استخراج بازنمائی‌های مقاوم مبتنی بر مدل فیزیولوژی گوش انسان برای سیستم‌های بازشناسی خودکار گفتار

امین فاضل دهکردی[1] afazel@mehr.sharif.edu
سید کمال‌الدین غیاثی[1] ghiathi@mehr.sharif.edu
خسرو حسین‌زاده[1] hoseinzadeh@ce.sharif.edu
کارو لوکس[2] lucas@ece.ut.ac.ir

[1] دانشکدهٔ کامپیوتر دانشگاه صنعتی شریف
[2] دانشکدهٔ برق و کامپیوتر دانشگاه تهران

چکیده: همیشه موجودات زنده بهترین الگوها برای تکولوژی بشر بوده‌اند. بر خلاف سالهای متمادی تحقیق و مطالعه، هنوز به طور واضح معلوم نیست که سیستم شنوائی انسان به چه شکل با دامنهٔ وسیعی از اصوات عمل می‌کند[1] و همچنین مشخص نیست که انسان به چه شکل گفتار را در حضور صداهای دیگر می‌تواند بفهمد[2]. یک سیستم بازشناسی گفتار که بتواند به شکل موثری کارائی سیستم شنوائی انسان را داشته باشد باید تضمین کند که اطلاعات پردازش شدهٔ گفتار دارای بازنمائی‌هائی شبیه به سیستم شنوائی انسان می‌باشد. معمولاً در بسیاری از سیستمهای بازشناسی گفتار امروزی بازنمائی‌های استخراج شده بیشتر بر اساس سیستم تولید گفتار می‌باشد و به نظر می‌رسد که مدل کردن فیزیولوژی گوش انسان می‌تواند در اسخراج ویژگی‌های مؤثرتر مفید باشد. در این مقاله با الهام گرفتن از فیزیولوژی گوش انسان قسمتی از گوش داخلی انسان مدل شده است.نتایج حاصل از آزمایشات بر روی دادگان اعداد متصل انگلیسی نشان دهندهٔ آن است که استفاده از استخراج ویژگی‌های مبتنی بر یافته‌های فیزیولوژیک گوش انسان می‌تواند دقت سیستم‌های بازشناسی گفتار را به خصوص در شرایط نویزی بالاتر ببرد.

کلمات کلیدی: بازشناسی مقاوم گفتار، بازنمائی‌های مقاوم، فیزیولوژی گوش انسان.

۱- مقدمه

بازشناسی خودکار گفتار یکی از تکنولوژی‌های در حال ظهور برای افزایش رابطهٔ بین انسان و ماشین می‌باشد. با وجود اینکه امروزه محصولاتی در این زمینه به وجود آمده است ولی هنوز استفادهٔ همه منظوره از این تکنولوژی عملی نشده است. مهمترین مشکل وجود عدم تطابق بین محیط آموزش و محیط آزمایش می‌باشد. همهٔ سیستمهای بازشناسی خودکار گفتار دارای یک طبقه‌بندی کننده می‌باشند که بر روی سیگنالهای گفتار (معمولاً فاقد نویز) آموزش داده می‌شود با این هدف که بتواند محتوای آوائی گفتار را برای بازشناسی گفتار(معمولاً نویزی) یاد بگیرد. محیط

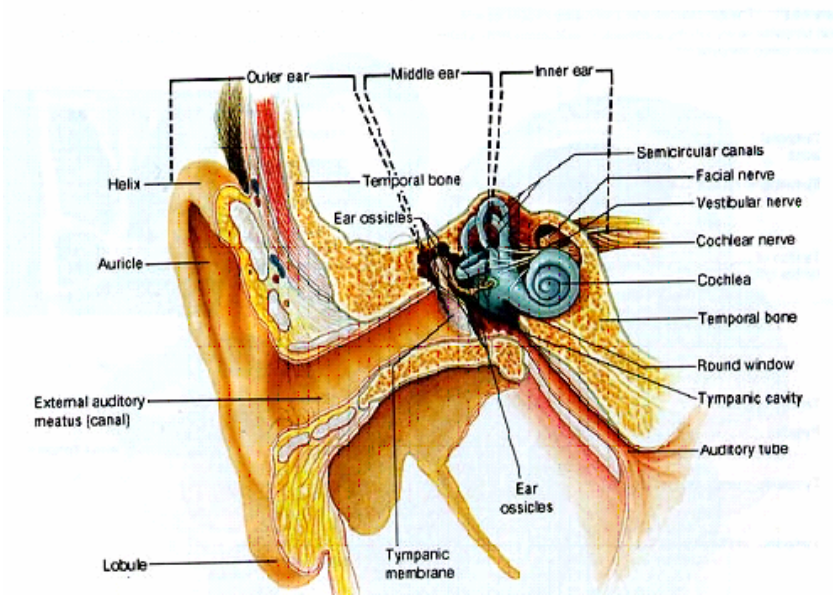
آزمایش ممکن است با محیط آموزش تفاوتهائی از نظر تنوع گوینده‌ها، نـویز پـس‌زمینـه و اثرهـای کانال انتقال گفتار داشته باشد.

همانطور که گفته شد مشکل بزرگ سیستمهای بازشناسی گفتار مقاوم نبودن آنها نسبت به نویز است که باعث شده نتوان بـه طـور فراگیـری از ایـن سیسـتمها اسـتفاده کـرد. از آنجـائی کـه بازشناسی گفتار در انسان در مقایسه با سیستمهای بازشناسی گفتار امروزی در مقابـل نـویز بسـیار مقاوم می‌باشد، به نظر می‌رسد که یکـی از راههـای مقـاوم کـردن ایـن سیسـتمها الهـام گـرفتن از بیولوژی موجود در طبیعت می‌باشد. بسیاری از سیستمهای بازشناسی از بازنمائی MFCC اسـتفاده می‌کنند[3] که یک الگوریتم مبتنی بر فیلتر بانک می‌باشد کـه فیلترهـای آن بـر روی فرکانسـهای لگاریتمی در دامنهٔ Mel قرار گرفته‌اند. از طرفی باندهای فرکانسی هر یک از این فیلترها به وسـیلهٔ مکان هر یک از آنها تعیین می‌گردند و از روی یافته‌های بیولوژیکی نمی‌باشند.

در این مقاله یک روش جدید بـا الهـام گـرفتن از گـوش انسـان و مـدل‌کـردن حرکتهـای مکانیکی غشای پایه[1] و سلولهای مژکدار[2] برای دریافت فرکانسهای مختلف صدا بر اساس یافته‌هـای فیزیولوژی ارائه می‌شود.

۲- فیزیولوژی سیستم شنوائی انسان

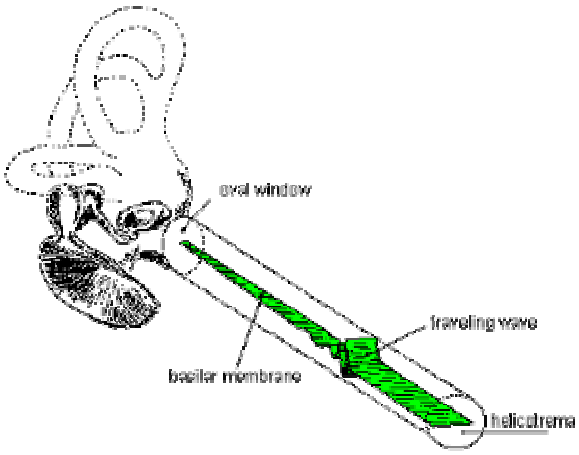
گوش اندام حسی تبدیل کنندهٔ صدا است. گوش انسان از سه بخش گوش بیرونـی، گـوش میانی و گوش درونی تشکیل شده است (شکل ۱). گوش بیرونی امواج صوتی را به پردهٔ صماخ، که در حد فاصل گوش بیرونی و گوش میانی است، راه می‌دهد. امواج صوتی پردهٔ صماخ را بـه ارتعـاش در می‌آورند. ارتعاشهای پردهٔ صماخ از طریق سه استخوان کوچک موجود در گوش میانی به گـوش درونی انتقال می‌یابند[4].



[1] Basilar Membrane
[2] Hair Cells

شکل (۱). ساختار کلی گوش

گوش درونی حفرهٔ پر از مایعی است که حلزونی گوش[1] نامیده می‌شود. حلزونی گوش انسان دارای یک ساختار مخروطی شکل است که فنروار در اطراف یک هستهٔ استخوانی پیچیده شده است. نیروی مکانیکی تقویت شده که از طرف گوش میانی به گوش داخلی فرستاده می‌شود بلافاصله توسط حلزونی گوش به فشار هیدرولیکی تبدیل می‌شود.این فشار هیدرولیکی، حرکت حاصل را به مجراء حلزونی گوش[2] و اندام کورتی[3] میرساند. در سراسر طول حلزون، نوعی غشاء امتداد یافته است که غشای پایه نامیده می‌شود. این غشاء پردهٔ خم‌پذیری است که هرگاه صدائی به پردهٔ صماخ برخورد کند آن را به ارتعاش در می‌آورد. ارتعاش غشای پایه به صورت جابجائی موجمانندی است که آن غشاء را تغییر شکل می‌دهد.

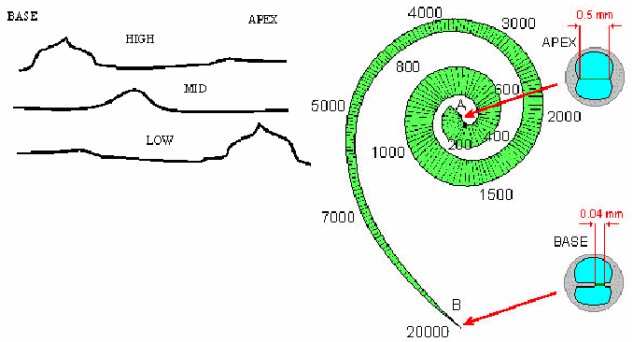


شکل (۲). حرکت موج جابجائی در در طول غشای پایه

هنگامی که موج جابجائی در طول غشای پایه پیش می‌رود (شکل ۲)، بر دامنهٔ آن افزوده می‌شود تا اینکه در نقطه‌ای به اوج برسد. اما وقتی که موج جابجائی غشای پایه از نقطهٔ اوج جلوتر می‌رود، دامنهٔ آن کاهش می‌یابد.محل اوج دامنهٔ موج جابجائی در غشای پایه، بستگی به فرکانس صوت دارد. اصوات با فرکانس بالا سبب می‌شوند که موج جابجائی در قاعدهٔ حلزون به اوج خود برسند در حالیکه اصوات با فرکانس پایین سبب می‌شوند که موج جابجایی در رأس حلزون به اوج خود برسد (شکل ۳). بنابراین، طرح مکانیکی غشای پایه چنان است که به اصوات دارای فرکانسهای مختلف با تغییر دادن شکل خود پاسخ می‌دهد[4]. تغییر شکل غشای پایه را سلولهای مژکدار به پتانسیل کار تبدیل می‌کنند. وقتی که غشای پایه مرتعش می‌شود و به اوج تغییر شکل خود می‌رسد، سلولهای مژکدار نسبت به جسم آن سلول خم می‌شوند و از این طریق عصب شنوائی،

[1] Cochlea
[2] Cochlea duct
[3] Organ of Corti

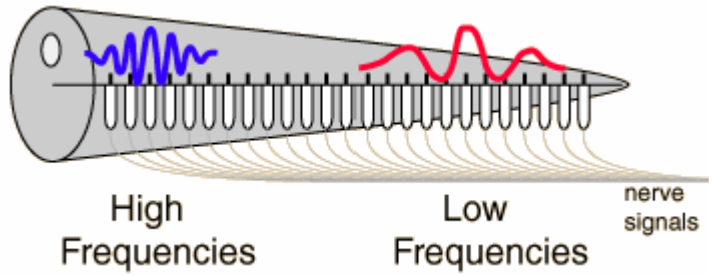
پتانسیل کار تولید می‌کند که به مغز منتقل می‌شود و اطلاعات مربوط به فرکانس صوت را به مغـز می‌رساند.



شکل(۳) حساسیت حلزون گوش نسبت به فرکانسهای مختلف

۳- مدلسازی حرکتهای مکانیکی غشاء پایه و سلولهای مژکدار

از آنجائی که غشاء پایه و سلولهای مژکدار با همکاری یکدیگر فرکانسهای مختلف صدا را بـه انرژی مکانیکی تبدیل می‌کنند با تقریب خوبی می‌توان هر قسمت از آن را بـه صـورت گسسـته بـه شکل یک دیاپازون در نظر گرفت. به این دیاپازون‌ها در واقع می‌توان به شکل سیستم‌هـای نوسـانی میرا که توسط یک نیروی خارجی تحریک می‌گردند، نگاه کرد.



شکل (۴) مدل کردن حرکتهای مکانیکی غشاء پایه و سلولهای مژکدار با دیاپازون

مدل هر یک از این سیستم‌های نوسانی میرا را می‌توان با رابطه(۱) نمـایش داد[5]، [6]، [7] و [8]،

$$\ddot{x} + \frac{b}{m}\dot{x} + \omega_0^{\,2} x = \frac{F}{m}\cos\omega t \qquad (۱)$$

که در این رابطه b نشان‌دهندهٔ فاکتور میرائی، ω_0 فرکانس طبیعی هر کدام از این دیاپازون‌ها و F نیروی خارجی است که توسط صوت با فرکانس ω بر هر یک ار آنها اعمال می‌شود. از آنجـائی کـه این مدلسازی باید به صورت گسسته انجام شود، در اینجا به تعداد ۲۵۶ عدد از ایـن مـدلها در نظـر گرفته شده است. از آنجایی که هر یک از این مدلها باید به فرکانس خاصی حساس باشـند و در آن فرکانس حالت تشدید داشته باشند، برای هر یک از آنها یک باند فرکانسی خاص در نظر گرفته شده است. این باند فرکانسی تعیین کننده مقدار b بر اساس رابطه زیر می‌باشند،

$$b = \frac{m\omega_0}{Q} \qquad (۲)$$

که در رابطهٔ فوق Q فاکتوری است که تعیین کنندهٔ میزان انرژی از دست رفته است و همچنین k را به عنوان ضریبی تعریف می‌کنیم که به وسیلهٔ رابطهٔ $k = m\omega_0^2$ به جنس هر یک از دیاپازون‌ها وابسته است.

برای استفاده از این مدل در استخراج بازنمائی‌ها باید به ازای تعدادی نمونهٔ صوت شتاب هریک از دیاپازون‌ها را به دست بیاوریم و از روی این شتاب و میزان انحراف آنها انرژی را برای تک‌تک آنها محاسبه کنیم. این انرژی به ازای تک‌تک دیاپازون‌ها می‌تواند به عنوان بردار بازنمائی برای سیستم بازشناسی گفتار مورد استفاده قرار بگیرد.

شتاب هر یک از این مدلها را بر اساس سرعت و مکان قبلی و همچنین نیروی اعمال شدهٔ جدید که از طرف نمونه‌های صوت بر آن‌ها وارد می‌شوند، طبق رابطه زیر بدست می‌آید

$$a = \frac{F - bv_{pr} - kx_{pr}}{m} \qquad (۳)$$

از روی این شتاب سرعت $v = v_{pr} + at$ و مکان $x = x_{pr} + at$ جدید به دست می‌آید. سپس انرژی از طریق رابطهٔ $E = \frac{1}{2}mv^2 + \frac{1}{2}kx^2$ به دست می‌آید و از آن برای استخراج بازنمائی‌ها استفاده می‌کنیم.

۴- نتایج حاصل از پیاده‌سازی

برای نشان دادن تأثیر بازنمائی‌های مبتنی بر مدل فیزیولوژیک گوش آزمایشهائی برای مقایسه با بازنمائی‌های MFCC طراحی شد. برای آزمایش از سیستم بازشناسی گفتار HTK[1] استفاده شد[9]. با استفاده از سیستم بازشناسی مذکور مدل‌سازی اعداد انگلیسی با استفاده از مدل مخفی مارکوف انجام داده شد. HMM های به کار برده شده از نوع چگالی پیوسته و با تلفیق‌های گوسی می‌باشند و پرش بین حالت‌های آن فقط به صورت چپ به راست مجاز می‌باشد. هر عدد زبان انگلیسی با یک HMM دارای ۱۶ حالت و ۳ تلفیق گوسی مدل شد. برای آموزش این سیستم از دادگان اعداد متصل زبان انگلیسی که شامل ۱۸ گوینده که هر گوینده ۷۷ رشته عددی را بیان می‌کند، استفاده شد. هنگام آزمایش از ۲۰۰ رشته عددی که به وسیلهٔ گوینده‌هائی غیر از گوینده‌های داده‌ای آزمون بیان شده‌اند، استفاده شد. دادگان تست به چهار دستهٔ ۵۰تائی تقسیم شدند و به هر دسته چهار نوع نویز مختلف با SNRهای متفاوت اضافه شد. در جدول (۱) خلاصهٔ نتایج حاصل از آزمایش‌ها آمده است. همانطور که در جدول مشخص است در اغلب موارد آزمایش بهبودهای در نتیجه حاصل از بازشناسی به دست آمد.

[1] Hidden Markov Toolkit

Word Error Rate								
Noise	Feature	20 db	15 db	10 db	5 db	0 db	-5 db	Overall
Car	MFCC	13.29	24.05	53.16	80.38	93.67	93.67	59.7
	Bio Model	13.29	23.42	55.06	74.05	82.28	93.67	56.96
Exhibition	MFCC	37.75	48.34	59.6	77.48	90.73	91.39	67.55
	Bio Model	33.77	36.42	64.24	78.15	86.09	89.4	64.68
Babble	MFCC	26.95	44.31	72.46	88.02	94.61	99.4	70.96
	Bio Model	23.95	34.73	63.47	76.05	83.23	88.62	61.68
Subway	MFCC	46.75	60.36	71.01	88.17	89.94	90.53	74.46
	Bio Model	47.93	57.4	72.78	83.43	87.57	91.12	73.37

جدول (۱) نتایج حاصل از مقایسهٔ بازنمائی‌های MFCC با بازنمائی‌های مبتنی بر مدل فیزیولوژیک

۵- جمع‌بندی و نتیجه‌گیری

در این مقاله مدلی بر مبنای یافته‌های فیزیولوژیک برای استخراج ویژگی جهت سیستم‌های بازشناسی گفتار معرفی شده است. نتایج بدست آمده حاکی از آن است که مدل ارائه شده برای غشای پایه و سلولهای مژکدار دارای قابلیت‌هائی برای بهبود نتایج بازشناسی گفتار در حضور نویز می‌باشد. همچنین می‌توان بیان داشت که رویکردهای فیزیولوژیک و مدل‌کردن گوش انسان نسبت به استخراج ویژگی‌های مرسوم و سنتی دارای نتایج بهتری می باشد.

۶- مراجع

[1] D.W. Massaro, ***"Speech Perception by Ear and Eye: a Paradigm for Psychological Inquiry"***, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1987.

[2] B. Dodd & R. Campbell, Eds., ***"Hearing by Eye:The Psychology of Lip-Reading"***, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1987.

[3] M. D. Skowronski and J. G. Harris ***"INCREASED MFCC FILTER BANDWIDTH FOR NOISE-ROBUST PHONEME RECOGNITION"***, Proc. IEEE Int. Conf. Acoustic, Speech, Signal Processing, (Orlando, FL, U.S.A.), May 2002.

[4] R. Rhoades and R. Pflanzer, ***Human Physiology***, 3rd Edition, Saunders College Publishing, 1996.

[5] R. Chabay and B. Sherwood, ***Matter and Intractions,*** John Wiley & Sons, 2002.

[6] C. H. Edwards, D. E. Penney, ***Differential Equations and Boundary Value Problems: Computing and Modeling,*** 2nd Edition, Prentice Hall, 1999.

[7] J. Stewart, ***Calculus, Early Transcendentals,*** 4th Edition, Brooks Cole, 1999.

[8] H. D. Young, R. A. Freedman, ***University Physics with Modern Physics with Mastering Physics,*** 11th Edition, Benjamin Cummings, 2004.

[9] S. Young, G. Everman, D. Kershaw, G Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, ***The HTK Book (for HTK Version 3.2)***, Cambridge University, 2002.