Paramedical Sciences and Military Health
Volume 10, Number 2, Autumn 2015 (Serial Number 18)

Original Article

Designing a New Non-parallel Training Method to Voice Conversion with Better Performance than Parallel Training

Mostafa Ghorbandoost, Valiallah Saba*

Department of Radiology, Faculty of Paramedicine, AJA University of Medical Sciences, Tehran, Iran

Abstract

Introduction: The art of voice mimicking by computers, has with the computer have been one of the most challenging topics of speech processing in recent years. The system of voice conversion has two sides. In one side, the speaker is the source that his or her voice has been changed for mimicking the target speaker's voice (which is on the other side). Two methods of parallel and non-parallel training are used for voice conversion. In parallel method, both source and target speakers express the same sentences while different sentences are expressed in non-parallel method. Most of the voice conversion researchers prefer to use parallel training; however, there is not always the possibility of collecting parallel data. Therefore, there is a need for using non-parallel methods.

Methods and Materials: Source and target speakers' voice was recorded and then analyzed. Voice features of both speakers were extracted by signal processing. Then the action of alignment has been done and the function of voice conversion was obtained. Source voice has been analyzed and the action of extracting feature has been carried out in order to convert source voice to the target. Voice conversion function from the previous section was applied on the extracted features. Then, the reverse action of features was done and finally, the voice synthesis took place. Moreover, the synthesized voice is the voice of target person

Results: The results of both numerical and objective experiments demonstrated that our proposed method is better than parallel training methods. It was observed that this superiority holds for different sizes of training material from 5 to 40 training sentences, both in terms of quality and similarity to the target speaker.

Discussion and Conclusion: It seems that our proposed method is a serious competitor of parallel training method for from alignment.

Keywords: Voice conversion, Speech analysis\synthesis, Non-parallel training system, INCA algorithm, Gaussian Mixture Model (GMM), Universal Background Model (UBM), Real-time voice conversion

* (Corresponding Author) Valiallah Saba, Department of radiology, Faculty of paramedicine, AJA university of medical sciences, Etemad zade street, Fatemi Street, Tehran, Iran, Postal code: 1411718541; tel: +98-21-43822449.

مقاله تحقیقاتی

مجله علوم پیراپزشکی و بهداشت نظامی
سال دهم، شماره دوم، پاییز ۱۳۹۴، مسلسل ۱۸ صفحات ۷ تا ۱۶

طراحی یک روش آموزش ناموازی جدید برای تبدیل گفتار با عملکردی بهتر از آموزش موازی

مصطفی قرباندوست[۱]، ولی اله صبا[۱]*

[۱] گروه تکنولوژی پرتوشناسی، دانشکده پیراپزشکی، دانشگاه علوم پزشکی آجا، تهران، ایران

چکیده

مقدمه: هنر تقلید صدای انسان با کامپیوتر، یکی از چالش برانگیزترین موضوعات پردازش گفتار در سال‌های اخیر بوده است. یک سیستم تبدیل گفتار دارای دو سمت است. در یک سمت آن، گوینده مبدا قرار دارد که صدایش برای تقلید صدای گوینده هدف (که در سمت دیگر سیستم قرار دارد) تغییر داده می‌شود. برای تبدیل گفتار فرد مبدا به فرد هدف از دو روش موازی و ناموازی استفاده می‌شود. در روش موازی گوینده مبدا و هدف جملات یکسانی بیان کرده و در روش ناموازی جملات متفاوتی بیان می‌کنند. بیشتر محققین تبدیل گفتار برای آموزش تابع تبدیل از دادگان آموزشی موازی استفاده کرده‌اند. با این حال، در عمل همیشه امکان جمع‌آوری دادگان موازی وجود ندارد و بنابراین نیاز استفاده از روش‌های ناموازی به وجود می‌آید.

مواد و روش‌ها: گفتار گوینده مبدا و هدف ضبط شده و سپس مورد آنالیز قرار گرفت. با پردازش سیگنال، ویژگی‌های گفتار هر دو نفر استخراج شد. سپس عمل هم ردیف سازی انجام شده و تابع تبدیل گفتار بدست آمد. برای تبدیل گفتار مبدا به هدف، گفتار مبدا آنالیز شده و سپس عمل استخراج ویژگی انجام شد. تابع تبدیل گفتار بدست آمده از قسمت قبل، بر ویژگی‌های استخراج شده اعمال شد. سپس عمل معکوس استخراج ویژگی انجام شده و در پایان سنتز گفتار صورت گرفت. صدای سنتز شده، صدای فرد هدف می‌باشد.

یافته‌ها: نتایج آزمایش‌های عددی و عینی مشخص کرد که روش پیشنهادی ما از روش آموزش موازی بهتر است. همچنین در آزمایش‌ها مشاهده شد که این برتری هم از لحاظ کیفیت و هم از لحاظ شباهت به گوینده‌ی هدف، برای اندازه‌های مختلف دادگان آموزشی از پنج تا چهل جمله صادق است.

بحث و نتیجه‌گیری: به نظر می‌رسد که روش پیشنهادی ما یک رقیب جدی برای روش‌های آموزش موازی برای همردیف سازی فریم است.

کلمات کلیدی: تبدیل گفتار، آنالیز و سنتز صدا، سیستم‌های آموزش ناموازی، الگوریتم INCA، مدل مخلوط گاوسی، مدل پس زمینه سراسری، تبدیل گفتار بلادرنگ

مقدمه

تبدیل گفتار، هنر تقلید صدای انسان با کامپیوتر، یکی از چالشی‌ترین موضوعات پردازش گفتار در سال‌های اخیر بوده است. یک سیستم تبدیل گفتار دارای دو سمت است. در یک سمت آن، گوینده مبدأ قرار دارد که صدایش برای تقلید صدای گوینده هدف (که در سمت دیگر سیستم قرار دارد) تغییر داده می‌شود. عملکرد یک سیستم تبدیل گفتار به کیفیت (طبیعی بودن) و شباهت (به گوینده‌ی هدف) صدای تبدیل شده‌ی آن بستگی دارد. بسته به اینکه گوینده‌های

* (نویسنده مسئول) ولی اله صبا، گروه تکنولوژی پرتوشناسی، دانشکده پیراپزشکی، دانشگاه علوم پزشکی آجا
شماره تماس: ۰۲۱۸۲۸۸۲۴۴۹

مبدأ و هدف جملات آموزشی یکسان یا متفاوتی را ادا کرده باشند، روش‌های تبدیل گفتار به ترتیب به روش‌های با دادگان موازی یا ناموازی تقسیم می‌شوند. بیشتر محققین این رشته ترجیح می‌دهند که از آموزش موازی استفاده کنند تا بتوانند روی دقت تابع نگاشت (رگرسیون) تمرکز کنند. به طور حتم، جمع‌آوری دادگان موازی برای همه‌ی سناریوهای عملی امکان پذیر نیست، بنابراین طراحی روش‌هایی که بتوانند با دادگان ناموازی کار کنند ضروری است.

در روش‌های آموزش موازی، همردیف سازی فریم‌های گویندگان مبدأ و هدف، با اعمال الگوریتم پیچش زمانی پویا (DTW) به جفت‌های جملات متناظر آنها صورت می‌گیرد. در مرحله‌ی بعد، یک تابع تبدیل دلخواه از روی زوج ویژگی‌های جفت شده، تخمین زده می‌شود. چند نمونه از توابع تبدیل عبارتند از: نگاشت مبتنی بر مدل مخلوط گوسی (GMM) (۱)، رگرسیون کمترین مربعات جزئی با هسته‌ی پویا(DKPLS) (۲)، سیستم‌های دینامیکی خطی (LDS) (۳) و شبکه‌های عصبی عمیق (۴). بحث در مورد مزیت‌ها و کمبودهای این روش‌ها در این تحقیق نمی‌گنجد، چون که تحقیق ما روی روش‌های آموزش ناموازی متمرکز است. همچنین ویژگی‌های طیفی مختلفی برای تشکیل تابع تبدیل مورد استفاده قرار می‌گیرند. بررسی کامل و مقایسه‌ی این ویژگی‌های طیفی، در تحقیق اخیر ما (۵) ارائه شده است.

روش‌های تبدیل گفتار ناموازی، به دو دسته تقسیم می‌شوند. دسته اول روش‌هایی هستند که سعی نمی‌کنند فریم‌های ناموازی مبدأ و هدف را همردیف کنند. آنها مستقیما از داده‌های مبدأ و هدف برای ساختن یا تطبیق تابع تبدیل استفاده می‌کنند. در اینجا ما به طور اختصار در مورد برخی از این روش‌ها بحث می‌کنیم:

در مقاله‌ی (۶)، یک مدل مخفی مارکف (HMM) برای گوینده‌ی هدف و یک GMM برای گوینده‌ی مبدأ ساخته می‌شود. تابع تبدیل به صورت مخلوط تبدیلات خطی ساده فرض می‌شود و با بیشینه کردن احتمال بردارهای تبدیل شده‌ی مبدأ نسبت به HMM هدف، آموزش داده می‌شود. این روش جذاب است اما در عین حال کمبودهایی دارد. محدودیت اصلی آن، نیاز به اطلاعات آوایی برای دادگان گفتاری است. مشکل دیگر وقتی به وجود می‌آید که الگوریتم بیشینه سازی انتظار (EM) اعمال می‌شود. در هر حالت از HMM هدف، یک GMM وجود دارد که برای تعیین احتمالات پسین اولیه‌ی این GMMها برای هر بردار تبدیل شده، بردار مبدأ به آنها داده می‌شود. این آغازسازی، دقت تخمین تابع تبدیل را کاهش می‌دهد. دلیل این است که وقتی فضای آکوستیکی مبدأ از فضای آکوستیکی هدف دور است (برای مثال در مورد تبدیل مرد به زن)، این آغازسازی فضاهای آکوستیکی را با هم مخلوط می‌کند.

روش (۷) فرض می‌کند که علاوه بر دادگان ناموازی، مقدار کمی داده‌ی موازی نیز موجود است. بنابراین، این روش، یک تابع تبدیل مبتنی بر GMM شبیه به (۱) می‌سازد. در مرحله‌ی تبدیل، ابتدا دنباله‌ی فریم‌های تست مبدأ با این تابع، تبدیل می‌شوند. سپس، این فریم‌های نیمه تبدیل یافته، با بهترین دنباله‌ی فریم‌های منطبق خود در دادگان آموزشی گوینده‌ی هدف، جایگزین می‌شوند. با این کار، فریم‌های تبدیل شده‌ی نهایی به دست می‌آیند. روشی که برای انتخاب بهترین دنباله‌ی فریم‌های منطبق استفاده می‌شود، انتخاب واحد (Unit Selection) است که از TTS قرض گرفته شده است. واضح است که محدودیت اصلی روش (۷)، نیاز آن به مقداری داده‌ی آموزشی موازی است. بنابراین، ما این گونه از روش‌ها را نیمه–ناموازی می‌نامیم.

این نکته را باید مد نظر قرار داد که تمامی روش‌های آموزش ناموازی مطرح شده در بالا یک ویژگی مشترک دارند. آنها قادرند فقط یک تابع تبدیل خاص را آموزش دهند و این استفاده‌ی آنها را محدود می‌کند. دسته‌ی دوم روش‌ها، روش‌هایی هستند که تلاش می‌کنند دادگان (فریم‌های) ناموازی مبدأ و هدف را همردیف کنند. بنابراین بر خلاف دسته‌ی اول، آنها یک تابع تبدیل خاص برای خودشان فرض نمی‌کنند و هر تابع تبدیل دلخواهی می‌تواند بعد از همردیف سازی فریم‌ها آموزش داده شود. در اینجا به طور اختصار در مورد بعضی از این روش‌ها بحث می‌شود:

در مقاله‌ی (۸)، فریم‌های مبدأ و هدف به صورت جداگانه خوشه بندی می‌شوند. سپس با محاسبه‌ی فاصله‌ی اقلیدسی بین مراکز خوشه‌ها (بعد از اعمال پیچش فرکانسی به مراکز)، خوشه‌های منطبق با هم یافت می‌شوند. بعد از آن، فریم‌های نرمالیزه شده با میانگین زوج خوشه‌های منطبق، با اعمال الگوریتم جستجوی نزدیکترین همسایه با هم جفت می‌شوند. این روش یک نقطه‌ی شروع برای روش‌هایی بود که تلاش می‌کنند همردیف سازی فریم‌ها را انجام دهند.

در مقاله‌ی (۹)، فریم‌های مبدأ با استفاده از الگوریتم انتخاب واحد با بهترین دنباله‌ی فریم‌ها از دادگان گوینده‌ی هدف، همردیف می‌شوند. متأسفانه وقتی اندازه پایگاه داده‌ی هدف بزرگ می‌شود، عملکرد این روش کاهش می‌یابد. دلیل این است که فریم‌های مبدأ تمایل دارند فریم‌های هدف شبیه به خود را بیابند، ولی وقتی اندازه پایگاه داده‌ی هدف بزرگ می‌شود، این شباهت زیاد می‌تواند منجر به تشکیل تابع تبدیلی شود که فریم‌های مبدأ را به خودشان نگاشت می‌کند. این مشکل وقتی جدی‌تر می‌شود که فضاهای آکوستیکی دو گوینده از هم دور باشند، برای مثال در تبدیلات مرد به زن. روش (۱۰) هم از الگوریتم انتخاب واحد برای همردیف کردن فریم‌ها استفاده می‌کند اما قبل از انجام همردیفی، دنباله‌ی حالات HMM مبدأ و هدف را با استفاده از یک سیستم بازشناسی گفتار مستقل از گوینده می‌یابد. بنابراین همردیفی بین حالات متناظر انجام می‌شود. روش مورد بحث، نویدبخش‌تر از روش (۹) است ولی میزان دقت سیستم بازشناسی گفتار، عملکرد آن را محدود می‌کند.

الگوریتم همردیفی با ترکیب مکرّر مرحله‌ی جستجوی نزدیکترین همسایه و مرحله‌ی تبدیل (INCA) (۱۱)، آخرین روش مبتنی بر همردیف سازی فریم مورد بحث در اینجاست. جزئیات این روش در فصل ۲ مورد بررسی قرار می‌گیرد. این روش، یک روش بدون نظارت مؤثر و ساده است، اما وقتی فضاهای آکوستیکی مبدأ و هدف متفاوت باشند (مانند تبدیلات مرد به زن)، این روش با مشکل آغازسازی مواجه می‌شود. این مشکل در مرحله‌ی اول الگوریتم جستجوی نزدیکترین همسایه به وجود می‌آید و در تمام مراحل بعدی انتشار پیدا می‌کند و در نهایت منجر به همردیف سازی فریم غیر دقیق می‌شود.

در این تحقیق ما یک روش آموزش ناموازی جدید برای تبدیل گفتار پیشنهاد می‌دهیم که هیچ کدام از مشکلات روش‌های بیان شده در بالا را ندارد. این روش بدون نظارت است و بنابراین به برچسب‌های آوایی نیازی ندارد. همچنین، نیازمند هیچ دادگان موازی از پیش ذخیره شده‌ای نیست. این روش همردیف سازی را در سطح فریم انجام می‌دهد، بنابراین هر تابع تبدیل دلخواهی بعد از همردیف سازی می‌تواند آموزش داده شود. و در نهایت به عنوان مهمترین ویژگی، این روش همردیف سازی را با نسخه‌ی تبدیل شده‌ی بردارهای ویژگی مبدأ آغاز سازی می‌کند. این آغاز سازی بی بدیل، به طور مؤثری مشکل همردیف سازی اشتباه در آغاز سازی را کاهش می‌دهد. آغاز سازی مخصوص ما، همردیف سازی مبتنی بر پس زمینه (BAM) نام دارد و از مدل‌های پس زمینه‌ی جهانی (UBM) استفاده می‌کند که از تحقیقات رشته‌ی تأیید گوینده قرض گرفته شده است (۱۲). روش کامل ما شامل آغازسازی BAM به علاوه‌ی الگوریتم INCA است و ما آن را الگوریتم (BAMI+INCA) BAMINCA می‌نامیم. در واقع، ما از BAM برای عنوان آغازسازی INCA استفاده می‌کنیم و مشکل همردیف سازی اشتباه در آغاز سازی را کاهش می‌دهیم. نتایج آزمایشات عینی و عددی مشخص می‌کنند که روش پیشنهادی ما (BAMINCA) نه تنها بر الگوریتم INCA، بلکه بر آموزش موازی مبتنی بر DTW نیز غلبه می‌کند. آزمایشات عینی این بهبودها را برای روش ما تأیید می‌کنند.

الگوریتم INCA

فرض کنید که ما دو مجموعه‌ی بردار ویژگی $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_T]$ و $\mathbf{Y}=[\mathbf{y}_1,\mathbf{y}_2,...,\mathbf{y}_N]$ داریم که به ترتیب فریم‌های آموزشی مبدأ و هدف را نشان می‌دهند. هدف الگوریتم INCA همردیف کردن اعضای $\mathbf{X}$ و $\mathbf{Y}$ است. الگوریتم INCA شامل پنج مرحله است (۱۱):

۱) آغازسازی: فرض کنید که ما یک مجموعه از بردارهای ویژگی کمکی $\mathbf{X}'=[\mathbf{x}'_1,\mathbf{x}'_2,...,\mathbf{x}'_T]$ داریم که عناصر آن متناظر با مجموعه‌ی مبدأ $\mathbf{X}$ هستند (اعضای $\mathbf{X}$ و $\mathbf{X}'$ اندیس‌های مشترکی دارند). این بردارهای ویژگی میانی به صورت $\mathbf{x}'_t=\mathbf{x}_t$ آغازسازی می‌شوند.

۲) همردیف سازی نزدیک‌ترین همسایه: برای هر عضو $\mathbf{x}'_t$ از مجموعه‌ی $\mathbf{X}'$، نزدیک‌ترین همسایه اش $\mathbf{y}_{NN}$ را از مجموعه‌ی $\mathbf{Y}$ بیابید و زوج‌های $\mathbf{x}_t,\mathbf{y}_{NN}$ را در مجموعه‌ی جدید $\mathbf{XY}_{NN}$ قرار دهید. همچنین برای هر عضو $\mathbf{y}_t$ از مجموعه‌ی $\mathbf{Y}$، نزدیک‌ترین همسایه اش $\mathbf{x}'_{NN}$ از مجموعه ی $\mathbf{X}'$ را بیابید و زوج‌های $\mathbf{x}_{NN}\mathbf{y}_t$ را در مجموعه‌ی جدید $\mathbf{X}_{NN}\mathbf{Y}$ قرار دهید. مجموعه‌های $\mathbf{XY}_{NN}$ و $\mathbf{X}_{NN}\mathbf{Y}$ را با هم تلفیق کنید و زوج‌های تکراری حاضر در هر دو مجموعه را حذف کنید.

۳) آموزش: یک GMM مشترک با پارامترهای (α,μ,Σ) برای بردارهای ویژگی تلفیق شده از مرحله‌ی قبل آموزش دهید. حال، یک تابع تبدیل کمکی با اعمال معیار خطای میانگین

کمترین مربعات (MMSE) به این GMM به دست می‌آید

$$F_{aux}(\mathbf{x}) = \sum_{i=1}^{M} P(C_i \mid \mathbf{x})[\boldsymbol{\mu}_i^{\,y} + \boldsymbol{\Sigma}_i^{\,yx}\boldsymbol{\Sigma}_i^{\,xx^{-1}}(\mathbf{x} - \boldsymbol{\mu}_i^{\,x})], \quad \boldsymbol{\mu}_i = \begin{bmatrix} \boldsymbol{\mu}_i^{x} \\ \boldsymbol{\mu}_i^{y} \end{bmatrix}, \boldsymbol{\Sigma}_i = \begin{bmatrix} \boldsymbol{\Sigma}_i^{xx} & \boldsymbol{\Sigma}_i^{xy} \\ \boldsymbol{\Sigma}_i^{yx} & \boldsymbol{\Sigma}_i^{yy} \end{bmatrix}, \quad (۱)$$

که $P(C_i \mid \mathbf{x})$ و M به ترتیب نشان دهنده‌ی احتمال پسین i اُمین مخلوط و تعداد مخلوط‌ها هستند. همچنین، $\boldsymbol{\mu}_i$ و Σ_i به ترتیب بردار متصل میانگین و ماتریس کوواریانس i اُمین مخلوط هستند.

۴) تبدیل: مجموعه‌ی بردارهای ویژگی کمکی $\mathbf{X}'$ را با اعمال تابع F_{aux} به هر عضو $\mathbf{x}_t$ از $\mathbf{X}$ به روز کنید:

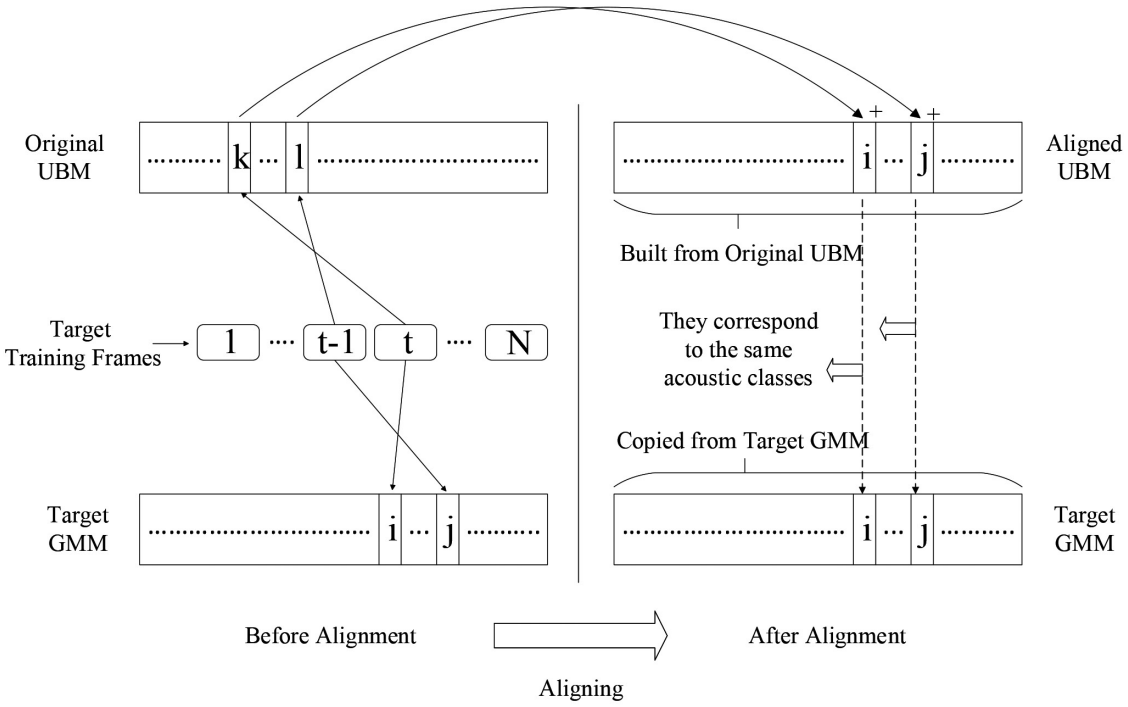
$$\mathbf{x}'_t = F_{aux}(\mathbf{x}_t)$$

۵) بررسی همگرایی: فاصله‌ی میانگین بین بردارهای ویژگی تبدیل شده از مرحله‌ی قبل و نزدیک‌ترین همسایه‌های متناظر با آنها از مجموعه‌ی $\mathbf{Y}$ را که در مرحله‌ی ۲ پیدا شده بودند، بیابید. اگر فاصله‌ی میانگین از یک مقدار آستانه کمتر شد، الگوریتم را متوقف کنید. حال، بردارهای همردیف شده‌ی مطلوب $\mathbf{X}$ و $\mathbf{Y}$ آنهایی هستند که در مرحله‌ی ۲ به دست آمدند. اگر شرط آستانه ارضا نشده است، به مرحله‌ی ۲ برگردید و الگوریتم را از آنجا تکرار کنید، تا وقتی که شرط آستانه ارضا شود.

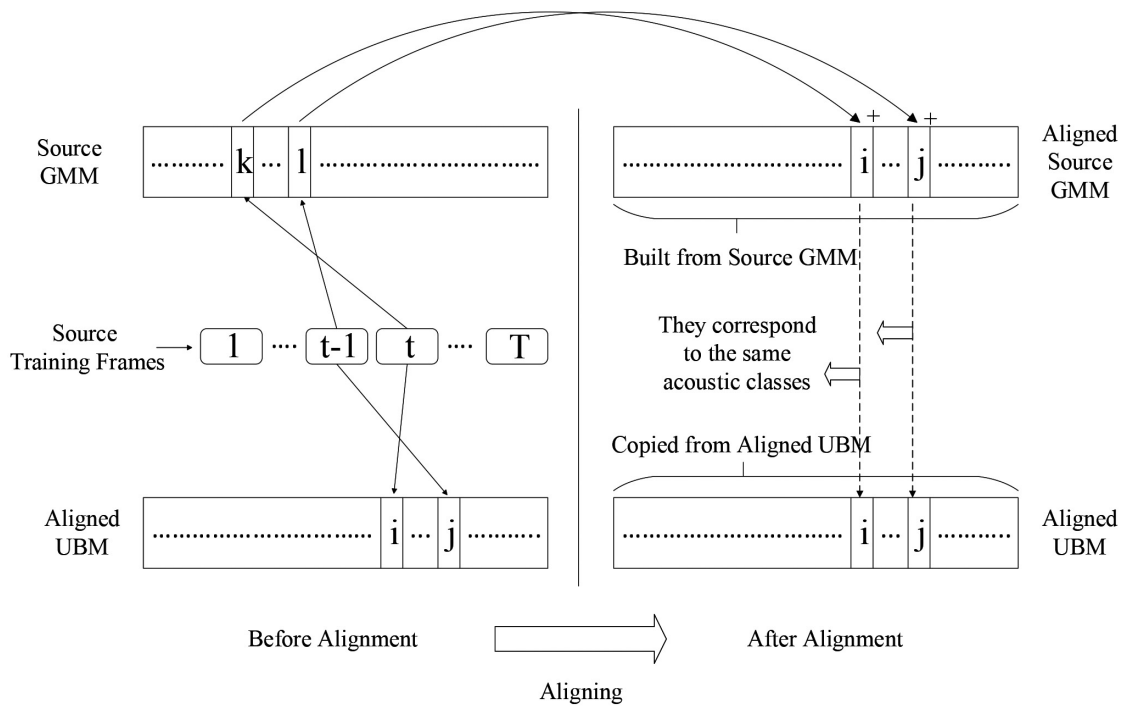
همردیف سازی مبتنی بر پس زمینه (BAM)

حال زمان آن فرا رسیده است که ما ایده‌ی پیشنهادی مان برای همردیف سازی مدل را که در بخش بعد برای آغازسازی الگوریتم INCA مورد استفاده قرار می‌گیرد، توصیف کنیم. فرض کنید که بردارهای آموزشی ناموازی مبدأ و هدف به ترتیب $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_T]$ و $\mathbf{Y}=[\mathbf{y}_1,\mathbf{y}_2,...,\mathbf{y}_N]$ باشند. ما دو GMM با M مخلوط (با ماتریس‌های کوواریانس قطری) به صورت مجزا برای گویندگان مبدأ و هدف آموزش می‌دهیم. هدف BAM همردیف سازی این دو GMM مجزا است، به طوری که مخلوط‌های متناظرشان دسته‌های آوایی شبیه به هم را نمایندگی کنند. برای انجام همردیفی، ما از یک مدل پس زمینه‌ی کمکی (UBM) با تعداد مخلوط های M استفاده می‌کنیم، که با دادگان گفتاری ناموازی از تعداد زیادی گوینده‌ی مرد و زن از پیش ذخیره شده، آموزش داده شده است. الگوریتم BAM از یک تئوری معروف هندسه‌ی اقلیدسی استفاده می‌کند. این تئوری بیان می‌کند که اگر دو خط با خط سومی موازی باشند، آنگاه آن دو خط با هم نیز موازی اند. بنابراین، مراحل ایده‌ی ما به این صورت است. ابتدا GMM هدف را با UBM همردیف می‌کنیم. سپس UBM را با GMM مبدأ همردیف می‌کنیم. قدم آخر بسیار ساده است. چون که حالا GMMهای مبدأ و هدف با UBM موازی هستند، بنابراین با هم نیز موازی اند. حال سؤال این است که: چگونه باید GMM هدف را با UBM و UBM را با GMM مبدأ همردیف کنیم؟

پروسه‌ی همردیفی GMM هدف با UBM با استفاده از فریم‌های آموزشی هدف $\mathbf{Y}$، در شکل ۱ نشان داده شده است. در اینجا ما GMM هدف را تغییر نمی‌دهیم و به جای آن، یک UBM همردیف شده از UBM اصلی می‌سازیم. توجه کنید که UBM همردیف شده در ابتدای الگوریتم خالی می‌باشد. این بدان معناست که پارامترهای آن (α,μ,Σ) برابر با صفر هستند. فرض کنید که احتمال پسین i اُمین و k اُمین مخلوط ($P(C_i \mid \mathbf{y}_t)$ و $P(C_k \mid \mathbf{y}_t)$) برای یک فریم هدف $\mathbf{y}_t$، به ترتیب در GMM هدف و UBM اصلی بیشینه هستند. این فریم با اضافه کردن k اُمین بردار میانگین و کوواریانس UBM اصلی به بردار میانگین و کوواریانس i اُمین مخلوط UBM همردیف شده، در ساختن UBM همردیف شده مشارکت می‌کند. این پروسه برای فریم $\mathbf{y}_{t-1}$ نیز نمایش داده شده است. این پروسه برای همه‌ی N فریم گوینده‌ی هدف تکرار می‌شود. در نهایت، بردار میانگین و کوواریانس مخلوط i اُم UBM همردیف شده، با تقسیم کردن مقادیر تجمیع شده‌ی آنها بر تعداد بردارهایی که به این مخلوط نسبت داده شده‌اند (N_i) به دست می‌آید. برای همردیف کردن UBM همردیف شده با GMM مبدأ با استفاده از فریم‌های آموزشی مبدأ $\mathbf{X}$، پروسه‌ی مشابهی مورد استفاده قرار می‌گیرد، اما این بار UBM همردیف شده تغییر نمی‌یابد و GMM مبدأ همردیف شده از روی GMM مبدأ ساخته می‌شود. شکل ۲ این پروسه را نمایش می‌دهد. حال GMM هدف و GMM مبدأ همردیف شده با UBM همردیف شده موازی هستند و با استفاده از اصل بیان شده‌ی بالا از هندسه‌ی اقلیدسی، GMM هدف و GMM مبدأ همردیف شده با هم نیز موازی اند. پس ما از این مرحله به بعد، دیگر نیازی به UBM همردیف شده نداریم و تنها GMM هدف و GMM مبدأ همردیف



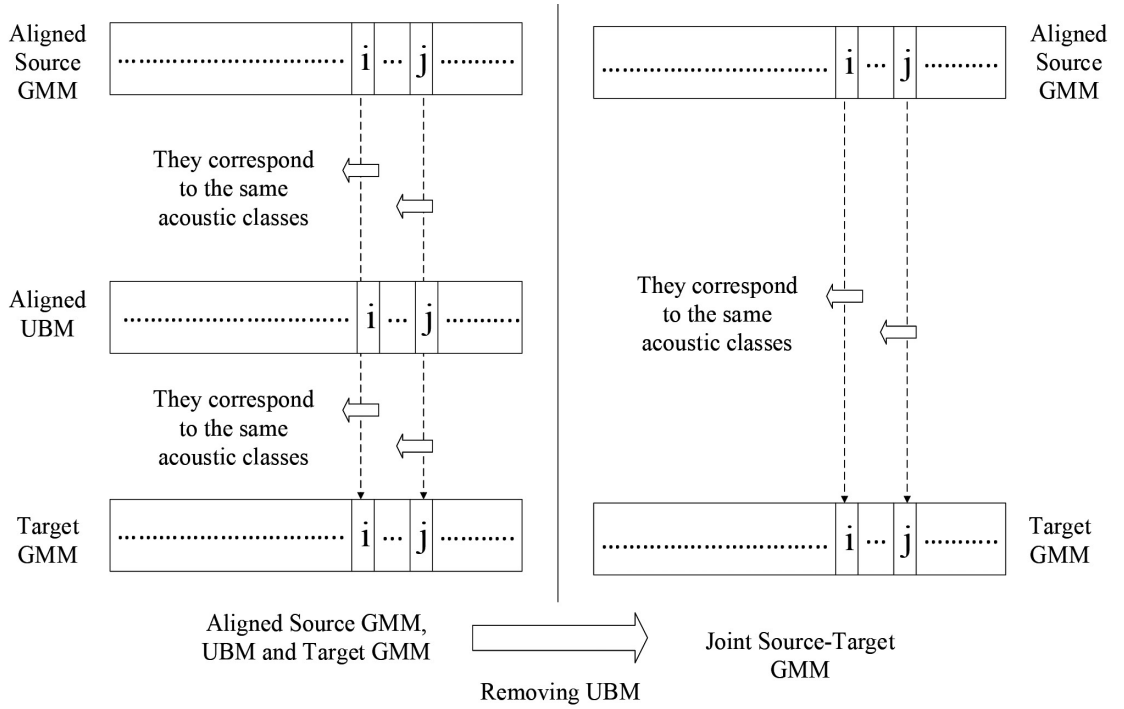
شکل ۱- هم‌ردیف سازی GMM هدف با UBM اصلی



شکل ۲- هم‌ردیف سازی GMM مبدأ با UBM هم‌ردیف شده

شده باقی می‌مانند. این در شکل ۳ نشان داده شده است. سرانجام، مدل‌های همردیف شده‌ی ما آماده هستند و ما مدل GMM مشترکمان را به دست آورده ایم. به علاوه، وزن مخلوط‌های (α) ی GMM مشترک از GMM هدف (که تغییری نکرده است) کپی می‌شوند.

الگوریتم همردیف سازی بالا از تصمیم‌گیری سخت برای نسبت دادن فریم‌ها به مخلوط‌ها استفاده می‌کند. الگوریتم پیچیده‌تر، مبتنی بر تصمیم‌گیری نرم است و نحوه‌ی به دست آوردن فرمول‌های آن از پروسه‌ی بالا سرراست است. در واقع مراحل تجمیع و تقسیم برای



شکل ۳- تشکیل GMM مشترک با حذف UBM هم‌ردیف شده

به دست آوردن بردارهای میانگین و کوواریانس i اُمین مخلوط UBM همردیف شده، در یک فرمول بیان می‌شوند:

$$\boldsymbol{\mu}_i^{A-UBM} = \frac{\sum_{t=1}^{N} P(C_i^{Taret} \mid \mathbf{y}_t) \sum_{k=1}^{M} P(C_k^{O-UBM} \mid \mathbf{y}_t) \boldsymbol{\mu}_k^{O-UBM}}{\sum_{t=1}^{N} P(C_i^{Taret} \mid \mathbf{y}_t)},$$

$$\boldsymbol{\Sigma}_i^{A-UBM} = \frac{\sum_{t=1}^{N} P(C_i^{Taret} \mid \mathbf{y}_t) \sum_{k=1}^{M} P(C_k^{O-UBM} \mid \mathbf{y}_t) \boldsymbol{\Sigma}_k^{O-UBM}}{\sum_{t=1}^{N} P(C_i^{Taret} \mid \mathbf{y}_t)},$$

که اندیس‌های $A-UBM$ و $O-UBM$ به ترتیب نمایشگر UBM همردیف شده و UBM اصلی هستند. برای تأیید این فرمول‌ها، ما مورد خاص تصمیم‌گیری سخت را بررسی می‌کنیم. در این مورد، معادلات بالا به معادلات زیر کاهش می‌یابند:

$$\boldsymbol{\mu}_i^{A-UBM} = \frac{\sum_{t=1}^{N_i} \boldsymbol{\mu}_{k(t)}^{O-UBM}}{N_i},$$

$$\boldsymbol{\Sigma}_i^{A-UBM} = \frac{\sum_{t=1}^{N_i} \boldsymbol{\Sigma}_{k(t)}^{O-UBM}}{N_i},$$

که $k(t)$ اندیس مخلوطی از UBM است که بیشترین احتمال پسین را برای فریم $\mathbf{y}_t$ دارد. واضح است که این دو فرمول با پروسه‌ی شکل ۱ که قبلا توضیحش دادیم، سازگار هستند. فرمول‌های مشابهی برای محاسبه‌ی بردارهای میانگین و کوواریانس i اُمین مخلوط GMM مبدأ همردیف شده وجود دارد:

$$\boldsymbol{\mu}_i^{A-Source} = \frac{\sum_{t=1}^{N} P(C_i^{A-UBM} \mid \mathbf{x}_t) \sum_{k=1}^{M} P(C_k^{Source} \mid \mathbf{x}_t) \boldsymbol{\mu}_k^{Source}}{\sum_{t=1}^{N} P(C_i^{A-UBM} \mid \mathbf{x}_t)},$$

$$\boldsymbol{\Sigma}_i^{A-Source} = \frac{\sum_{t=1}^{N} P(C_i^{A-UBM} \mid \mathbf{x}_t) \sum_{k=1}^{M} P(C_k^{Source} \mid \mathbf{x}_t) \boldsymbol{\Sigma}_k^{Source}}{\sum_{t=1}^{N} P(C_i^{A-UBM} \mid \mathbf{x}_t)},$$

$$\boldsymbol{\mu}_i^{A-Source} = \frac{\sum_{t=1}^{T_i} \boldsymbol{\mu}_{k(t)}^{Source}}{T_i},$$

$$\boldsymbol{\Sigma}_i^{A-Source} = \frac{\sum_{t=1}^{T_i} \boldsymbol{\Sigma}_{k(t)}^{Source}}{T_i},$$

توجه کنید که دلیل ما برای استفاده از UBM برای همردیف کردن GMMهای مبدأ و هدف این است که فضای آکوستیکی UBM، میانگین فضاهای آکوستیکی تمام صداهای موجود در یک زبان است. بنابراین

یک فریم مبدأ یا هدف می‌تواند به مخلوط‌های UBM به عنوان یک مدل میانگین، نسبت داده شود. قطعا نسبت دادن مستقیم فریم‌های مبدأ به مخلوط‌های GMM هدف یا بالعکس، عملی اشتباه است، چون با این کار فضاهای آکوستیکی با هم مخلوط می‌شوند. این، قدرت UBM است، که آن را قادر می‌سازد به عنوان یک مدل میانی برای هر دو گوینده‌ی مبدأ و هدف ایفای نقش کند. در انتهای این فصل، این نکته باید بیان شود که با توجه به بهبودهایی که با استفاده از همردیف سازی نرم نسبت به همردیف سازی سخت به دست می‌آیند، ما از این به بعد تنها از همردیف سازی نرم استفاده می‌کنیم. در واقع، همردیف سازی سخت واسطه‌ای برای ساده سازی و تأیید توضیح ما در مورد همردیف سازی نرم بود.

آغازسازی BAM به علاوه‌ی الگوریتم INCA (BAMINCA)

حال که الگوریتم INCA و الگوریتم همردیف سازی مدل پیشنهادی خودمان (BAM) را توصیف کرده ایم، می‌توانیم روش کامل آموزش ناموازی خودمان را معرفی کنیم. همانطور که در مراحل INCA دیدیم، برای یافتن نزدیکترین همسایه‌های بردارهای مبدأ $\mathbf{X}$ از مجموعه‌ی بردارهای هدف $\mathbf{Y}$، یک مجموعه‌ی بردار ویژگی کمکی $\mathbf{X}'$ مورد استفاده قرار می‌گیرد. اگر چه $\mathbf{X}'$ با تبدیل به روز می‌شود، اما آغازسازی اش به صورت $\mathbf{X}' = \mathbf{X}$ است. متأسفانه، این آغازسازی ساده باعث غیرهمردیفی‌های بسیار جدی در مرحله‌ی یافتن نزدیکترین همسایه می‌شود. به خصوص وقتی فضای آکوستیکی مبدأ و هدف از هم دور باشد (که در مورد تبدیلات مرد به زن و زن به مرد رخ می‌دهد)، اثر منفی این آغازسازی بر روی همردیف سازی، آشکارتر می‌شود. برای کاهش این مشکل، ما آغازسازی $\mathbf{X}'$ را به صورت نسخه‌ی تبدیل شده‌ی $\mathbf{X}$ پیشنهاد می‌دهیم. تابع تبدیل مورد استفاده برای آغازسازی، از خروجی BAM ساخته می‌شود، بنابراین ما روش پیشنهادی مان را آغازسازی BAM به علاوه‌ی الگوریتم INCA (BAMINCA) می‌نامیم. در فصل قبل توضیح دادیم که چگونه یک GMM مشترک از GMMهای مجزای مبدأ و هدف ساخته می‌شود. همانطور که از مقاله‌ی (۱) می‌دانیم، اگر GMM مشترکِ مبدأ و هدف را داشته باشیم، می‌توانیم با استفاده از معادله‌ی ۱، یک تبدیل از مبدأ به هدف بیابیم. برای استفاده از تابع تبدیل معادله‌ی ۱، باید مقادیر ماتریس‌های کواریانس متقابل Σ_i^{yx} را بدانیم، اما متأسفانه GMM مشترکِ حاصل از BAM، شامل این پارامترها نمی‌شود. این GMM مشترک، تنها شامل بردارهای میانگین و ماتریس‌های کوواریانس مبدأ و هدف است. برای غلبه بر این مشکل، از یک تابع تبدیل که با اندکی تغییر به دست آمده، برای آغازسازی استفاده می‌کنیم:

$$F_{init}(\mathbf{x}) = \sum_{i=1}^{M} P(C_i \mid \mathbf{x})[\boldsymbol{\mu}_i^{\,y} + \boldsymbol{\Sigma}_i^{yy}\boldsymbol{\Sigma}_i^{xx^{-1}}(\mathbf{x} - \boldsymbol{\mu}_i^{\,x})].$$

این معادله جمع وزن دار تبدیل‌های نرمالیزاسیون گوسی در هر مخلوط است. پس پیشنهاد BAMINCA این است که برای آغازسازی $\mathbf{X}'$، آن را برابر $\mathbf{X}$ قرار ندهیم، بلکه آن را برابر نسخه‌ی تبدیل یافته‌ی $\mathbf{X}$ که توسط معادله‌ی بالا به دست می‌آید، قرار دهیم. بقیه‌ی مراحل BAMINCA دقیقا مشابه با مراحل ۲ تا ۵ INCA است.

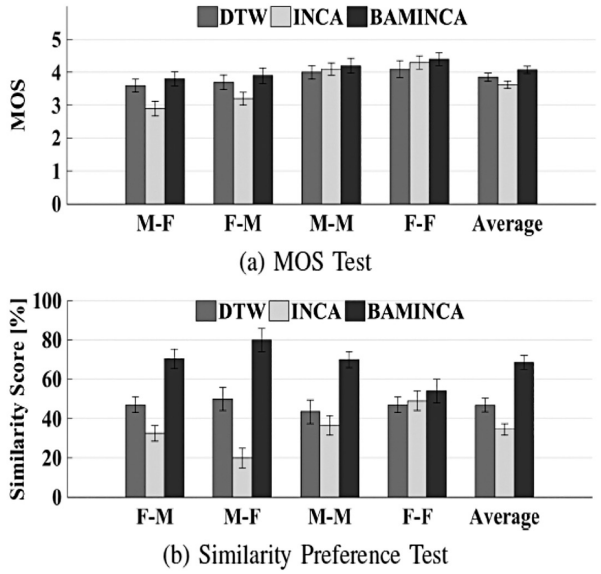
ارزیابی‌های آزمایشگاهی

در این فصل ما عملکرد روش آموزش ناموازی پیشنهادی خودمان (BAMINCA) را با الگوریتم آموزش ناموازی INCA و آموزش موازی مبتنی بر DTW مقایسه می‌کنیم. برای مقایسه‌ی این روش‌ها، آزمایش‌های عینی طراحی شده‌اند. برای آموزش مدل پس زمینه‌ی BAMINCA از پایگاه داده‌ی مشهور TIMIT استفاده شده است که ۱۰۰ گوینده‌ی مرد و ۱۰۰ گوینده‌ی زنِ این پایگاه داده برای آموزش دادن UBM مورد استفاده قرار گرفته‌اند. همچنین دو گوینده‌ی مرد (BDL و RMS) و دو گوینده‌ی زن (CLB و SLT) از پایگاه داده‌ی CMU ARCTIC برای آموزش و تست سیستم تبدیل گفتار مورد استفاده قرار گرفته‌اند. ۱۲ زوج تبدیل مختلف بین این چهار گوینده وجود دارد که آزمایش‌های ما تمام آنها را در بر می‌گیرد. تعداد جملات آموزشی بین ۵ تا ۴۰ جمله متغیر است. مجموعه‌ای از ۱۰ جمله بررای تست سیستم مورد استفاده قرار می‌گیرد که از مجموعه‌ی جملات آموزشی متمایز است. وُکُدِر STRAIGHT (۱۳) با شیفت فریم ۵ میلی ثانیه به عنوان الگوریتم آنالیز و سنتز مورد استفاده قرار می‌گیرد. ۲۴ ضریب MCC از طیف STRAIGHT به عنوان بردار ویژگی استخراج می‌شوند. برای آموزش تابع تبدیل اصلی، روش اخیر تبدیل گفتار پیشنهادی ما (۳) که مبتنی بر LDS با در نظر گرفتن واریانس سراسری (GV) است، مورد استفاده قرار می‌گیرد. دلیل اول ما برای استفاده از این روش، عملکرد قابل رقابت آن با روش‌هایی

مشهوری چون MLE+GV است. دلیل دیگر این است که در این روش، نیازی به تغییر k (تعداد حالت‌های مخفی) با تغییر اندازه‌ی دادگان آموزشی نیست. پس برای تعیین مقدار بهینه‌ی پارامترهای LDS، نیازی نیست که آن را چندین بار آموزش دهیم. همانطور که در مقاله‌ی (۳) بیان شده است، مقدار بهینه‌ی k برای تمامی اندازه‌های داده‌های آموزشی، $k = 50$ است. برای آموزش GMMها، تعداد تکرارهای الگوریتم بیشینه کردن انتظار (EM) برابر با ۱۰ است و از الگوریتم k-means برای آغازسازی پارامترهای GMM استفاده شده است. فرکانس گام گوینده‌ی مبدأ با استفاده از نرمالیزاسیون گوسی در حوزه‌ی لگاریتم (۱) به فرکانس گام گوینده‌ی هدف تبدیل می‌شود که در تحقیقات تبدیل گفتار بسیار معمول است.

برای قضاوت و مقایسه‌ی عملکرد BAMINCA، INCA و DTW به صورت عینی، ما تست‌های ترجیح شباهت و امتیاز نظر میانگین (MOS) را انجام می‌دهیم. ۱۰ گوینده در آزمایش‌ها شرکت می‌کنند. تعداد جملات تست نیز برابر با ۱۰ است. در تست MOS، هر شنونده به طور تصادفی به یک جمله‌ی تبدیل شده با یکی از سه روش گوش می‌دهد و با توجه به طبیعی بودن آن، به آن امتیازی از ۱ تا ۵ می‌دهد. در تست ترجیح شباهت، هر شنونده ابتدا به جمله‌ی اصلی گوینده‌ی هدف گوش می‌دهد و پس از گوش دادن به جمله‌ی تبدیل شده با دو روش، صدایی که شباهت بیشتری به هدف دارد را انتخاب می‌کند. توجه کنید که در یک تست MOS، سه روش یکجا با هم مقایسه می‌شوند، بنابراین یک تست برای قضاوت آنها کافی است. با این حال، در یک تست ترجیح شباهت، فقط دو روش می‌توانند با هم مقایسه شوند و در نتیجه برای قضاوت سه روش، باید سه تست انجام شود. بنابراین نتایج نهایی امتیازهای ترجیح شباهت، میانگین سه تست هستند. شکل ۴ نتیجه‌ی تست‌های MOS و ترجیح شباهت را برای تمام زوج‌های تبدیل با ۱۰ جمله‌ی آموزشی نشان می‌دهد. با توجه به شکل می‌توان متوجه شد که در تمام زوج‌های تبدیل، BAMINCA بر INCA و DTW غلبه کرده است. همچنین مشخص است که تفاوت عملکرد روش‌ها در تبدیل‌های مرد به زن و زن به مرد، بسیار قابل توجه‌تر از تبدیل‌های مرد به مرد و زن به زن است و این کاملا منطبق با نتایج عددی ماست.

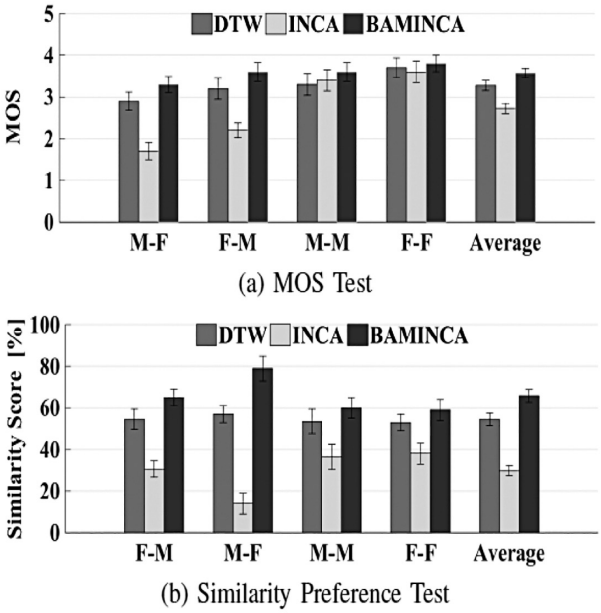
به علاوه غیر از تبدیل مرد به مرد که عملکرد INCA از لحاظ کیفیت (و نه هویت) کمی بهتر از DTW است، DTW همیشه بهتر از INCA



شکل ۴- نتایج تست‌های MOS و ترجیح شباهت برای چهار زوج تبدیل با ۱۰ جمله‌ی آموزشی

عمل می‌کند. این پدیده عجیب نیست چرا که همانطور که از نتایج عددی می‌دانیم، عملکرد سه روش در تبدیل‌های مرد به مرد و زن به زن بسیار نزدیک به هم است و این نزدیکی در تست‌های عددی، نزدیکی در تست‌های عینی را نیز القا می‌کند.

نتایج عینی بالا برای ۱۰ جمله‌ی آموزشی بودند. با افزایش تعداد جملات آموزشی از ۱۰ به ۴۰، نتایجی حاصل می‌شوند که در شکل ۵ نشان داده شده‌اند. برتری BAMINCA بر INCA و DTW با ۴۰ جمله‌ی آموزشی نیز مشخص است. اگرچه اختلاف‌ها در نمودار MOS در تبدیل‌های مرد به زن و زن به مرد نسبت به MOS با ۱۰ جمله‌ی آموزشی کمتر شده است ولی عملکرد عالی BAMINCA در تست ترجیح شباهت تغییری نکرده است. فرق دیگر با ۱۰ جمله‌ی آموزشی، عملکرد بهتر INCA نسبت به DTW از جهت شباهت و کیفیت در تبدیل زن به زن است. برای توضیح این پدیده از این واقعیت استفاده می‌کنیم که در این مورد به دلیلی نزدیکی بیشتر فضاهای آکوستیکی دو زن به هم (نسبت به تبدیل‌های مرد به مرد، زن به مرد و مرد به مرد)، آغازسازی ساده‌ی INCA عملکرد آن را کاهش نمی‌دهد. در حقیقت، وقتی میزان داده‌ی آموزشی کافی باشد (مثلا ۴۰ جمله) و فضاهای آکوستیکی دو گوینده به اندازه‌ی کافی به هم نزدیک باشند (مثلا CLB و SLT)، انتظار می‌رود که عملکرد INCA با DTW قابل مقایسه شود.



شکل ۵- نتایج تست‌های MOS و ترجیح شباهت برای چهار زوج تبدیل با ۴۰ جمله‌ی آموزشی

بحث و نتیجه‌گیری

در این مقاله ما یک روش آموزش ناموازی مبتنی بر همردیفی فریم برای تبدیل گفتار ارائه دادیم. ما از الگوریتم مشهور INCA برای ارائه‌ی روش جدیدمان استفاده کردیم. در واقع ما با معرفی یک آغازسازی جدید به نام BAM، عملکرد INCA را بهبود دادیم. در نتیجه روش کامل ما BAMINCA نام گرفت. این آغازسازی مبتنی بر همردیف سازی مدل با مدل‌های پس زمینه است و به طور مؤثری مشکل تفاوت فضاهای آکوستیکی دو گوینده که مسئول کاهش عملکرد INCA است را کاهش می‌دهد. BAM از یک تئوری هندسه‌ی اقلیدسی استفاده می‌کند که طبق آن اگر دو خط با خط سومی موازی باشند، با هم نیز موازی اند. بنابراین BAM، GMM های مبدأ و هدف را به منظور تشکیل GMM مشترک، با استفاده از UBM (خط سوم) موازی می‌کند. تابع تبدیل این GMM مشترک برای آغازسازی INCA استفاده می‌شود. نتایج آزمایش‌های عینی و عددی مشخص کردند که BAMINCA در تمام زوج تبدیل‌های ممکن بر INCA غلبه می‌کند. به خصوص در تبدیل‌های مرد به زن و زن به مرد، بهبودها بسیار قابل توجهند زیرا تفاوت بین فضاهای آکوستیکی مرد و زن زیاد است. همچنین به طرز اعجاب آوری BAMINCA بر آموزش موازی مبتنی بر DTW نیز غلبه می‌کند. آزمایش‌های عینی متعددی به ازای تعداد جملات آموزشی مختلف انجام شدند و مشاهده شد که BAMINCA همیشه از INCA و DTW بهتر است. اگرچه، این برتری تعداد جملات آموزشی محدود باشد، مشهودتر است. پس به عنوان مکملی برای INCA، BAMINCA یک روش آموزش ناموازی جدید و مؤثر برای تبدیل گفتار است که مشکل آغازسازی INCA را حل می‌کند. این بهبود در آغازسازی BAMINCA را قادر می‌سازد که از آموزش موازی مبتنی بر DTW نیز بهتر عمل کند. بنابراین به صرفه است که برای همردیف سازی فریم در تبدیل گفتار، به ازای کمی هزینه‌ی محاسباتی بیشتر، BAMINCA را جایگزین DTW کنیم. کاری که می‌توان در آینده انجام داد، اعمال BAMINCA به تبدیل گفتار چند زبانه است.

References

1- Stylianou Y, Cappe O, and Moulines E.Continuous probabilistictransform for voice conversion. IEEE Trans. Speech Audio Process.Mar. 1998;6 (2): 131-142.
2- Helander E, Silen H, Virtanen T, and Gabbouj M. Voice conversion using dynamic kernel partial least squares regression.IEEE Trans. Audio,Speech, Lang. Process.Mar. 2012;20 (3): 806-817.
3- Ahangar M ,Ghorbandoost M , Sheikhzadeh H, Raahemifar K, ShahrebabakiA S , and Amini J. Voice Conversion Based on State Space Model and Considering Global Variance. In IEEE Intl. Symp.SignalProcessing and Information Technology (ISSPIT) ;2013. p. 416-421.
4- Chen L H, Ling Z H , Liu L J, and Dai L R. Voice conversion using deep neural networks with layer-wise generative training.IEEE/ACMTransactions on Audio, Speech and Language Processing (TASLP).2014; 22 (12): 1859-1872.
5- Ghorbandoost M ,Sayadiyan A, Ahangar M, Sheikhzadeh H, Shahrebabaki A S, and Amini J.Voice conversion based on feature combination with limited training data.Speech Commun.2015; 67: 113-128.
6- Ye H, and Young S. Quality-enhanced voice morphing using maximum likelihood transformations. EEE Transactions on Audio, Speechand Language Processing. 2006;14 (4): 1301-1312.
7- Dutoit T, Holzapfel A, Jottrand M, Moinet A, Perez J, and Stylianou Y. Towards a voice conversion system based on frame selection.inProc. ICASSP. 2007;4: IV-513-IV-516.
8- Sundermann D, Bonafonte A, Ney H, and Hoge H. A first step towards text-independent voice conversion. In: Proc. Int. Conf. on SpokenLanguage Processing. National Cheng

Kung University, Tainan, Taiwan; 2004.p. 1173-1176.
9- Sundermann D, Hoge H, Bonafonte A, Ney H, Black AW, and Narayanan S. Text-independent voice conversion based on unit selection. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing. Toulouse; 2006. Vol.1. p.81-84.
10- Y eH, and Young S. Voice conversion for unknown speakers. In: Proc. Int. Conf. on Spoken Language Processing;2004. p. 1161-1164.
11- Erro D, Moreno A, and Bonafonte A. INCA algorithm for training voice conversion systems from nonparallel corpora. IEEE Trans. Audio,Speech, Lang. Process.2010; 18 (5): 944-953.
12- Reynolds D A, Quatieri T F, and Dunn R B. Speaker verification using adapted gaussian mixture models. Dig. Sig. Proc.2000; 10 (1): 19-41.
13- Kawahara H, Masuda-Katsuse I, and de Cheveigne A. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F extraction: Possible role of a repetitive structure in sounds.Speech Commun.1999; 27 (3): 187-207.