پیاده سازی سیستم بازشناخت گوینده

مرجان شاهچرا، حسام شینی

دانشگاه آزاد اسلامی واحد نجف آباد

E-mail:unison_mcd@yahoo.com

خلاصه – بازشناخت گوینده ، شناخت گوینده فرایندی است که شخص را بر اساس صدای او شناسایی می کند.شناخت گوینده شامل دو جزء تایید هویت و تعیین هویت می شود. در پیاده سازی ارائه شده سیستم بازشناسی گوینده مبتنی بر تحلیل کپستروم سیگنال گفتار با روش *MFCC* و مدلسازی وتطبیق خصیصه با روش *LBG-VQ* اجرا شده است ، الگوریتم *MFCC* دارای دقت بسیار بالا در بین دیگر الگوریتم ها است ولی به قیمت زمان اجرای بیشتر می انجامد. شناسایی بر اساس خصوصیات وابسته به صوت فرد است و تحلیل کوتاه مدت باعث کیفیت این پردازش می شود. در این مقاله،با تعاریف اصلی در زیست سنجی و *DSP* شروع کرده و سپس به سمت استخراج خصیصه وتکنیک های مدلسازی گوینده می رویم و در آخرکلیه مراحل وتکنیک های توصیف شده در محیط برنامه ی *MATLAB* پیاده سازی و تست می شود.

کلما ت کلیدی- سیستمهای تشخیص گفتار، بازشناسی گوینده ، MFCC ، VQ

1- مقدمه

زیست‌سنجی عبارت است از دانش و فن‌آوری اندازه‌گیری و تحلیل آماری داده‌های زیستی در فن‌آوری اطلاعات واژه‌ی زیست‌سنجی به مجموعه فن‌آوریهایی اطلاق می‌گردد که در آنها از اندازه‌گیری و تحلیل ویژگیهایی از بدن انسان همچون اثر انگشت، اثر کف دست، شبکیه و عنبیه‌ی چشم، الگوهای صوتی، الگوهای مربوط به رخسار ، دمانگاری صورت، شکل دست یا گوش، داده‌های به دست آمده از گام، الگوهای وریدی، دی.ان.ای و یا ویژگیهایی همچون دستخط (امضا) و دینامیک ضربه زدن به صفحه‌کلید برای تأیید هویت اشخاص استفاده می‌شود .این فن‌آوریها در تلاشند تا اندازه گیری و مقایسه‌ی ویژگیهای برشمرده شده را به منظور بازشناسی افراد به صورت خودکار درآورند.

عملیات سیستمهای زیست‌سنجی در بر دارنده‌ی دو مرحله‌ی مجزا می‌باشد: ثبت کاربر و بازشناسی کاربر. در مرحله‌ی اول اطلاعات مربوط به کاربر به سیستم وارد می‌شوند و در مرحله‌ی دوم اطلاعات ورودی حاضر با اطلاعات ذخیره شده مقایسه می‌گردند.

تشخیص گوینده عبارت است از فرایند تشخیص خودکار هویت شخص صحبت‌کننده بر اساس اطلاعات یکتای

موجود در موج صوتی صحبت او. این فن‌آوری امکان تشخیص هویت شخص گوینده و در نتیجه امکان کنترل دسترسی او در هنگام استفاده از خدماتی همانند شماره‌گیری صوتی، بانکداری تلفنی، خرید تلفنی، خدمات دسترسی به پایگاه داده‌ها، خدمات اطلاعاتی، پست الکترونیکی صوتی، کنترل امنیتی برای ورود به قلمروهای اطلاعاتی محرمانه و دسترسی از راه دور به کامپیوترها را فراهم می‌آورد[7].

سیستمهای تشخیص گوینده از لحاظ روش استفاده، عموماً در دو دسته‌ی سیستمهای تأیید هویت گوینده و سیستمهای بازشناسی هویت گوینده قرار می‌گیرند.در یک سیستم تأیید هویت گوینده، شخص عموماً با انتخاب یا وارد کردن نام یکی از کاربران خاص سیستم ادعا می‌کند که او همان کاربر ثبت‌شده‌ی سیستم است. در این حالت سیستم وظیفه دارد ویژگیهای صوتی شخص مدعی را با ویژگیهای صوتی ذخیره شده‌ی کاربر ثبت شده‌ی مورد ادعا مقایسه نموده و با استفاده از نتیجه‌ی به دست آمده ادعای شخص را بپذیرد یا رد کند.در یک سیستم بازشناسی هویت گوینده، شخص صحبت کننده ادعای هویت یک کاربر خاص ثبت شده را نمی‌نماید و این سیستم است که وظیفه دارد که او را در میان کاربران ثبت شده‌ی سیستم بازشناسی نماید و یا تشخیص دهد که ویژگیهای صوتی او با هیچ یک از کاربران ثبت شده همخوانی ندارد.

سیستمهای تشخیص گوینده از دیدگاه دیگری به دو دسته‌ی سیستمهای تشخیص گوینده‌ی وابسته به متن و سیستمهای تشخیص گوینده‌ی مستقل از متن تقسیم می‌شوند .روش اول نیازمند آن است که گوینده کلمات کلیدی یا جمله‌های ثابتی را چه در مرحله‌ی یادگیری و چه در آزمونهای تشخیصی بیان کند، در حالی که دومی وابسته به جمله یا کلمه‌ی خاصی نیست [8].

تقریباً در تمامی سیستمهای تشخیص هویت با استفاده از فرایندی که به تشخیص الگو شهرت دارد شباهت هر زوج نمونه نمره‌گذاری می‌شود. استفاده از این روش نیازمند وجود دسته‌ای از خصایص منحصر به فرد و قابل مقایسه که از ویژگی انتخاب شده به عنوان ورودی سیستم استخراج شده می‌باشد.ویژگیهای فیزیکی افراد نظیر ساختار اندامهای صوتی، اندازه‌ی چاله‌ی بینی و ویژگیهای تارهای صوتی منحصر به فرد بوده و از طریق الگوریتمهای پردازش سیگنال به صورت پارامترهای خصیصه‌ای یا مجموعه‌ی خصایص قابل استخراج می‌باشند. این حقیقت پایه‌ی روشهای پیاده‌سازی سیستمهای تشخیص صحبت می‌باشند. مهمترین گلوگاه سیستمهای تشخیص گوینده (و به تبع هم خانواده بودن مهمترین گلوگاه سیستمهای تشخیص صحبت) نحوه‌ی عملکرد آنها در مکانهای دارای شرایط متفاوت با شرایط آزمایشگاهی که از ویژگیهای عمده‌ی آنها می‌توان به حضور نویز در سیستم اشاره کرد می‌باشد. برای غلبه بر این مشکل از روشهای هنجارسازی استفاده می‌گردد که این روشها نیز انواع مختلفی دارند و در سیستمهای تجاری موجود، اغلب نمود پیدا می‌کنند.

با تحلیل یک موج صوتی می‌توان خصیصه‌های اندامهای گفتاری گوینده را تخمین زد که این خصیصه‌ها راهکاری برای تشخیص هویت و تصدیق آن به روش زیست‌سنجی فراهم می‌آورند.

یک سیستم تشخیص الگو شامل دو جزء است: یک استخراج کننده‌ی خصیصه‌ها و یک طبقه‌بندی کننده یا مدل کننده. ایده‌آل آن است که وقتی داده‌ها به فضای داده‌های خصیصه‌ها انتقال پیدا کرد به سمت طبقه‌ای کشیده شود که از همه به آن نزدیک‌تر است و از طرف طبقه‌های متفاوت دیگر بازپس زده شود. وقتی که به طبقه‌بندی کننده آموزش داده شد که بین طبقه‌ها در این فضای انتقال داده شده از خصیصه‌ها تمایز قائل شود یک سیستم تشخیص نیازمند آن است که تنها داده‌های ورودی را از طریق همان سیستم استخراج خصیصه‌ها انتقال دهد و مشخص کند که در کدام طبقه یک مشاهده‌ی جدید رخ می‌دهد.

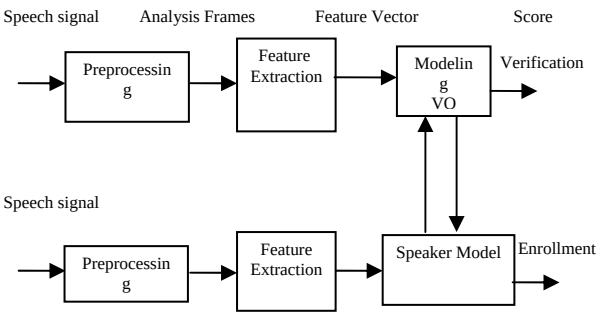
ساختارهایی که برای هر دو نوع سیستم ارائه شد هر دو دارای یک مرحله برای تشخیص میزان شباهت الگوهای متعلق به گوینده‌ی حاضر با گوینده‌ی مورد ادعا (نوع اول) یا همه‌ی گویندگان است که با استفاده از آن معیاری برای

تصمیم گیری در اختیار ما قرار داده می‌شود. همچنان که برای تشخیص الگو، الگوریتمهای متعدد و روشهای گوناگون وجود دارد الگوریتمهای گوناگونی نیز برای یافتن میزان شباهت میان الگوها وجود دارد. انتخاب یک روش به ویژگیهای سیستم هدف بستگی دارد. بعضی از روشهای موجود تنها می‌توانند فقط برای سیستمهای وابسته به متن یا فقط برای سیستمهای مستقل از متن مورد استفاده قرار گیرند و بعضی می‌توانند برای هر دو نوع مورد استفاده قرار گیرند.دو روش MFCC وLPC روشهایی برای استخراج خصیصه است وروشهای TDW وVQو HMM وGMM ازجمله روشهایی برای مدلسازی یا طبقه بندی است .

2-روش پیاده سازی سیستم بازشناخت گوینده

در پیاده سازی دو فاز ثبت کاربر و بازشناسی کاربر بسیار نزدیک به هم هستند که برای مثال در هر دو فاز، استخراج خصیصه ومدلسازی با الگوریتم مشابه پیاده سازی شده ومورد مقایسه برای شناسایی گوینده قرار می گیرد.

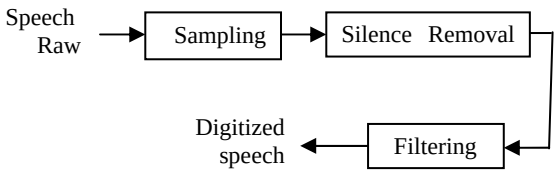
سیستم بازشناخت گوینده از نظر عملیات روی سیگنال گفتار در هردو فاز ثبت کاربر و بازشناسی شامل زیر سیستم های دیگری می شود که در بلوک دیاگرام زیر مشخص شده است.



2-1- زیر سیستم پیش پردازش سیگنال

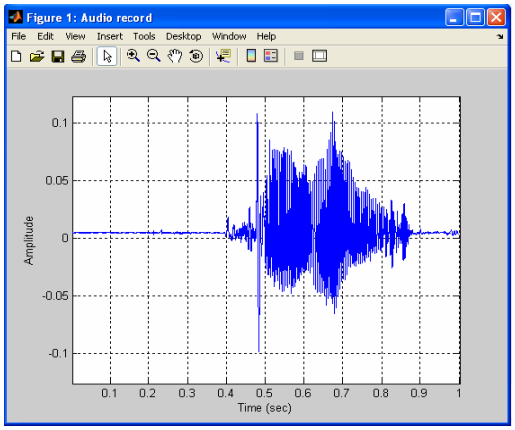
ورودی سیستم سیگنال گفتار خام است وخروجی برداری شامل سمپل های نمونه برداری شده است.

زیر سیستم پیش پردازش سیگنال شامل مراحل نشان داده شده در بلوک دیاگرام زیر است.



2-1-1- نمونه برداری

زیر سیستم پیش پردازش سیگنال از سیگنال خام وآنالوگ گفتار با دستور waverecord یک ثانیه از صوت کاربر با فرکانس fs=۲۲۰۵۰ هرتز و به صورت تک کاناله ضبط شده و با دستور wavwrite صدای ضبط شده با فرکانس fsو۱۶بیتی در فایل user.wav ذخیره می شود.



شکل۲ صوت ثبت شده، که کاربر کلمه سلام را بیان کرده

2-1-2-فیلترینگ

فیلتری برای بالا بردن کیفیت سیگنال به منظور جلوگیری از حساسیت سیگنال در مراحل بعدی استفاده شد که به کمک دستورwden فیلتر ویولت روی سیگنال اجرا شده و که این دستور پروسه ی اتوماتیک حذف نویز سیگنال یک بعدی را به کمک ویولت انجام می دهد.

```
[m2,CXD,LXD]=wden(m,'minimaxi','s','mln',8,'db10');
```

m نام برداری است که حذف نویز روی آن انجام شده وm2 برداری است که پس از حذف نویز ایجاد می شود ،با

انتخاب minimaxi آستانه گذاری مینیماکس ،s آستانه گذاری نرم یا soft ، mln استفاده از تخمین نویز به صورت مستقل در سطوح تجزیه و db10 رشته ای که محتوی نام ویولت مادر از نوع Daubechies با vanishing moments درجه ی ۸ است.

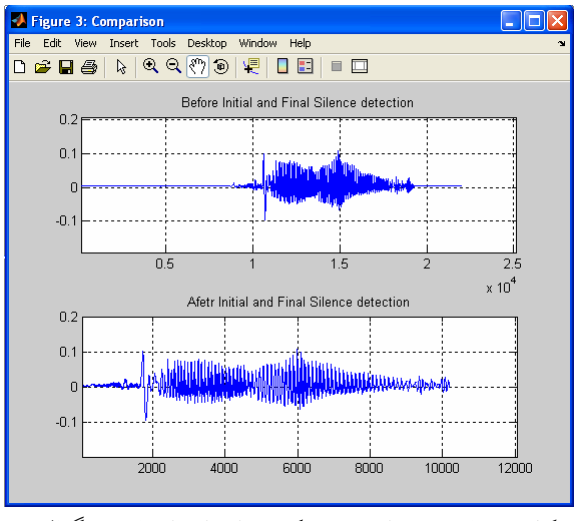
3-1-2-برداشتن سکوت

با برداشتن سکوت ابتدا وانتهای گفتار به طور زیادی سرعت بازشناخت گفتار بهبود بخشیده می شود,به دلیل اینکه مقدار وحجم زیادی از فریم ها کاهش پیدا می کند . تشخیص سکوت معمولا مبنی بر خصوصیات سیگنال است، یکی از روش ها محاسبه ی کوتاه مدت انرژی از فرمول (1) محاسبه و مقایسه می شود[1]:

$$E_n = \sum_{k=1}^{Wn} |x[k]|^2 w[n-k] \qquad (1)$$

$$E_{avg} = \frac{1}{N} \sum_{k=1}^{N} |x[k]|^2$$

در معادله(1), N تعداد کل سگمنت هایی است که سیگنال برای میانگین گیری به آن تقسیم شده است.



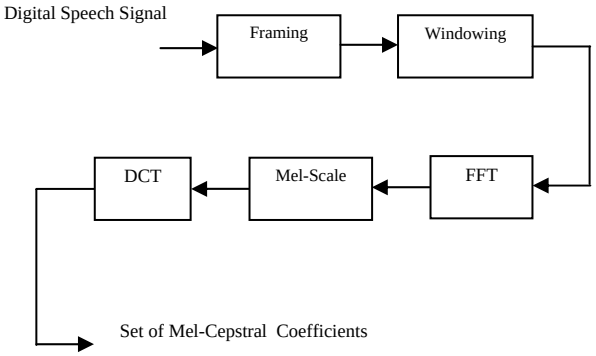
شکل۲ :نمایش جدا کردن سکوت ابتدا و انتهای سیگنال

2-2- زیر سیستم استخراج خصیصه

ورودی سیستم سیگنال گفتاردیجیتال است که به صورت برداری شامل همه ی ارزشهای نمونه برداری شده است.

وخروجی به صورت برداری از صوت شنیده شده و تشخیص داده شده است.

مراحل پیاده سازی استخراج خصیصه در بلوک دیاگرام زیر نشان داده شده است.



حجم زیادی از اطلاعات در حین گفتاربوجود می آید ولی مشخصات ضروری فرآیند گفتار نسبتا آرام تغییر می کند.بنابر این استخراج خصیصه در واقع فرآیندی است که اطلاعات را ساده می کند و کاهش می دهد در حالی که اطلاعات قابل تبعیض گوینده را حفظ ونگه می دارد.

به خاطر طبیعت سیگنال گفتار به آرامی تغییر می کند یا شبه ایستا (quasi-stationary) است.

به این معنی که هنگامی که گفتارمورد تست قرار می گیرد در دوره زمانی به مقدار کافی کوتاه (10-30 msec) دارای خصوصیات صوتی کاملا پایدار وایستایی است وچون اپراتورهای آماری تنها درفواصل ایستا قابل استفاده هستند و این منجربه مفهوم مفیدی در توصیف سیگنال گفتار انسان است به نام تحلیل کوتاه مدت (short-term analysis)می شود .

در فریم بندی کردن (framing) بردار ارزشهای نمونه برداری شده به بلوک های روی هم افتاده و هم پوشانی شده ای تقسیم می شود. هر بلاک تقریبا 16 msec وشامل ۲۵۶ سمبل با ۱۲۸ سمبل هم جوار است.

برای جلوگیری از دست رفتن اطلاعات,هم پوشانی ۵۰-۳۰٪ نیاز است والگوریتم های FFT زمانی دارای حداقل پیچیدگی است که اندازه سیگنال توانی از ۲ باشد و انتخاب ۲۵۶ به همین دلیل است.

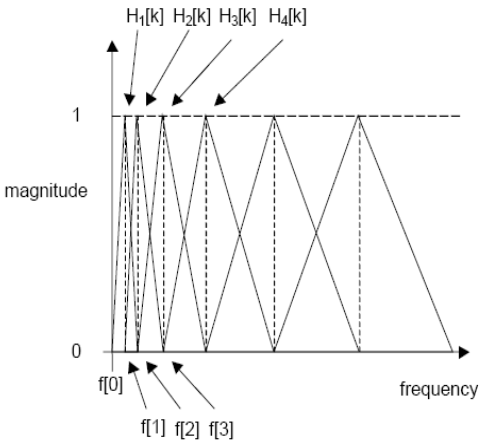
در (windowing) برای اجتناب از تغییرات تند و ناگهانی در نقطه آخر فریم با اضافه کردن و ضرب تابع پنجره از اعوجاج وناپیوستگی سیگنال جلوگیری می شود.
تابع های پنجره متعددی در محدوده بازشناخت گوینده استفاده می شود که از یکی از معروف ترین آنها تابع پنجره همینگ استفاده شد که با معادله (2) تعریف می شود [2].

$$w(n) = 0.54 - 0.64\cos\left(\frac{2n\pi}{N-1}\right) \qquad (2)$$

در معادله (2) ، N سایز فریم است .
در این مرحله از فرایند به خصیصه های استخراج شده از یک فریم بردار خصیصه(feature vector) گفته می شود.
مرحله بعدی محاسبه کپستروم (cepstrum) است که بیان کننده سیگنال است در جایی که دو جزء به دو بخش جمع پذیر تقسیم شده اند و تقسیم مطلوب به اجزاء جمع پذیر است، با رابطه زیر [3].

Cepstrum(frame)=IDFT(log(|DFT(frame)|))

در واقع تکنیک MFCC یا Mel-Frequency Cepstrum Coefficient مشابه محاسبه کپستروم است به جز یک قدم اضافه شده مهم که محور فرکانس بر طبق mel-scale منحرف می شود.
مل (mel) واحد زیر و بمی صوت است و(mel-scale) مقیاسی لگاریتمی است که همانند آنچه گوش انسان درک میکند ومی شنود، عمل می کند.برای محاسبه یک راه مفید استفاده از بانک فیلتر(filter bank) است،یک فیلتر برای هر جزء فرکانسی مل مطلوب استفاده شده که در پیاده سازی از ۲۰ فیلتر استفاده شد.هر فیلتری در این بانک پاسخ فرکانسی میان گذر مثلثی دارد و هر فیلتری میانگین کپستروم اطراف مرکز فرکانسی را با افزایش پهنای باند محاسبه می کند به نحوی که در شکل ۳ نشان داده شده است[4].



شکل ۳: نمایش ساختار فیلتر بانک

در پیاده سازی آخرین قدم که محاسبه ی IDFT است برای کیفیت محاسبات با Discrete Cosine Transform (DCT) جایگزین می شود.

2-3- زیر سیستم فشرده سازی وتطبیق خصیصه ها

ورودی بردار شنیده شده است وبا پیاده سازی الگوریتم(Linde Buzo Gray Vector Quantization) خروجی مفهومی به نام دفتر کد (code book) است.[6] vector quantization (VQ) عملیاتی است که بردار ها را از یک بردار فضا به یک مرکز ثقل(centroid) از ناحیه ای از آن فضا نگاشت می کند.
اساساً روشی است که برای فشرده‌سازی داده‌های آموزش دهنده‌ی سیستم تا اندازه‌ای قابل مدیریت و کارا می‌باشد. با استفاده از یک دفتر کد می‌توان داده‌های اولیه را به مجموعه‌ی کوچکی از نقاط نمایانگر کاهش داد. مقدارگزینی برداری هم در سیستمهای وابسته به متن و هم در سیستمهای مستقل از متن قابل استفاده است.
انحراف تدریجی (quantization distortion)با الگوریتم فاصله اقلیدسی بین بردار ومرکز ثقل محاسبه شدکه در معادله (3) زیر نمایش داده شده[5].

$$d_E(x,y) = \sum_{i=1}^{N} (x_i - y_i)^2 \qquad (3)$$

در معادله(3) ،N تعداد بردار ها است.

4-2- زیر سیستم تصمیم

این عملیات بستگی به روش تطبیق و مدل سازی گوینده دارد.با توجه به روش تطبیق، تصمیم روی فاصله ی محاسبه شده است ومدل گوینده با کوچکترین امتیاز انتخاب می شود وبدین صورت گوینده شناسایی می شود.

3- نتیجه گیری

برای ارزیابی سیستم کلمه ای یکسان توسط سه کاربر متفاوت گفته و به عنوان مرجع مقایسه ذخیره شده و در بار دوم دوباره از هر کاربر درخواست شد که کلمه مورد نظر را دوباره تکرار کنند.برای تصمیم گیری ومقایسه کمترین فاصله محاسبه شده، بیشترین امتیاز برای شباهت است.جدول ۱ نتایج فاصله محاسبه شده در هر مقایسه را نمایش میدهد.در هر سطر قسمت مشخص شده با حاشیه تیره تر مقادیری است که کمترین میزان فاصله را دارد و مشاهده می شود که خود شخص کاربر دارای کمترین مقدار فاصله نسبت به همه ی گویندگان دیگر است و این نمایانگر این است که هر گوینده می تواند به طور موفقیت آمیزی از میان دیگران انتخاب شود.

	a0	a1	r0	r1	s0	s1
a0	0	0.1226	0.3664	0.3297	0.4009	0.4685
a1	0.1226	0	0.5887	0.3258	0.4086	0.4894
r0	0.3664	0.5887	0	0.0989	0.3299	0.4243
r1	0.3297	0.3258	0.0989	0	0.367	0.4287
s0	0.4009	0.4086	0.3299	0.367	0	0.1401
s1	0.4685	0.4894	0.4243	0.4287	0.1401	0

جدول ۱: نمایش فاصله ها برای کاربر های آزمایش شده

سپاسگزاری:

از استاد راهنمای این پروژه جناب آقای مهندس حمیدرضا مراتب کمال تشکر را داریم واز پدر و مادر عزیز و مهربانمان قدردانی میکنیم.

مراجع:

[1] Rabiner and Schafer. Digital Processing of Speech
Signals. Prentice Hall International, 1978.

[2] S. Furui, Digital Speech Processing, Synthesis and Recognition, New York, Marcel Dekker, 2001.

[3] J. R. Deller, J. H. L. Hansen, J. G. Proakis, Discrete-Time Processing of
Speech Signals, Piscataway (N.J.), IEEE Press, 2000.

[4] Karpov. Evgeny, Real-Time Speaker Identification, University of Joensuu,
Department of Computer Science,Master's Thesis,2003.

[5] Lau. Kelvin, Chin. Scott,leu. Lindsey, A Speaker Verification System
, University of Victoria,Master's Thesis, 2002.

[6]Y. Linde, A. Buzo, and R. M. Gray, ``An Algorithm for Vector Quantizer Design," IEEE Transactions on Communications, pp. 702--710, January 1980.

[7] Prabhakar, S., Pankanti, S., and Jain, A. Biometric recognition: security
and privacy concerns. IEEE Security & Privacy Magazine 1 (2003), 33–42.

[8] Li, Q., Juang, B.-H., and Lee, C.-H. Automatic verbal information verification
for user authentication. IEEE Trans. on Speech and Audio Processing 8 (2000), 585–596.

[9] "Matlab VOICEBOX"
http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html