دانشگاه صنعتی نوشیروانی بابل

جهت اخذ درجه کارشناسی ارشد

رشته مهندسی برق-گرایش الکترونیک

موضوع:

جداسازی صوت تک کانال با روش Bayesian

استاد راهنما:

دکتر محمدرضا کرمی

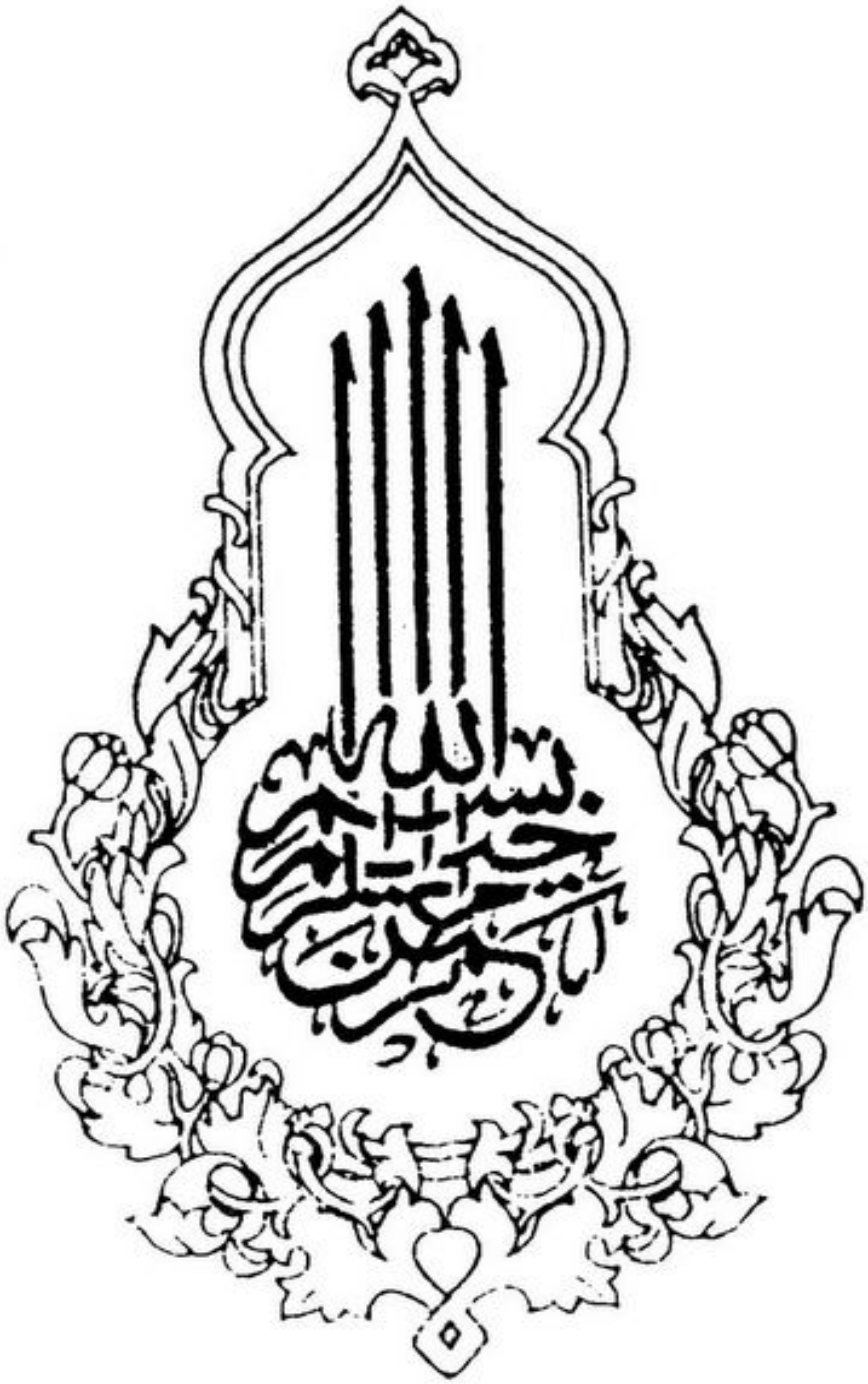
استاد مشاور:

دکتر محمدرضا ذهابی

نگارش:

سونای کمی

تابستان ۹۰


بسم الله الرحمن الرحيم

دانشگاه صنعتی نوشیروانی بابل

جهت اخذ درجه کارشناسی ارشد

رشته مهندسی برق-گرایش الکترونیک

موضوع:

جداسازی صوت تک کانال با روش Bayesian

استاد راهنما:

دکتر محمدرضا کرمی

استاد مشاور:

دکتر محمدرضا ذهابی

نگارش:

سونای کمی

تابستان ۹۰

سپاسگزاری

بر خود لازم می دانم از زحمات و راهنمایی های بی دریغ استاد راهنمای گرانقدر، جناب آقای دکتر محمدرضا کرمی که رهنمود های عالمانه ایشان من را در این راه یاری نموده نهایت تشکر و قدردانی را داشته باشم.

همچنین از جناب آقای دکتر محمدرضا ذهابی که مشاوره این پایان نامه را به عهده داشته اند کمال تشکر را دارم.

تقدیم به:

روح پاک پدرم که عالمانه به من آموخت تا چگونه در عرصه زندگی،

ایستادگی را تجربه نمایم

و به مادرم، دریای بی کران فداکاری و عشق که وجودم برایش همه

رنج بود و وجودش برایم همه مهر

چکیده

سیگنال های گفتار به ندرت به صورت خالص برای کاربرد های پردازش گفتار موجود می باشند و اغلب به وسیله‌ی تداخل آوایی نظیر نویز پس زمینه، اعوجاج، سیگنال گفتار گوینده‌ی دیگر و ... مخدوش می شوند. در چنین حالتی لازم است که ابتدا سیگنال گفتار از پس زمینه جدا شود. به ویژه عمل جداسازی گفتار چند گوینده که به عنوان جداسازی گفتار شناخته می شود امری چالش برانگیز است زیرا شامل جداسازی سیگنال هایی است که دارای مشخصات آوایی و آماری بسیار مشابهی هستند. چالش برانگیز ترین حالت جداسازی گفتار، جداسازی سیگنال های گفتار ناشی از رکورد تک کاناله است؛ مسئله ای که معمولاً به عنوان جداسازی گفتار تک کانال شناخته می شود. سیستم های جداسازی گفتار تک کانال با کیفیت بالا هنوز از اهمیت زیادی برخوردارند و به عنوان یک مسئله‌ی حل نشده باقی مانده اند. هدف این پایان نامه، ارائه‌ی راهکار هایی برای بهبود سیستم های جداسازی گفتار تک کانال می باشد. یک سیستم مؤثر جداسازی گفتار تک کانال، پیش پردازشی مهم در کاربرد های بسیاری نظیر شناسایی اتوماتیک گفتار و تعیین هویت گوینده به شمار می رود. این پایان نامه روی جداسازی مبتنی بر مدل گفتار تک کانال متمرکز میشود که به تکنیک هایی اطلاق می شود که از مدل های آموزش داده شده‌ی منابع استفاده می کنند تا منابع را از رکورد تک کاناله‌ی ترکیب خطی آنها جدا کنند. اولین راهکار پیشنهادی در این پایان نامه جداسازی منابع در سطح زیربخش و اعمال این رویکرد به روش مبتنی بر تخمین حداکثر احتمال پسین و روش مبتنی بر کوانتیزاسیون بردار می باشد. در راهکار پیشنهادی دوم، روش مبتنی بر تخمین حداکثر احتمال پسین برای حالتی که در آن سیگنال های تست و آموزشی سطح انرژی مختلفی نسبت به هم دارند (منابع بهره مختلفی دارند) بسط داده می شود. این حالت در بسیاری از روش های مبتنی بر مدل جداسازی گفتار تک کانال در نظر گرفته نشده است. روش هایی که این حالت را در نظر می گیرند به روش های وفق بهره موسوم اند. اگرچه روش های وفق بهره‌ی موجود در تخمین بهره‌ی منابع موفق عمل می کنند اما از پیچیدگی محاسباتی و زمان پردازش بالایی برخوردارند. روش پیشنهادی سوم بر مبنای جداسازی زیربخشی و کوانتیزاسیون بردار تلاشی برای غلبه بر این محدودیت است. نتایج تجربی نشان می دهند که راهکار های پیشنهادی، عملکرد سیستم جداسازی گفتار تک کانال را بهبود می دهند.

واژه های کلیدی:

جداسازی گفتار تک کانال، مدل ترکیبی گوسی، کوانتیزاسیون بردار، تخمین حداکثر احتمال پسین، تخمین حداقل میانگین مربعات خطا.

عنوان	فهرست مطالب	صفحه

عنوان	فهرست شکل ها	صفحه

ACF	تابع خود همبستگی (Autocorrelation Function)
AM	مدولاسیون دامنه (Amplitude Modulation)
ASA	تحلیل صحنه شنوایی (Auditory Scene Analysis)
ASR	شناسایی اتوماتیک گفتار (Automatic Speech Recognition)
BSS	جداسازی کور منابع (Blind Source Separation)
CASA	تحلیل محاسباتی صحنه شنوایی (Computational Auditory Scene Analysis)
CSM	مدل منبع مرکب (Composite Source Modeling)
EM	بیشینه سازی امید (Expectation Maximization)
GMM	مدل ترکیبی گوسی (Gaussian Mixture Model)
HMM	مدل مخفی مارکوف (Hidden Markov Model)
ICA	تحلیل مؤلفه های مستقل (Independent Component Analysis)
IDFT	تبدیل فوریه معکوس گسسته (Inverse Discrete Fourier Transform)
MAP	حداکثر احتمال پسین (Maximum A Posteriori)
ML	حداکثر شباهت (Maximum Likelihood)
MMSE	حداقل میانگین مربعات خطا (Minimum Mean Squared Error)
MSE	میانگین مربعات خطا (Mean Squared Error)
OLA	افزودن همپوشانی (Overlap Add)
PDF	تابع چگالی احتمال (Probability Density Function)
SACF	تابع خود همبستگی اختصاری (Summary Autocorrelation Function)
SDR	نسبت سیگنال به اعوجاج (Signal to Distortion Ratio)
SE	بهبود گفتار (Speech Enhancement)
SID	تعیین هویت گوینده (Speaker Identification)
SNR	نسبت سیگنال به نویز (Signal to Noise Ratio)
SPL	سطح شدت صدا (Sound Pressure Level)
SRR	نسبت سیگنال به مانده (Signal to Residual Ratio)
TIR	نسبت هدف به تداخل (Target to Interference Ratio)
VQ	کوانتیزاسیون بردار (Vector Quantization)

فصل اول
مقدمه

۱- مقدمه

۱-۱- انگیزه پژوهش

در بسیاری از کاربرد های پردازش گفتار نظیر شناسایی اتوماتیک گفتار[۱] (ASR)، تعیین هویت گوینده[۲] (SID) و بهبود کیفیت سیگنال گفتار[۳] (SE)، سیگنال ورودی اغلب به نویز آوایی محیطی آغشته می شود که این عمل سبب کاهش کیفیت سیگنال گفتار و در نتیجه باعث کاهش کارایی کلی الگوریتم پردازش گفتار می شود. هنگامی که تداخل آوایی، شامل سیگنال های صحبت گوینده های دیگر باشد (معمولاً به عنوان اثر مهمانی شلوغ[۴] شناخته می شود)، آنگاه به علت مشابهت در ماهیت سیگنال های خواسته و ناخواسته، این کاهش کارایی بیشتر می شود. بنابراین در چنین شرایطی یک الگوریتم جداسازی صوت، پیش پردازشی ضروری به شمار می رود تا کیفیت سیگنال گفتار را برای پردازش بعدی بهبود ببخشد. اگر ما بتوانیم سیگنال گفتار مورد نظر را قبل از پردازش آن از سیگنال اولیه جدا کنیم، این عمل در بهبود کارایی کلی الگوریتم پردازش گفتار کمک شایانی خواهد کرد.

در بعضی از شرایط عملی که تنها یک کانال موجود است تکنیک های جداسازی تک کانال باید مورد استفاده قرار گیرد. این حالت ممکن است به وسیله‌ی سیستم مورد استفاده (مانند کاربردهای مبتنی بر تلفن) یا به علت فراهم بودن سیگنال مورد نظر (مانند کاربردهای از پیش ضبط شده) اعمال شود. این شرایط به علت سهولت در نصب میکروفن جالب هستند اما محدودیت اصلی روش های تک کانال این است که هیچ سیگنال مرجعی برای تداخل[۵] موجود نیست. بنابراین چگالی طیف قدرت[۶] سیگنال تداخلی باید تنها بر اساس سیگنال گفتار تک کانال موجود تخمین زده شود و این چیزی است که آن را به امری چالش برانگیز تبدیل می کند. این مسئله معمولاً به مسئله‌ی جداسازی صوت تک کانال[۷] اطلاق می شود.

ایده‌ی جداسازی صوت تک کانال، پردازش اتوماتیک سیگنال ترکیبی به منظور بازیابی سیگنال گفتار اولیه‌ی هر گوینده است. حداقل کردن آرتیفکت در سیگنال گفتار پردازش شده امری کلیدی است به ویژه اگر هدف نهایی استفاده از گفتار بازیابی شده در کاربردهای مبتنی بر ماشین نظیر سیستم های شناسایی اتوماتیک گفتار و تشخیص گوینده باشد. بنابراین هدف الگوریتم جداسازی صوت تک کانال عبارت است از:

---

[1] Automatic speech recognition
[2] Speaker identification
[3] Speech enhancement
[4] Cocktail party effect
[5] Interference
[6] Power spectral density
[7] Monaural speech separation problem

* بهبود جنبه های ادراکی سیگنال گفتار مخدوش شده

* بهبود عملکرد سیستم نهایی پردازش گفتار

* افزایش توانمندی سیستم های پردازش گفتار مبتنی برماشین

اگرچه کاملا واضح است که سیستم شنوایی انسان در تمرکز روی یک یا چند گوینده‌ی خاص در میان جمعی از گویندگان که همزمان در حال صحبت هستند بسیار تواناست [۱]، الگوریتم های کامپیوتری که برای انجام همان کار طراحی شده اند درجه‌ی محدودی از موفقیت را نشان داده اند.

۱-۲- هدف پایان نامه

هدف این پایان نامه ارائه‌ی راهکار هایی برای بهبود جداسازی گفتار تک کانال می باشد. در میان رویکرد های موجود برای حل مسئله‌ی جداسازی گفتار تک کانال، رویکرد های مبتنی بر مدل[1] توانایی بیشتری از خود نشان داده اند به طوریکه در چند سال گذشته مورد توجه محققین قرار گرفته اند. در این پایان نامه ما بر روی رویکرد های مبتنی بر مدل متمرکز می شویم. راهکار های ارائه شده در این پایان نامه به صورت زیر می باشد

۱- بهبود جداسازی گفتار تک کانال با استفاده از جداسازی زیربخشی

۲- وفقی کردن بهره[2] در روش تخمین حداکثر احتمال پسین

۳- ارائه‌ی روشی مؤثر و سریع بر مبنای جداسازی زیربخشی و کوانتیزاسیون بردار[3] برای جبران تفاوت بهره[4] سیگنال های ورودی

۱-۳- ساختار پایان نامه

ادامه‌ی این پایان نامه به صورت زیر سازمان دهی شده است

فصل دوم مسئله‌ی جداسازی گفتار تک کانال را معرفی می کند و به مرور سریع رویکرد های مختلفی می پردازد که برای حل این مسئله ارائه شده اند.

[1] Model based approaches
[2] Gain adaptation
[3] Vector quantization
[4] Gain

فصل سوم به تفصیل به بررسی برخی از روش های مبتنی بر مدل که برای حل مسئله‌ی جداسازی گفتار تک کانال ارائه شده اند می پردازد. به کمک این روش ها با شیوه های مختلف مدل سازی[1] و تخمین آشنا می شویم. در نهایت، روش های پیشنهادی با روش های ارائه شده در این فصل مقایسه خواهند شد.

در فصل چهارم راهکار هایی برای بهبود جداسازی گفتار تک کانال ارائه می شود. این راهکارها شامل جداسازی زیربخشی، وفقی کردن بهره در روش تخمین حداکثر احتمال پسین و ارائه‌ی یک روش وفق بهره بر مبنای جداسازی زیربخشی می باشد.

[1] Modeling

فصل دوم
پیش زمینه و مروری بر کارهای گذشته

۲- پیش زمینه و مروری بر کار های گذشته

۲-۱- مقدمه

صوت تک کانال هنگامی اتفاق می افتد که سیگنال های گفتار مربوط به چند گوینده از طریق یک کانال با هم ترکیب می شوند. فرایند استخراج یا جداسازی سیگنال گفتار مورد نظر از سیگنال ترکیبی معمولاً به عنوان جداسازی صوت تک کانال شناخته می شود. استفاده از یک الگوریتم جداسازی صوت تک کانال به عنوان پیش پردازش در سیستم های پردازش گفتار می تواند در بهبود سیگنال گفتار هدف[1] در بسیاری از کاربردها تاثیر زیادی داشته باشد. این کاربردها شامل تکنیک های شناسایی اتوماتیک گفتار (ASR)، تعیین هویت گوینده (SID) و بهبود کیفیت گفتار (SE) هستند. به طور قابل ملاحظه ای، سیستم شنوایی انسان می تواند مسئله جداسازی صوت تک کانال را با دقت بیشتری نسبت به سیستم های جداسازی مبتنی بر ماشین فعلی حل کند. از لحاظ تاریخی، سیستم های جداسازی صوت تک کانال با استفاده از الگوریتم هایی توسعه یافته اند که سیگنال های تداخلی را حذف می کنند، سیگنال هدف را بهبود می دهند، یا هر دو سیگنال را به طور همزمان تخمین می زنند. تاکنون، تکنیکی که به طور رضایت بخش در تمام حالات کار کند وجود نداشته است. در این فصل دورنمای مختصری از مسئله‌ی جداسازی صوت تک کانال و مروری بر کارهای انجام شده برای حل این مسئله ارائه می شود به این ترتیب که بخش ۲-۲ به توضیح درباره‌ی این مسئله و کاربردهای اصلی آن می پردازد و مروری بر رویکردهای گذشته برای حل این مسئله در بخش ۲-۳ ارائه می شود.

۲-۲- مسئله‌ی جداسازی صوت تک کانال

صوت تک کانال به صورت سیگنال ترکیبی حاصل از دو یا چند گوینده تعریف می شود. این پدیده معمولا به علت ترکیب سیگنال های گفتار منابع مستقل و همزمان به صورت یک سیگنال منفرد در گیرنده (میکروفن منفرد) اتفاق می افتد. شرایط متداولی که صوت تک کانال ممکن است اتفاق بیافتد می توان به صورت زیر خلاصه کرد:

۱- هنگام ضبط صدای دو نفر که به طور همزمان نسبت به یک میکروفن در حال صحبت هستند. یک مثال در این مورد، سیگنال های گفتار ضبط شده در جعبه‌ی ضبط صدا در کابین خلبان می باشد.

۲- هنگامی که سیگنال گفتار هدف همزمان با صداهای پس زمینه‌ی شنیده شدنی گوینده های دیگر به سمعک افراد کم شنوا می رسد.

[1] Target speech signal

۳- هنگامی که به علت انتقال سیگنال گفتار از طریق یک کانال ارتباطی معیوب، تداخل صحبت[1] اتفاق می افتد.

سیستم شنوایی انسان در تمرکز روی یک یا چند گوینده‌ی خاص در جمعی از گویندگان (معمولا به عنوان اثر مهمانی شلوغ شناخته می شود) توانایی های قابل ملاحظه ای از خود نشان می دهد. به طرز شگفت آوری، شنوندگان انسانی حتی هنگامی که فقط از یک گوش استفاده می کنند یا با دو گوش به صدای ضبط شده از طریق یک کانال منفرد[2] گوش می دهند قادر به انجام این کار هستند. از طرف دیگر، الگوریتم های کامپیوتری طراحی شده برای انجام این کار، درجه‌ی محدودی از موفقیت را از خود نشان داده اند. آزمایش های میلر [۲] و بروکس و نوتبوم [۳] نشان داده اند که شنوندگان انسانی می توانند به درجه‌ی بالایی از جداسازی صداهای ترکیب شده از طریق یک کانال برسند. در حقیقت، انسان شامل فاکتورهای زیادی برای جداسازی صداهای هم زمان است که ممکن است برای ماشین موجود نباشد. بعضی از این فاکتورها همانطور که به وسیله ی چری [۴] بیان شده اند عبارتند از:

۱- اطلاعات فضایی درباره‌ی جهت منبع صوتی

۲- اطلاعات بصری نظیر اشارات در هنگام سخن گفتن و حرکت لب

۳- اطلاعات درباره‌ی آهنگ صدا (براساس سرعت گفتار، جنس گوینده و...)

۴- اطلاعات درباره لهجه ها

۵- اطلاعات درباره‌ی احتمالات گذرا[3] (بر اساس درک گفتار، علم نحو و ...)

به جز آخرین فاکتور، فاکتورهای دیگر می توانند در حالتی از صوت تک کانال که، به عنوان مثال، دو پیام از یک گوینده به طور همزمان روی یک نوار ضبط می شوند حذف شوند. با این وجود، گوش انسان هنوز قادر به جداسازی گفتار در چنین حالت حدی است و علت آن حافظه‌ی وسیع درباره‌ی احتمالات گذرا است که به انسان اجازه می دهد تا دنباله‌ی کلمات را پیشگویی کند.

هدف جداسازی صوت تک کانال بازیابی یک یا هر دو سیگنال گفتار از سیگنال ترکیبی است. مدل دو گوینده ای مسئله‌ی جداسازی صوت تک کانال که مدل مورد بررسی در این پایان نامه است در شکل (۲-۱)

[1] Cross talk
[2] Single channel
[3] Transition probabilities

نشان داده شده است. همانطور که از شکل پیداست، سیگنال گفتار تک کانال ($z(t)$) مجموع سیگنال گفتار هدف ($x(t)$) و سیگنال گفتار تداخل ($y(t)$) است.

$$z(t) = x(t) + y(t) \qquad , \quad t = 1,2, \dots, T \tag{۱-۲}$$

که t نشان دهنده‌ی نمونه در حوزه‌ی زمان است. نسبت هدف به تداخل[1] ورودی (TIR) به صورت نسبت توان $x(t)$ به $y(t)$ تعریف می شود.

$$\text{TIR [dB]} = 10 \log_{10} \frac{\sum_t x^2(t)}{\sum_t y^2(t)} \tag{۲-۲}$$

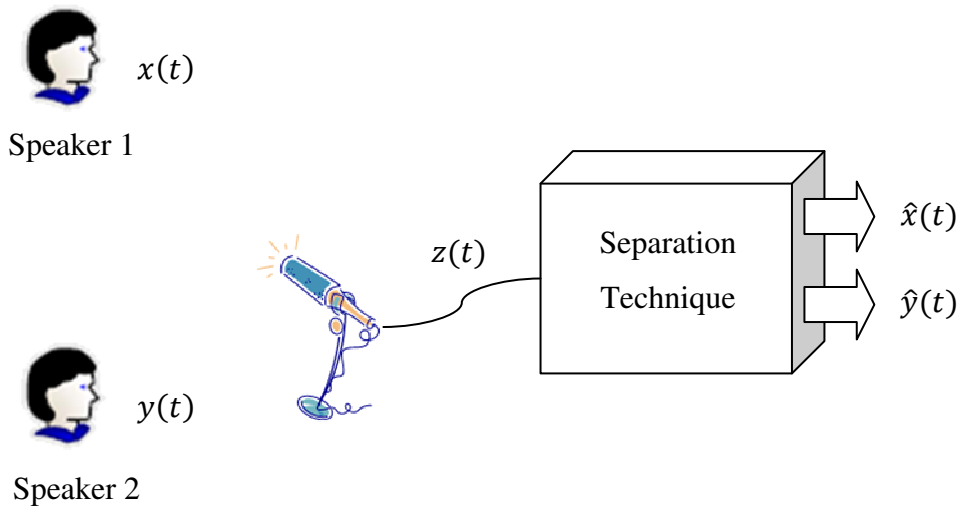
در زمینه‌ی پردازش سیگنال می توان به جداسازی صوت تک کانال از چهار منظر مختلف نگریست

۱- بهبود سیگنال گفتار هدف

۲- استخراج سیگنال گفتار هدف

۳- حذف سیگنال گفتار تداخلی

۴- تخمین هر دو سیگنال گفتار



شکل ۲-۱- مسئله‌ی جداسازی صوت تک کانال

[1] Target to interference ratio