مجله علمی-پژوهشی «علوم و فناوری‌های پدافند غیرعامل»
سال سوم، شماره ۱، بهار ۱۳۹۱: ص ۳۸- ۲۹

یک روش سریع برای تقطیع گفتار گویندگان بر اساس بسامد گام گفتار (RPSS)

بهروز عبدالعلی*[۱]، حسین صامتی[۲]، محمدحسین قزل ایاغ [۳]

۱- کارشناس ارشد، ۳- استادیار، گروه الکترونیک، دانشگاه جامع امام حسین (ع) ۲- دانشیار، گروه کامپیوتر، دانشگاه صنعتی شریف

(دریافت: ۱۳۹۰/۰۵/۱۲، پذیرش: ۱۳۹۱/۰۲/۲۰)

چکیده

تقطیع و خوشه‌بندی گویندگان فرآیندی است که طی آن قطعه‌بندی و برچسب‌گذاری برای گفتار حاصل از یک جلسه که شامل چند گوینده است انجام می‌شود و دنباله صوتی به بخش‌هائی تقسیم می‌شود که هر بخش شامل فقط یک گوینده است و با برچسب‌گذاری مشخص می‌شود که هر بخش مربوط به‌کدام گوینده است. تشخیص فعالیت گفتاری، تقطیع گفتار و خوشه‌بندی گویندگان، حداقل مراحل اصلی سامانه‌های تقطیع و خوشه‌بندی گفتار بر اساس گوینده محسوب می‌شوند. برای مرحله‌ی تقطیع روش‌های متعددی وجود دارد که تقطیع بر مبنای BIC یکی از روش‌های مرسوم است. این روش به دلیل محاسبات آماری آن، نیاز به زمان محاسبات طولانی دارد. هدف اصلی این مقاله ارائه‌ی روش تقطیع بر اساس بسامد گام گفتار است، که هم دارای دقتی در حد روش‌های مرسوم است و هم دارای سرعت محاسبات بالا است، به طوری که در مقایسه با تقطیع بر اساس BIC به طور میانگین دارای مزیت ۲/۴ برابری در سرعت و افزایش یک درصد در دقت است.

کلید واژه‌ها: تقطیع و خوشه‌بندی گویندگان، تقطیع گفتار، بسامد گام گفتار.

A Method for Rapid Pitch-Based Speaker Segmentation (RPSS)

B. Abdolali*, H. Sameti, M. H. Ghezeayagh

Department of electronic , Imam Hossein University

(Received: 08/03/2011, Accepted: 05/09/2012)

Abstract

Audio Segmentation has got a widespread application in underwater signal processing specially observation of passing objects using their acoustic signals and also in audio annotation of a recorded meeting, which also used by intelligence services that is one of instances of passive defense.Audio Segmentation is the process of partitioning an audio stream into regions each of which corresponds to only one audio source or one speaker.There are various methods for speaker segmentation the most common methods of which are based on BIC criteria. These methods needs heavy statistical computations and are very time consuming.The main goal of this paper is to propose a new audio segmentation method based on pith frequency with acceptable accuracy along with a higher computation speed than BIC-based methods. This algorithm is about 2.4 times faster than the BIC-based segmentation an about %1 higher than that of the BIC-based method in accuracy.

Keywords: Speaker Segmentation and Clustering, Speech Segmentation, Pitch-based Speaker Segmentation

* Corresponding author E-mail: g8712905738@ihu.ac.ir

۱. مقدمه

سامانه تقطیع[1] گفتار، به طور وسیعی در کاربردهای پردازش گفتار مورد استفاده قرار می‌گیرد. یکی از کاربردهای آن در سامانه تقطیع و خوشه‌بندی گویندگان[2] است که به‌عنوان یکی از بخش-های اساسی این سامانه مورد استفاده قرار می‌گیرد. هدف اصلی این سامانه، تقسیم گفتار به بخش‌هایی است که هر بخش شامل گفتار تنها یک گوینده باشد و با اندیس‌گذاری بخش‌های گفتار هر گوینده مشخص شود. تشخیص فعالیت گفتاری، تقطیع گفتار و خوشه‌بندی گویندگان، حداقل ملزومات اصلی سامانه‌های تقطیع و خوشه‌بندی گویندگان محسوب می‌شود.

اولین بخش این سامانه، تشخیص فعالیت گفتاری[3] است. در این مرحله با استفاده از استخراج ویژگی‌هایی از بخشی از گفتار در مورد گفتار یا غیر گفتار بودن آن بخش تصمیم‌گیری می‌شود و گفتار از غیر گفتار جدا می‌شود. دومین بخش این سامانه، تشخیص تغییر گوینده[4] یا تقطیع گفتار است یعنی نقاط تغییر گوینده از نظر زمانی مشخص می‌شوند، به بیان دیگر هدف تقطیع، مشخص کردن نقاط تغییر گوینده یا منبع صوتی است. برای این مرحله، روش‌های مختلفی ارائه شده است. این روش‌ها به چند دسته‌ی اصلی تقسیم می‌شوند که عبارتند از روش‌های تقطیع بر اساس فاصله[5]، تقطیع بر اساس مدل و روش‌های ترکیبی. هر کدام از این روش‌ها مزایا و معایبی دارند که در بخش‌های بعدی به تشریح آنها خواهیم پرداخت. با ترکیب این مرحله با مرحله‌ی تشخیص فعالیت گفتاری، قطعاتی از دنبالهٔ صوتی که فقط حاوی گفتار هستند مشخص می‌شوند و در بخش آخر سامانه این قطعات خوشه بندی می‌شوند. یعنی روی قطعاتی که برای یک گویندهٔ واحد بوده‌اند اندیس مشترک زده می‌شود.

رایج‌ترین روش‌هایی که امروزه برای تقطیع گفتار بر اساس گوینده استفاده می‌شوند، روش‌های مبتنی بر BIC[6] هستند که جزء روش‌های تقطیع بر اساس فاصله محسوب می‌شوند. این روش‌ها علیرغم داشتن دقت مناسب، به علت حجم بالای محاسبات، زمان‌گیر و کند هستند. به همین دلیل روش‌هایی برای تقطیع گفتار بر اساس گوینده ارائه می‌شوند که ضمن داشتن دقت مناسب و قابل قبول، محاسبات کمتری نیاز داشته باشند و بتوانند به سرعت‌های بالاتری دست یابند. برای این هدف نیز روش‌هایی ارائه شده است.

در این مقاله قصد داریم به معرفی روشی برای تقطیع گفتار بپردازیم که به مانند روش‌های تقطیع بر اساس فاصله، نیاز به اطلاعات قبلی از گویندگان نداشته باشد و دارای محاسبات سنگین نیز نباشد. به بیان دیگر به دنبال ارائه یک روش تقطیع سریع، و حل مشکل کندی سامانه‌های تقطیع بر اساس فاصله هستیم. ما برای حل این مشکل استفاده از اطلاعات بسامد گام گفتار را پیشنهاد داده‌ایم و بر اساس تابع تغییرات بسامد گام گفتار در طول دنباله صوتی نقاط تغییر گوینده را تشخیص می-دهیم. البته استفاده از این تغییرات مشکلاتی هم دارد که باعث افزایش خطا می‌شود که با استفاده از روش‌هایی خطا را کاهش داده‌ایم که در ادامه توضیح داده خواهند شد.

در بخش دوم به تشریح روش‌های مختلف تقطیع گفتار پرداخته-ایم، سپس در بخش سوم مباحث نظری پیرامون بسامد گام گفتار به تفصیل بیان شده است. در بخش چهارم این مقاله، روش ارائه شده برای تقطیع گفتار بر اساس بسامد گفتار (RPSS) توضیح داده شده است و در بخش پنجم روش بهبود RPSS را ارائه کرده-ایم که مشکلات کاهش دقت روش RPSS را رفع کرده است. بعد از آن در بخش ششم به توضیح روش‌های ارزیابی و دادگان‌های مورد استفاده برای سامانه‌های تقطیع پرداخته شده است و نتیجه‌های مقایسه‌ی روش ارائه شده با روش مبتنی بر BIC نشان داده شده است.

۲. انواع روش‌های تقطیع گفتار

اولین و ابتدایی‌ترین الگوریتم تقطیع گفتار، مبتنی بر انرژی است که وابسته به یک مقدار آستانه برای آشکارسازی تغییر گویندگان است. اما این روش از نظر دقت، بسیار ضعیف است. بنابراین روش‌های پیشرفته‌تری برای تقطیع گفتار ارائه شده است که به طور کلی به سه دسته تقسیم می‌شوند. تقطیع بر اساس مدل، تقطیع بر اساس فاصله و روش‌های مرکب که از ترکیبی از مدل و فاصله استفاده می‌کنند[۱].

۲-۱. تقطیع بر اساس مدل

در روش‌های مبتنی بر مدل، مجموعه‌ای از مدل‌ها تعریف می‌شود و دنباله‌ی صوت ورودی بر اساس این مدل‌ها و گویندگان مختلف تقطیع می‌شود. بنابراین در این روش داشتن اطلاعات پیشین از گویندگان ضروری است. به‌طورمعمول یک مدل پس زمینه عمومی (UBM) آموزش داده می‌شود تا یک مدل عمومی از گویندگان ایجاد شود. .یک مدل عمومی دیگر مدل گوینده نمونه (SSM) است که یک مدل اولیه مستقل از گوینده ارائه می‌کند. مدل‌ها می‌توانند توسط مدل مخفی مارکوف (HMM) یا روش‌های دیگر نیز ساخته شوند. این مدل‌ها در پی این هستند که نرخ دقت تقطیع را افزایش دهند[۴].

۲-۲. تقطیع بر اساس فاصله

در این روش ابتدا یک معیار فاصله برای دو قطعه‌ی صوتی تعریف می‌شود و بر اساس آن به دنبال میزان ناهمانندی یا فاصله‌ی بین پنجره‌های تحلیل گفتار هم‌جوار است که بر روی جریان صوتی

---

[1] Segmentation
[2] Speaker Diarization
[3] Voice Activity Detection
[4] Speaker Change Detection
[5] Distance-Based Segmentation
[6] Bayesian Information Criterion

تعیین شده است[۵]. دو پنجره‌ای که تابع فاصله در آن نقطه دارای بیشینه‌ی محلی باشد، در واقع متعلق به گفتارهای دو گوینده متفاوت خواهد بود. برای یافتن بیشینه‌ی محلی لازم است که یک مقدار آستانه تعریف شود که مقادیر بیشتر از آن به عنوان نقطه بیشینه‌ی محلی استخراج شوند.

روش‌های مبتنی بر فاصله نیازی به داشتن اطلاعات پیشین از تعداد گویندگان، هویت آنها یا مشخصات سیگنال صوتی ندارند و تنها با اندازه‌گیری معیار فاصله، تغییر گوینده را تشخیص می‌دهند. معیارهای فاصله متعددی تاکنون برای تقطیع معرفی شده است. معیار مجذور فاصله اقلیدسی وزن‌دار ، معیار همگرایی کالبک لیبلر ، معیار همگرایی KL2 (یا KL متقارن)، معیارهای آماری مرتبه دوم مانند معیار کروی بودن، و معیار آماری Hotelling T^2 همچنین معیار نسبت شباهت تعمیم یافته(GLR) نیز از معیارهایی هستند که برای بخش بندی استفاده می‌شوند. اما رایج‌ترین معیاری که برای بخش بندی استفاده می‌شود، معیار BIC است[۴].

به همین دلیل لازم است که روش‌های جدیدی که در حوزه تقطیع گفتار ارائه می‌شود، با این روش مقایسه شوند. بدین منظور در بخش ۲-۳ به تفصیل به معرفی این روش پرداخته خواهد شد.

۲-۳. تقطیع گفتار بر اساس معیار BIC

BIC یک رویکرد برای انتخاب مدل برای یک دسته داده است که توسط شوارز[۲] ارائه شد[HYPERLINK \l "GSc" 6]. فرض می‌کنیم که یک دسته داده X داریم و یک مدل برای بیان این داده‌ها به نام M وجود دارد، مقدار امتیاز BIC برای این مدل به صورت زیر نشان داده می‌شود:

$$BIC_M = \log p(X|M) - \lambda \frac{\#M}{2} \log N \qquad (۱)$$

که در این فرمول، $P(X|M)$ مقدار شباهت[۳] داده X نسبت به مدل M را نشان می‌دهد و $\#M$ نشان دهنده تعداد پارامترهای آزاد در مدل M است و N نشان دهنده تعداد نمونه‌های داده‌های X است.

به بیان دیگر BIC میزان شباهت مدل را به داده‌ها اندازه می‌گیرد و به مدل امتیاز می‌دهد[۷]. در این رابطه به λ عامل جریمه[۴] گفته می‌شود، اگر λ برابر صفر قرار گیرد BIC به GLR تبدیل می‌شود[۸]. برای دست‌یابی به کارایی مورد نظر در یک دادگان می‌توان λ را تنظیم کرد[۵].

همان‌طور که در [۶] توضیح داده شده است، بیشینه شدن BIC منجر به بیشینه شدن مقدار مورد انتظار شباهت مدل و داده‌ها می‌شود بنابراین از BIC می‌توان برای انتخاب بهترین مدل برای بیان یک دسته داده استفاده کرد[۶،۷].

فرض کنید $X = x_i \in R^d \ , \ i = 1,2,...,N$ دنباله‌ای از بردارهای ویژگی با بعد d بر اساس فریم استخراج شده از روی دنباله صوتی باشد. در این کاربرد به طور معمول از بردارهای ویژگی MFCC استفاده می‌شود. البته معیار BIC هیچ فرضی درباره‌ی روش استخراج ویژگی ندارد، بنابراین این روش قابل تعمیم به مواردی است که از روش‌های دیگر استخراج ویژگی استفاده می‌کنند. فرض می‌کنیم که در این دنباله‌ی صوتی حداکثر یک مرز قسمت وجود دارد. مسئله تعیین اینکه آیا یک تغییر گوینده (مرز قسمت) در فریم $b \in (I, N)$ وجود دارد یا نه، می‌تواند به یک مسئله انتخاب مدل تبدیل شود[۵].

برای این کار حول زمان فرضی b دو پنجره همسایه X و Y را در نظر می‌گیریم، می‌خواهیم تصمیم بگیریم که تغییر گوینده در این زمان اتفاق افتاده یا نه[۴].

مدل M_1 فرض می‌کند که همه نمونه‌های X مستقل هستند و به طور یکسان توسط یک فرآیند گاوسین چند متغیره[۵] توزیع می‌شود.

$$M_1 : \quad Z = z_1, z_2, \dots, z_N \sim N(\mu_z, \Sigma_z) \qquad (۲)$$

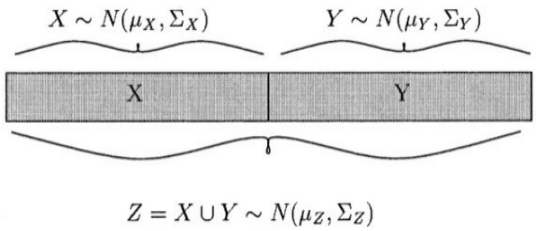
مدل M_2 فرض می‌کند که X توسط دو فرآیند گاوسین چند متغیره ایجاد شده است، یکی تا نقطه‌ی b و یکی از این نقطه تا انتهای داده‌های آن محدوده.

$$M_2: \quad Z = X + Y \qquad (۳)$$

$$X = z_1, z_2, \dots, z_b \sim N(\mu_x, \Sigma_x) \qquad (۴)$$

$$Y = z_{b+1}, z_{b+2}, \dots, z_N \sim N(\mu_y, \Sigma_y) \qquad (۵)$$

این دو فرض در شکل(۱) به خوبی نشان داده شده است.



شکل۱. مدل‌های فرضی بری تقطیع یک قطعه از دنباله‌ی صوتی

[1] Kullback-Leiber
[2] Schwarz
[3] Likelihood
[4] Penalty Factor
[5] Multivariate Gaussian Process

$$\Delta BIC\{x,y\} = BIC(M_2,Z) - BIC(M_1,Z)$$

$$= log\, p(X|\,\widehat{\mu_x},\widehat{\Sigma_x}) + log\, p(Y|\,\widehat{\mu_y},\widehat{\Sigma_y}) - log\, p(Z|\,\widehat{\mu_z},\widehat{\Sigma_z}) - \frac{1}{2}\lambda\left(d+\frac{1}{2}d(d+1)\right)log\, n$$

$$= \frac{n}{2}\log|\widehat{\Sigma_z}| - \frac{n_x}{2}\log|\widehat{\Sigma_x}| - \frac{n_y}{2}\log|\widehat{\Sigma_y}| - \frac{1}{2}\lambda\left(d+\frac{1}{2}d(d+1)\right)\log n \quad (۶)$$

که $\widehat{\Sigma_y},\widehat{\Sigma_x},\widehat{\Sigma_z}$ برآوردهای ماتریس کواریانس با بیشینه درست نمایی از روی داده نظیر هستند و عمل‌گر $|.|$ به معنای دترمینان ماتریس مربوطه است و d بعد بردار ویژگی کپسترال است. در این رابطه به λ عامل جریمه[1] گفته می‌شود، اگر λ برابر صفر قرار معیار BIC به GLR تبدیل می‌شود[۸].

اگر $\Delta BIC = \text{BIC(M2)} - \text{BIC(M1)} > 0$ باشد به این معناست که امتیاز این که داده‌ها با دو توزیع گاوسی تعریف شوند (M2) بیشتر از حالتی است که داده‌ها توسط یک مدل واحد تعریف شوند(M1)، یعنی داده یکنواخت نیست و نقطه شکست (تغییر گوینده) داریم. در غیر این صورت داده یکنواخت بوده و نقطه شکست (تغییر گوینده) نداریم.

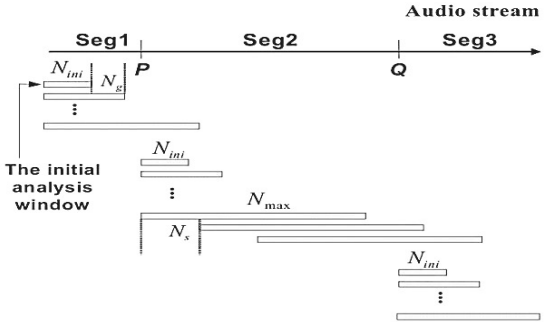
باید توجه داشت که BIC تنها برای به‌دست آوردن حداکثر یک نقطه تغییر آکوستیکی در داده‌های صوتی کاربرد دارد. بنابراین لازم است از الگوریتم‌هایی برای به‌دست آوردن نقاط شکست بیشتر استفاده نماییم. برای این منظور، الگوریتم‌های آشکارسازی ترتیبی پیشنهاد شده است [۸]. می‌توان تفاضل BIC یا ΔBIC را به عنوان تابعی از نقطه‌ی شکست b نوشت. اگر نقطه‌ی شکست را b فرض کنیم آنگاه تعداد بردارهای موجود در X یا همان n_x برابر با b خواهد شد و مقدار n_y هم $n-b$ خواهد شد و فرمول بر حسب b به صورت زیر خواهد شد.

$$\Delta BIC_b\{x,y\} = \frac{n}{2}\log|\widehat{\Sigma_z}| - \frac{b}{2}\log|\widehat{\Sigma_x}| - \frac{n-b}{2}\log|\widehat{\Sigma_y}| - \frac{1}{2}\lambda\left(d+\frac{1}{2}d(d+1)\right)\log n \quad (۷)$$

بر اساس ΔBIC تقطیع جریان صوتی به دو بخش در فریم b وقتی صحیح است که $\Delta BIC(b) > 0$ باشد. مقدار مثبت بدین معناست که مدل M_2 سیگنال را بهتر توصیف می‌کند و نقطه شکست b وجود دارد. در ادامه چند الگوریتم معمول برای پیدا کردن نقاط تغییر در طول دنباله‌ی صوتی توضیح داده خواهد شد. روش پنجره بزرگ شونده برای محاسبهٔ ΔBIC است.

این روش برای تشخیص بیش از یک نقطه تغییر در دنباله صوتی است. همان‌طور که در شکل (۲) در این روش یک اندازه اولیه برای پنجره در نظر گرفته می‌شود که به اندازه N_{ini} بردار ویژگی در آن موجود است.

این پنجره به صورت مداوم با اندازه N_g بزرگ می‌شود تا اینکه بر اساس معیار BIC یک نقطه تغییر یافت شود. البته یک مقدار حداکثر N_{max} نیز برای اندازه پنجره تعریف می‌شود تا از آن بیشتر نشود. اگر قبل از رسیدن اندازه پنجره به N_{max} یک نقطه شکست پیدا شود آن نقطه علامت گذاری شده و فرآیند از همان نقطه با اندازه پنجره اولیه دوباره شروع می‌شود و اگر پیدا نشود بعد از رسیدن اندازه به N_{max} پنجره به اندازه N_s نمونه جابجا می‌شود و فرآیند دوباره تکرار می‌شود تا اینکه به پایان دنباله صوتی برسد[۸]. البته ذکر این نکته مهم است که این روش دارای سرعت کم و حجم پردازش بالا است، چرا که با گذر زمان و افزایش طول پنجره و با اضافه شدن نمونه‌های جدید به حجم محاسبات افزوده می‌شود.



شکل۲. الگوریتم پنجره‌ی بزرگ شونده برای تشخیص نقاط تغییر[۵]

۲-۳-۲. روش پنجره ثابت لغزان[2] برای محاسبهٔ ΔBIC

در این روش یک پنجره با طول ثابت در نظر گرفته می‌شود و با لغزاندن آن در طول دنباله صوتی، محاسبات ΔBIC انجام می‌گیرد. طول پنجره بسته به طول دنباله صوتی متفاوت است و با تنظیم آن و همچنین مقدار λ می‌توان به سرعت و دقت خوبی دست یافت. چیزی که مبرهن است این است که این روش از حجم محاسبات کمتری نسبت به روش قبل برخوردار است اما دقت آن نسبت به روش قبل کمتر است. بر طبق آزمایش‌هائی که انجام گرفت مشخص شد که برای صوت‌های کوتاه در حدود چند دقیقه، طول ۱s مقدار خوبی برای طول پنجره است[۸]. ما در این مقاله روش اول را پیاده سازی نموده‌ایم تا بتوانیم بیشترین دقت‌های حاصل از روش‌های مبتنی بر BIC را با روش تقطیعی که خود ارائه کرده‌ایم مقایسه کنیم.

[1] Penalty Factor

[2] Fixed-size Sliding

۳. مباحث نظری پیرامون بسامد گام گفتار

در این بخش بعد از مطرح کردن ویژگی‌های بسامد گام گفتار، به بیان چند روش مهم برای استخراج آن پرداخته می‌شود.

۳-۱. بسامد گام گفتار و ویژگی‌های آن

بسامد گام گفتار همان هارمونیک اصلی سیگنال گفتار انسان است و همان فرکانس اصلی لرزش تارهای صوتی انسان است. هرچقدر شکل موج سیگنال، شبیه‌تر به شکل موج سینوسی باشد، مفهوم بسامد واضح‌تر و مفهوم گام ناواضح تر است.] همین طور هرچه مؤلفه‌های فرکانسی نسبت به هم هارمونیک باشند، مفهوم گام واضح‌تر و مفهوم بسامد نا واضح‌تر است[۱۲- ۱۰].

۳-۲. روش‌های استخراج بسامد گام گفتار

کارهای بسیاری برای تخمین بسامد اصلی یا f_0 سیگنال انجام شده است، اما ارائه یک روش تخمین که فارغ از محتوای سیگنال بتواند به خوبی f_0 را تخمین بزند، کاری مشکل است. بنابراین در زمینه‌ای که هم موسیقی وجود دارد و هم گفتار باید تخمین-گرهای دقیقی استفاده شود که در هر دو زمینه دقیق باشند. سختی تشخیص f_0 در یک شکل موج، وابسته به خود شکل موج است، یعنی اگر شامل هارمونیک‌های بالاتر کم‌تری در طیف بسامد باشد یا توان هارمونیک‌های بالاتر، پایین باشد، تشخیص f_0 راحت‌تر خواهد بود.

روش‌های تعیین بسامد گام که به نام PDA معروف هستند، برای بسیاری از الگوریتم‌های پردازش گفتار اهمیت بسیاری دارد. در این قسمت، روش‌های تشخیص بسامد گام از طریق روش تابع خود همبستگی ، روش کپسترروم ، روش کدگذاری پیش‌گوئی خطی (LPC) و تابع میانگین تفاوت دامنه توضیح داده می‌شود.

۳-۲-۱. تشخیص بسامد گام از طریق روش خود همبستگی

درک ما از بسامد گام به شدت با درک ما از تناوبی بودن شکل موج در حوزه زمان، ارتباط دارد. روشی که بتواند بسامد اصلی سیگنال را از روی شکل موج تشخیص دهد، همان روش خود همبستگی است[۱۴]. تابع خود همبستگی یک سیگنال s(n) به صورت زیر است [۱۵] که در آن τ مقدار تأخیر یا شیفت است. با محاسبه‌ی این معیار و تشخیص نقاط بیشینه می‌توان تخمینی از فرکانس گام را محاسبه کنیم.

$$R(\tau)=\sum_{n=0}^{N-1} s(n).s(n+\tau) \qquad (۸)$$

۳-۲-۲. تشخیص بسامد گام با استفاده از روش کپسترال

تحلیل کپسترال راهی برای تخمین بسامد گام فراهم می‌آورد. فرض کنیم که دنباله‌ای از نمونه‌های صوتی گفتار حاصل کانولوشن دنباله برانگیختگی ناشی از حنجره e[n] و پاسخ ضربه گسسته مربوط به تارهای صوتی θ[n] باشد. در حوزه فرکانس عمل‌گر کانولوشن به عمل‌گر ضرب تبدیل می‌شود. با استفاده از خاصیت تابع لگاریتم که log(A.B)=log(A)+log(B) عمل‌گر ضرب می‌تواند به عمل گر جمع تبدیل شود. در نهایت کپسترروم حقیقی یک سیگنال با فرمول s[n]=e[n]*θ[n] به صورت رابطه‌ی زیر خواهد بود:

$$c[n]=\frac{1}{2\pi}\int_{-\pi}^{\pi}\log\ |S(\omega)|e^{jn\omega}d\omega \qquad (۹)$$

که در این فرمول داریم:

$$S(\omega)=\sum_{n=-\infty}^{\infty} s[n]e^{-jn\omega} \qquad (۱۰)$$

بنابراین، کپسترروم در واقع حاصل تبدیل فوریه روی لگاریتم دامنه طیف سیگنال است. اگر لگاریتم دامنه‌ی طیف شامل هارمونیک‌های بسیاری باشد که به فواصل منظم از یکدیگر قرار گرفته‌اند باشد، تبدیل فوریه طیف دارای قله‌ای است که متناظر با فاصله بین هارمونیک‌ها است. این قله، در واقع فرکانس اصلی یا فرکانس گام را نمایش می‌دهد.

۳-۲-۳. تشخیص بسامد گام با استفاده از روش تابع میانگین تفاوت دامنه

مفهوم AMDF (تابع میانگین تفاوت دامنه) بسیار نزدیک به مفهوم ACF (تابع خود همبستگی) است، به‌جز اینکه در این تابع به جای تخمین شباهت بین فریم و نسخه تأخیر یافته‌اش، تفاوت دامنه بین فریم و نسخه تأخیر یافته‌اش را تخمین می‌زند. محاسبه‌ی AMDF در رابطه‌ی (۱۱) آمده است، در این رابطه τ محدوده زمانی بر حسب تعداد نمونه‌های صوتی است. مقدار τ که به ازای آن amdf(τ) در یک محدوده‌ی خاص کمینه می‌شود به عنوان دوره تناوب گام صوتی انتخاب می‌شود.

به عبارت دیگر نسخه‌ی تأخیر یافته را n بار جا بجا می‌کنیم و قدرمطلق مجموع تفاوت در بخش‌های هم‌پوشان را محاسبه می-کنیم تا n مقدار AMDF به‌دست آید. عدد فرکانس گام صوتی از تقسیم فرکانس نمونه‌برداری بر نمونه صوتی متناظر با مقدار اولین کمینهٔ محلی AMDF حاصل می‌شود.

[1] Pitch Determination Algorithm (PDA)
[2] Autocorrelation Function (ACF)
[3] Cepstrum
[4] Average Magnitude Difference Function (AMDF)
[5] Glottal Excitation
[6] Vocal Tract's Discrete Impulse Response

۴. تقطیع به روش پیشنهادی RPSS

روش‌های تقطیع که بر اساس فاصله عمل می‌کردند، مثل روش تقطیع بر اساس BIC ، از ویژگی‌های کپسترال مثل MFCC استفاده می‌کردند. اما بردارهای ویژگی دیگری نیز وجود دارند.علاوه بر این، ویژگی‌های عروضی[1] مثل فرکانس گام هم برای تشخیص گفتار از سکوت می‌توانند به ما کمک کنند، همچنین نمودار تغییرات فرکانس گام می‌تواند برای تشخیص گوینده مورد استفاده قرار گیرد [۱۶]. در اینجا به ۳ دلیل از فرکانس گام برای تقطیع استفاده شده است:

۱- مقدار فرکانس گام برای هر گوینده مقدار متفاوتی است.

۲- در زمان تغییر گوینده، فرکانس گام تغییرات تندی دارد.

۳- برای قطعه گفتارهای زیر ۱s روش‌های دیگر مثل BIC که از ویژگی‌های MFCC استفاده می‌کنند دقت خوبی ندارند زیرا در این قطعات کوچک اطلاعات کافی در دسترس نیست.

برای آنکه بتوانیم از اطلاعات گام گفتار برای تقطیع استفاده کنیم ابتدا باید به یکی از روش‌های پیش گفته مقادیر فرکانس گام را استخراج کنیم. به علت سرعت بیشتر روش AMDF، ما این روش را برگزیدیم.

بعد از استخراج مقادیر گام گفتار برای تک تک فریم‌های صوتی، حال نوبت به استفاده از این اطلاعات رسیده است. از آنجا که تغییر گوینده را با تغییرات ناگهانی فرکانس گام نیز می‌توان تشخیص داد. از این پس با استفاده از داده‌های فرکانس گام سعی می‌کنیم تقطیع انجام دهیم. فرض کنید ما N پنجره داریم که مقادیر فرکانس گام را برای هر کدام از آنها با استفاده از یکی از روش‌های گفته شده محاسبه کرده‌ایم، برای تشخیص تغییر فرکانس گام از تابع مشتق استفاده می‌کنیم که در رابطه‌ی زیر آمده است:

$$Diff(x) = |pitch(n+1) - pitch(n)| \qquad (۱۲)$$

آنگاه یک مقدار آستانه تعریف می‌کنیم که اگر تفاوت از آن بیشتر بود، به عنوان تغییر گوینده مد نظر قرار گیرد. ما این آستانه را به طور میانگین ۰/۷ بیشینهٔ اختلاف بین فرکانس‌های گام در نظر گرفته‌ایم. در نهایت مقدار اندیس n مربوط به اختلاف‌های بیشتر از آستانه را ذخیره می‌کنیم و با استفاده از آن شروع و پایان قطعه را به‌دست می‌آوریم، بر اساس این روش ما نقاط تغییر تند را به عنوان تغییر گوینده در نظر می‌گیریم. این روش سرعت بسیار خوبی دارد. اما در مقابل، دقت آن در مقایسه با روش مبتنی بر معیار BIC کمتر است. در قسمت بعد نحوه بهبود دقت این روش توضیح داده شده است.

۵. بهبود دقت روش RPSS

با نگاهی دقیق‌تر به تغییرات فرکانس گام متوجه می‌شویم که تغییرات آن بسیار تند است و حتی در طول گفتار یک گوینده امکان دارد تغییرات تند این فرکانس را داشته باشیم. در این صورت خطای اعلام غلط (FA) افزایش می‌یابد، چرا که ما هر تغییرات تندی را به عنوان تغییر گوینده در نظر گرفته‌ایم و این یکی از مشکلات تقطیع با استفاده از تغییرات فرکانس گام است.

مشکل دیگر این روش این است که امکان دارد که تغییر فرکانس گام از یک گوینده به گوینده دیگر آنچنان تند نباشد که از مقدار آستانه‌ی در نظر گرفته شده بیشتر شود که به عنوان نقطه‌ی تغییر معرفی شود که این باعث افزایش خطای تشخیص از دست رفته (MD) می‌شود.

برای حل این مشکل شاید این راه حل پیشنهاد شود که مقدار آستانه را کوچک در نظر بگیریم که این تغییرات در نظر گرفته شده از مقدار آستانه بزرگ‌تر شوند که تغییرات حس شوند.

پرواضح است که با این راه حل بسیاری از تغییراتی هم که ناشی از تغییر گوینده نبوده‌اند بزرگ‌تر از آستانه شده و به عنوان نقطه تغییر معرفی می‌شوند و با اینکه MD کاهش یافته است اما FA افزایش یافته و دقت کل را کاهش داده است.

برای حل این مشکل ما نیاز داریم که تابعی روی تغییرات فرکانس گام اعمال کنیم که تغییرات کوچک را بزرگ و تغییرات بزرگ را تغییر آنچنانی ندهد که از تابع تصحیح گاما استفاده می‌کنیم.

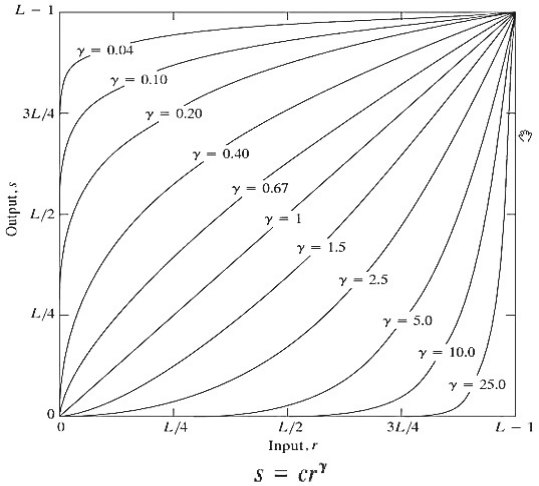
$$Gamma(f(x)) = C.f(x)^{\gamma} \qquad (۱۳)$$

این تابع را روی نمودار تغییرات فرکانس گام اعمال کردیم، با توجه به شکل (۳) و مشکل گفته شده،واضح است که برای کاربرد ما باید از $\gamma < 1$ استفاده شود که بهترین مقدار آن حدود ۰/۳ به‌دست آمد[۱۸].

با اعمال این تابع خطای MD کاهش می‌یابد اما مشکل زیاد بودن FA همچنان بر جای خود باقی است، چرا که به طور طبیعی از قبل تغییرات تند کاذب زیاد وجود داشت و با اعمال تابع تصحیح گاما هم مجدداً این مشکل تشدید شد چرا که تغییرات کوچک، بزرگ شدند.

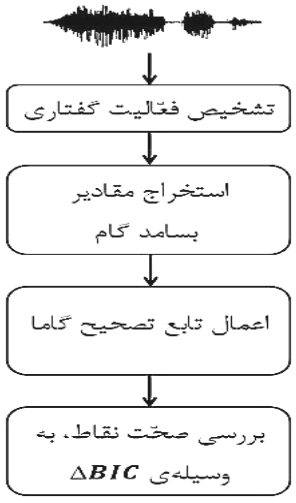
[1] Prosodic

[2] Gamma Correction Function



شکل ۳. تابع تصحیح گاما

برای کاهش خطای FA می‌توانیم از ایده‌ای که برای تقطیع به روش BIC استفاده کردیم استفاده کنیم به این صورت که همه تغییرات بالاتر از آستانه را به عنوان نامزد نقطه‌ی تغییر در نظر می‌گیریم و بعد از آن با استفاده از اعمال یک پنجره‌ی کوچک BIC به طول حدود 1s، صحت آن را بررسی می‌کنیم. نتایج آزمایش‌ها نشان داد که با اعمال این ایده بهبود فاحشی در دقت نهایی ایجاد خواهد شد و می‌توانیم بگوئیم که این ایده یکی از ایده‌های اساسی این مقاله است. در شکل (۴) فلوچارت الگوریتم پیشنهادی نشان داده شده است.



شکل ۴. فلوچارت تقطیع به روش پیشنهادیRPSS

۶. نتایج و بحث

واضح است که برای هر سامان‌های باید روش‌های ارزیابی استاندارد نیز تعریف شود، برای سامانه‌های تقطیع نیز روش‌های ارزیابی استاندارد تعریف شده‌اند. در این بخش به معرفی این معیارها و روش‌های ارزیابی سامانه‌های تقطیع پرداخته می‌شود. نتایج ارائه شده در این مقاله نیز بر اساس همین معیارها تهیه شده‌اند.

عملکرد سامانه می‌تواند در هر جلسه ضبط شده، مورد بررسی قرار گیرد اما نتایج معتبر آن است که نتایج آزمایش سامانه روی جلسات یک دادگان[1] گفتاری معتبر ارائه شود. بر همین اساس برای ارزیابی سامانه‌های گفتاری دادگان‌های متعددی تولید شده است، که می‌توان از دادگان جلسه‌ایNIST، AMI و TIMIT به عنوان مهم‌ترین آنها نام برد. برای ارزیابی سامانه‌های تقطیع، معیارهائی معرفی شده است که این معیارها مقایسه‌ای است بین نقاط تشخیص داده شده و نقاطی که در دادگان مورد آزمایش موجود بوده است. از مهم‌ترین معیارها $\boldsymbol{FD}$%و $\boldsymbol{FR}$% می‌باشند که تعاریف آنها در زیر ارائه شده‌اند[۱۷].

اعلام غلط[2] : تعداد نقاطی که در دادگان مرجع نقطه‌ی تغییر نبوده‌اند ولی توسط سامانه به عنوان نقطه‌ی تغییر معرفی شده‌اند. این خطا به نام اعلام غلط یا FA نیز شناخته می‌شود.

تعداد کل نقاط تغییر : کل تعداد نقاطی که توسط سامانه به عنوان نقطه‌ی تغییر معرفی شده‌اند.

نقاط تشخیص از دست رفته[4] : تعداد نقاطی که در دادگان مرجع نقطه‌ی تغییر بوده‌اند ولی توسط سامانه تشخیص داده نشده‌اند. این نقاط، نقاط تشخیص از دست رفته یا MD هستند.

total_amount_of_true_change_pointsتعداد کل نقاط تغییر صحیح : تعداد نقاطی که توسط سامانه به درستی به عنوان نقطه‌ی تغییر معرفی شده‌اند.

برای تعیین دقت تقطیع، معیار F به صورت زیر تعریف می‌شود :

$$F = \frac{2 * (1 - FD) * (1 - FR)}{2 - FD - FR} \qquad (۱۶)$$

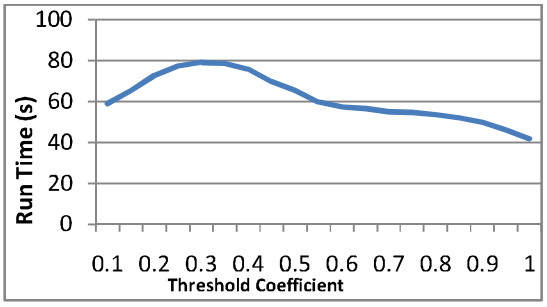
در این قسمت با توجه به نمودارهای حاصل از نتایج تغییر پارامترهای مهم دخیل در محاسبات روش بهبود یافته‌ی RPSS قصد داریم اثر این پارامترها را بر دقت و زمان اجرا روش پیشنهادی بررسی نماییم. در این مقاله نتایج آزمایش‌ها بر روی ۴ جلسه از جلسات دادگان AMI ارائه شده است. این جلسات به صورت تصادفی انتخاب شده‌اند و سامانه روی قسمتی از آنها آزمایش شده است که نام آنها در جداول نتایج مشخص است و نمودارها نیز با مقادیر میانگین این ۴ فایل تهیه شده‌اند.

[1]Corpus
[2]False Alarm
[3] Total Amount of Detection
[4] Missed Detection
[5] Total Amount of True Change Points

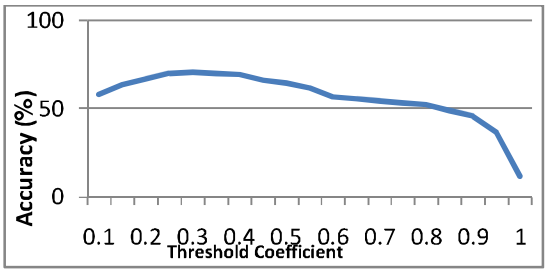
ضریب آستانهٔ تغییرات پارامتری است که اندازه‌ی تغییرات فرکانس گام قابل قبول نسبت به بیشینه‌ی جهانی را مشخص می‌کند، یعنی اگر این مقدار ۰/۷ باشد تغییراتی که بیش از ۰/۷ بیشینه‌ی جهانی هستند به عنوان نقطه‌ی تغییر معرفی می‌شوند. بنابراین باید اثر تغییرات این پارامتر را بر دقت و سرعت بررسی نماییم. همان‌طور که در نمودارهای شکل(۱) مشاهده می‌شود، هر چه این مقدار بیشتر باشد FA کاهش و MD افزایش می‌یابد و بالعکس. نتایج نشان می‌دهد که ۰/۷۵ مقدار خوبی برای این پارامتر است.

با توجه به نمودارها مشخص می‌شود که افزایش اندازه‌ی گام برای تصحیح گاما، تأثیر بسیار زیادی روی دقت دارد که برای این آزمایشات مقدار حدود۰/۳ بهترین دقت را نتیجه می‌دهد. جدول (۱) مقدار پارامترهای دخیل در دقت و سرعت روش پیشنهادی RPSS را برای هر یک از جلسات صوتی AMI به ازای دست‌یابی به بهترین دقت نشان می‌دهد. بعد از آنکه با تنظیم پارامترهای مختلف روش پیشنهادی RPSS به بهترین نقطه از نظر دقت و سرعت دست یافتیم، حال باید این روش را با روش مبتنی بر BIC که پیاده سازی نموده‌ایم مقایسه کنیم. جدول(۲) نتایج حاصل از دو روش را در حالتی که به بهترین دقت‌هایشان دست یافته‌اند، با یکدیگر مقایسه کرده است.

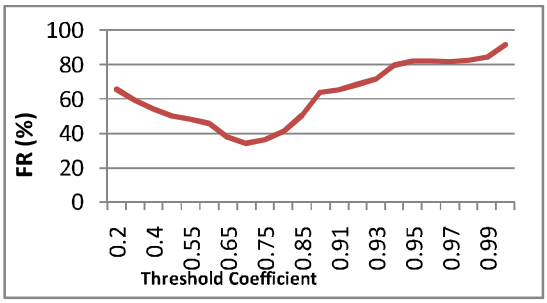
بر اساس جدول(۲) به این نتیجه می‌رسیم که تقطیع بر اساس روش پیشنهادی که بر مبنای تغییرات فرکانس گام استوار است در دادگان AMI به طور متوسط مزیت ۲/۴ برابری در سرعت انجام محاسبات و افزایش ۱٪ در دقت را نسبت به تقطیع بر اساس BIC با پنجره‌ی بزرگ شونده دارد.



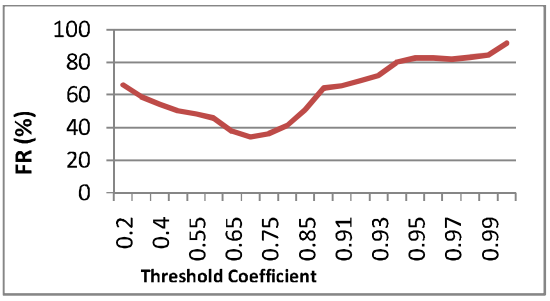
شکل۵. اثر تغییرات مقدار ضریب آستانه بر زمان اجرای روش RPSS



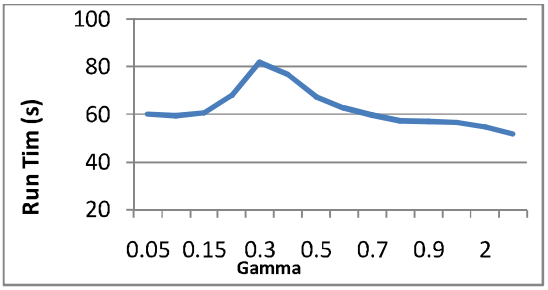
شکل ۶. اثر تغییرات مقدار ضریب آستانه بر دقت روش RPSS



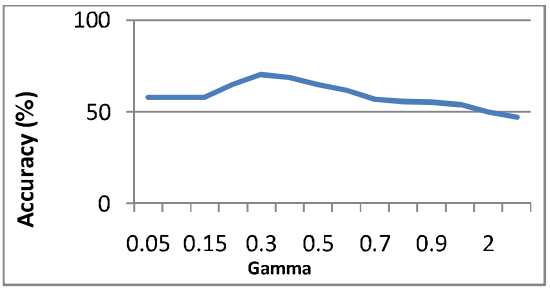
شکل ۷. اثرتغییرات مقدارضریب آستانه بر دقت روش RPSS



شکل ۸. اثر تغییرات مقدارضریب آستانه بر FR



شکل ۹. اثر تغییرات مقدار ضریب آستانه بر FD



شکل ۱۰– اثرتغییرات مقدارg بر دقت روشRPSS

جدول ۱. مقایسه دقت و سرعت روش پیشنهادیRPSS و روش مبتنی برBIC باپنجره‌ی بزرگ شونده

		BIC Based Segmentation				Pitch Based Segmentation (Proposed Method)				
File name (*.WAV)	Length (Sec)	%FD	%FR	%F	Run Time (Sec)	%FD	%FR	%F	Run Time (Sec)	Speed Up
ES2002a_p1	۳۶۵	۴۲/۸۷	۰/۶۳	۷۳/۵۵	۳۲۸/۹۸	۴۲/۰۷	۱/۲۶	۷۳/۰۲	۸۷/۸۶	۳/۷۴
ES2002b_p2	۳۹۰	۴۱/۶۲	۵/۶۲	۷۳/۵۶	۱۹۰/۸۵	۴۱/۰۲	۵/۶۱	۷۴/۰۴	۸۶/۵۷	۲/۲۰
ES2002c_p1	۴۰۷	۴۴/۰۷	۱/۴۵	۷۱/۳۶	۱۹۵/۵۲	۴۳/۹۳	۱۱/۵۹	۷۱/۵۰	۱۳۴/۹۳	۱/۴۴
ES2002d_p1	۳۶۰	۳۸/۸۸	۰	۷۵/۸۷	۱۷۶/۳۸	۳۷/۹۷	۱/۲۹	۷۶/۵۶	۸۰/۱۹	۲/۱۹

جدول ۲. بهترین مقادیر پارامترهای دخیل در روش RPSS

File name (*.WAV)	Pitch Thresh Coef.	Mfcc Win Size (Sample)	Mfcc Overlap Size	λ	BIC Win for Pitch (sec)	γ	%FD	%FR	%F	Run Time (Sec)
ES2002a_p1	۰/۳۲۵	۲۰۰	۱۲۰	۱	۰/۴	۰/۳	۴۲/۰۷	۱/۲۶	۷۳/۰۲	۸۷/۸۶
ES2002b_p2	۰/۳۲۵	۲۰۰	۱۲۰	۱/۲	۰/۲	۰/۳	۴۱/۰۲	۵/۶۱	۷۴/۰۴	۸۶/۵۷
ES2002c_p1	۰/۳۲۵	۲۰۰	۱۲۰	۱/۱	۱/۶	۰/۳۵	۴۳/۹۳	۱۱/۵۹	۷۱/۵۰	۱۳۴/۹۳
ES2002d_p1	۰/۳۲۵	۲۰۰	۱۲۰	۱	۰/۲	۰/۳	۳۷/۹۷	۱/۲۹	۷۶/۵۶	۸۰/۱۹

۷. نتیجه گیری

در این مقاله روشی سریع و دقیق برای تقطیع گفتار برای کاربرد تقطیع و خوشه‌بندی گویندگان ارائه شد. در این روش با استفاده از تغییرات فرکانس گام گفتار به تقطیع گفتار پرداخته می‌شود. این روش از نظر زمان انجام محاسبات بر روش تقطیع بر اساس BIC مزیت دارد اما دارای معایبی نیز هست، از جمله بالا بودن FA و MD که با استفاده از روش‌های پیشنهادی این معایب رفع شد. برای کاهش خطای FA یک مرحله به سامانه اضافه شد که نقاط تعیین شده را توسط معیار BIC دوباره مورد بررسی قرار می‌دهد و برای کاهش MD تابع تصحیح گاما به تغییرات فرکانس گام گفتار اضافه شد که تغییرات کوچک نادیده گرفته نشوند، با اضافه کردن این دو مرحله معایب روش پیشنهادی رفع شد و با استفاده از این روش هم به دقتی در حدود ۷۰٪ دست یافته شد،در حالی که سرعت انجام محاسبات با حدود همین دقت نسبت به روش BIC حدود ۲/۴ برابر شد.

۸. مراجع

[1] Moschou, V.; Kotropoulos, C.; Kotti, M. "Speaker Segmentation and Clustering."; Elsevier Journal of Signal Processing 2008, 88, 1091-1124.

[2] Zhu, X.; Meignier, S.; Gauvain, J. L.; Barras, C. "Multistage Speaker Diarization of Broadcast News."; IEEE Transactions on Audio, Speech, and Language Processing 2006, 14, 1505-1512.

[3] Lu,L.; Chen,K.; Zhang,H.; Wu,T. "UBM-Based Real-Time Speaker Segmentation for Broadcasting News."; in 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing 2003, pp. 193–196.

[4] Lu, L.; Chen, K.; Zhang, H.; Wu, T. "Universal Background Models for Real-Time Speaker Change Detection."; in Proc. of International Conference on Multimedia Modeling 2003, pp. 135–149.

[5] Cheng, S. S; Wang, H. M.; Fu, H. C. "BIC-Based Audio Segmentation by Divide-and-Conquer."; in IEEE International Conference on Acoustics, Speech and Signal Processing 2008, 4841 - 4844.

[6] Schwarz, G. "Estimating the Dimension of a Model."; The Annals of Statistic 1978, 6, 461-464.

[7] Hieu, N. T.; "Speaker Diarization in Meetings Domain."; Proposal for Admission to the Degree of Doctor of Philosophy, School of Computer Engineering, Nanyang Technological University, 2009.

[8] Cheng, S. S; Wang, H. M.; Fu, H. C. "BIC-Based Speaker Segmentation Using Divide-and-Conquer Strategies with Application to Speaker Diarization."; IEEE Transaction on Audio, Speech, and Language Processing 2010, 18, 141 – 157.

[9] Wellekens, C. J.; Delacourt, P. "DISTBIC: a Speaker-Based Segmentation for Audio Data Indexing."; Elsevier Journal on Speech Communication 2000, 32, 111-126.

[10] Truax, B.; "Handbook for Acoustic Ecology."; 2nd ed. Vancouver: A.R.C. Publication, 1999.

[11] Bregman, A.; "Auditory Scene Analysis."; MIT Press: Cambridge, 1990.

[12] Moore, B. C. M. "Hearing: Handbook of Perception and Cognition."; 2nd Ed. Academic Press:Toronto, 1995.

[13] Popper, A. N.; Fay, R. R.; Yost, W. A. "Human Psychophysics."; Springer-Verlag: New York, 1993.

[14] Gerhard, D. "Pitch Extraction and Fundamental Frequency: History and Current Techniques."; Technical Report TR-CS 2003-06, Department of Computer Science, University of Regina, 2003.

[15] Huang, X. "Spoken Language Processing: : A Guide to Theory, Algorithm and System Development."; Princeton Hall: New Jersey, 2001.

[16] Kondoz, A. M.; "Digital Speech:Coding for Low Bit Rate Communication Systems."; 2nd ed., John Wiley and Sons: 2004.

[17] Yang, Y.; Yang, M. "A Pitch-Based Rapid Speech Segmentation for Speaker Indexing.", in Proc. of IEEE International Symposium on Multimedia (ISM'05) 2005, 571-576.

[18] Woods, R. E.; Gonzalez, R. C. "Digital Image Processing."; 3rd ed., Prentice Hall: 2008.

[19] Anguera, X. "XBIC: Real-Time Cross Probabilities Measure for Speaker Segmentation."; ICSI Technical Report TR-05-008, International Computer Science Institute, 2005.