتار[۶]، قادر به تولید گفتار سنتز شده با انتخاب مناسب ترین دنباله واحدهای آکوستیکی از یک پایگاه گفتاری[۷] وابسته به گوینده و سپس بهره گیری از یک استراتژی هموارسازی[۸] در مکان اتصال واحدهای انتخاب شده

---

[1] - Linguistic Content
[2] - Environmental Noise
[3] - Voice Conversion
[4] - Text To Speech
[5] - Unit Selection
[6] - Corpus-Based TTS
[7] - Database
[8] - Smoothing Strategy

می باشد. بنابراین به منظور سنتز دسته های دیگر مانند گفتار احساسی[۱]، شعر[۲] و یا گفتار با صدای گویندگان متعدد بایستی نمونه های گفتار نماینده از قبل ضبط شده باشند. همچنین کیفیت بالای گفتار سنتز شده نیازمند در اختیار داشتن یک مجموعه گفتاری با حجم زیاد می باشد. از طرفی ضبط گفتار و پردازش اطلاعات برای یک TTS مستلزم صرف هزینه و زمان زیادی می باشد. VC به عنوان یک راه حل، وسیله ای سریع و ارزان برای تولید صداهای جدید برای یک TTS می باشد. با استفاده از VC امکان خواندن نامه های الکترونیکی و SMS[۳] با صدای فرستنده بدست می آید. همچنین امکان استفاده از صدای خود و یا سایر دوستان به شخصیت های بازی رایانه ای و یا برای تخصیص صداهای مختلف به کاربردهای مختلف کامپیوتر فراهم می گردد. همچنین VC می تواند با تغییر و اصلاح ویژگی های عروضی[۴] گفتار روی یک جمله طبیعی، آن را به یک گفتار احساسی تبدیل نماید.

۱-۳-۲-ترجمه خودکار گفتار به گفتار[۵]:

در کاربرد ترجمه برخط تلفنی، به این فناوری برای مواردی که نیازمند قابلیت تشخیص هویت گوینده در گفتار تبدیل شده توسط شنوندگان هستیم، بسیار ضروری است. برای مثال، در یک تماس تلفنی با بیش از دو نفر، توانایی شناسایی گویندگان مختلف توسط صدایشان بسیار مهم می باشد.

۱-۳-۳-آموزش[۶]:

هنگام یادگیری یک زبان خارجی، تلفظ صحیح جملات و ادای واج هایی که در زبان مادری وجود ندارد یکی از مشکل ترین موارد برای دانش آموزان است[۳۶] و [۳۷]. استفاده از VC می تواند برای یادگیری زبان موثر واقع شود، خصوصا در تلفظ تمرینات، هنگامی که دانش آموز به تلفظ زبان خارجی با صدای خودش گوش خواهد داد.

---

[1] - Emotional
[2] - Expressive
[3] - Short Message Service
[4] - Prosodic Modification
[5] - Automatic Speech to Speech Translation
[6] - Education

۱-۳-٤-یاری رسانی پزشکی[1]:

حوزه دیگر کاربرد VC در بازسازی صدای افرادی است که صدای آنها دچار عیب شده است[۲۷]. در این صورت از سیستم تبدیل گفتار برای بهبود قابلیت فهم گفتار غیر عادی تلفظ شده بوسیله گوینده ای که دستگاه صوتی او دچار مشکل شده است می توان استفاده کرد.

کاربرد دیگر سیستم VC در بازگرداندن قدرت شنوایی برای افرادی است که دچار اختلال شنوایی در بازه های مختلف فرکانسی می باشند. در این صورت به کمک فن آوری VC می توان با تغییر بازه فرکانسی قدرت بازشناسی فرد معلول را افزایش داد.

۱-۳-۵-سرگرمی[2]:

یکی از کاربردهای پر طرفدار VC در حوزه سرگرمی ، استفاده از نرم افزارهایی است که امکان خواندن یک متن با صدای دیگر افراد را فراهم می آورد. کاربرد دیگر این سیستم در صداگذاری بر روی فیلم ها و یا تغییر متن در محتوای فیلم است[۳۰]. در این کاربرد با بکارگیری تنها چند صداگذار، قادر به تولید صدای هنرپیشگان با زبان های مختلف می باشیم. کاربرد دیگر صداگذاری، بازگرداندن صدای بازیگرانی است که کیفیت صدای آنها بدلیل کهولت سن از دست رفته است. تلفیق سیستم VC با فن آوری پویا نمایی ۳ بعدی چهره[3]، امکان بوجود آوردن شخصیت های مجازی با ویژگی های مطلوب صوتی را برای کاربردهای گوناگون مانند بازی های ویدیویی فراهم می آورد.

بجز موارد نامبرده، تبدیل گفتار را می توان در بسیاری از کاربردهای دیگر مورد استفاده قرار داد. به عنوان مثال در کدینگ گفتار با پهنای خیلی کم[4]، گفتار ارسال شده بدون اطلاعات مربوط به هویت گوینده بوده و این اطلاعات در مرحله رمزگشایی در مقصد، به گفتار اضافه می گردد.

---

[1] - Medical Aids
[2] - Entertainment
[3] - Facial Animation
[4] - Very Low Bandwidth Speech Encoding

علاوه بر این، VC با در اختیار گذاردن اطلاعات سطح بالا در خصوص هویت گوینده، موجب افزایش کارآیی فن آوری های دیگر در حوزه گفتار مانند بازشناسی گوینده[۱] و بازشناسی گفتار[۲] می گردد.

۱-٤- اجزای سیستم تبدیل گفتار

هر سیستم تبدیل گفتار دارای دو بخش مجزا می باشد که عبارتند از: مرحله آموزش توابع تبدیل مناسب برای تغییر اصلاح ویژگی های طیفی و عروضی و همچنین مرحله تبدیل گفتار گوینده.

در مراحل آموزش و تبدیل نیازمند بخش های زیر می باشیم:

۱- تهیه پایگاه گفتاری مناسب برای فرآیند تبدیل گفتار که در آن می توان از انواع مختلف دادگان از قبیل کلمات، هجاها[۳]، نیمه هجا و یا آواهای موجود در یک زبان استفاده کرد.

۲- بهینه سازی و توسعه یک روش آنالیز/سنتز مناسب جهت تولید گفتار با کیفیت بالا

۳- استفاده از یک روش همترازی زمانی[۴] که اطلاعات طیفی و عروضی متناظر از پایگاه گفتاری دو گوینده را با حداکثر تطبیق زمانی همتراز نماید.

۴- تکنیک های طراحی تبدیل های مناسب اطلاعات طیفی و پروزدیک

در مرحله آموزش، نمونه های گفتار و اطلاعات وابسته به آن برای هر دو گوینده مبدا و مقصد مورد تحلیل قرار گرفته و ویژگی های متناظر با یکدیگر همتراز می شوند و از این اطلاعات برای آموزش توابع تبدیل استفاده می گردد. بنابراین، برای هر زوج جدید مبدا و مقصد نیازمند اجرای یک آموزش جدید روی مجموعه گفتاری می باشیم. بلوک دیاگرام مرحله آموزش سیستم تبدیل گفتار در شکل زیر نشان داده شده است.

---

[1] - Speaker Recognition
[2] - Speech Recognition
[3] - Syllables
[4] - Time Alignment