DOI: 10.22041/ijbme.2007.13499

New Biologically Inspired Connectionist Approaches to Improve Machine Speech Recognition

M.R. Yazdchi[1]*, S.A. Seyyed Salehi[2]

[1] Assistant Professor, Biomedical Engineering Department, Faculty of Engineering, University of Isfahan, Isfahan, Iran
[2] Assistant Professor, Biomedical Engineering School, Amirkabir University of Technology, Tehran, Iran

Abstract

One of the most important challenges in automatic speech recognition is in the case of difference between the training and testing data. To decrease this difference, the conventional methods try to enhance the speech or use the statistical model adaptation. Training the model in different situations is another example of these methods. The success rate in these methods compared to those of cognitive and recognition systems of human beings seems too much primary. In this paper, an inspiration from human beings' recognition system helped us in developing and implementing a new connectionist lexical model. Integration of imputation and classification in a single NN for ASR with missing data was investigated. This can be considered as a variant of multi-task learning because we train the imputation and classification tasks in parallel fashion. Cascading of this model and the acoustic model corrects the sequence of the mined phonemes from the acoustic model to the desirable sequence. This approach was implemented on 400 isolated words of TFARSDAT Database (Actual telephone database). In the best case, the phoneme recognition correction increased in 16.9 percent. Incorporating prior knowledge (high level knowledge) in acoustic-phonetic information (lower level) can improve the recognition. By cascading the lexical model and the acoustic model, the feature parameters were corrected based on the inversion techniques in the neural networks. Speech enhancement by this method had a remarkable effect in the mismatch between the training and testing data. Efficiency of the lexical model and speech enhancement was observed by improving the phonemes' recognition correction in 18 percent compared to the acoustic model.

Keywords: **Speech recognition; Speech enhancement; Inversion of neural networks; Bidirectional neural networks; Lexical modeling**

* **Corresponding author**
Address: *MohammadReza Yazdchi, Biomedical Engineering Department- Faculty of Engineering, University of Isfahan, Hezar Jarib St.,Isfahan, Iran*
Tel: *+98 9133152045*
Fax: *+98 311 2646997*
E-mail: *yazdchi@eng.ui.ac.ir*

© Copyright 2008 ISBME, http://www.ijbme.org

DOI: 10.22041/ijbme.2007.13499

روش‌هاي اتصال‌گراي جديد برگرفته از سامانه ادراك گفتار انسان به منظور بهبود بازشناسي گفتار ماشيني

محمدرضا يزدچي[1*]، سيدعلي سيدصالحي[2]

[1] استاديار گروه مهندسي پزشكي، دانشكده فني و مهندسي، دانشگاه اصفهان، اصفهان

[2] استاديار دانشكده مهندسي پزشكي، دانشگاه صنعتي اميركبير، تهران ssalehi@aut.ac.ir

چكيده

بازشناسي خودكار گفتار در شرايط عدم تطابق دادگان آموزش و آزمون، يكي از چالش‌هاي مهم در اين مورد است. به منظور كاهش هر چه بيشتر اين عدم تطابق، روش‌هاي مرسوم، سعي در بهسازي گفتار يا تطابق مدل آماري دارند. در اين زمينه از جمله روش‌هاي ديگر مي‌توان به آموزش مدل در شرايط مختلف اشاره كرد. موفقيت در اين روش‌ها، در مقابل كارايي سيستم درك و بازشناسي در انسان بسيار ابتدايي به نظر مي‌رسد.

در مقاله حاضر با الهام از سيستم درك و بازشناسي در انسان، شبكه عصبي دوسويه جديدي طراحي و پياده‌سازي شده‌است. اين شبكه قادر است تا با اتصالات بازگشتي ضمن مدل‌سازي دنباله واج نظير كلمات مجزا، طي تكرارهاي مختلف، دنباله واج استخراج‌شده از مدل صوتي را به دنباله واج مطلوب، اصلاح كند. براي اين منظور پس از پياده‌سازي روش مذكور بر روي ۴۰۰ كلمه مجزا از دادگان فارس دات تلفني، در بهترين حالت، افزايش ۱۶/۹٪ در صحت بازشناسي واج مدل صوتي مشاهده شد. تشكيل بستر جذب در اين شبكه عصبي دوسويه، از مزاياي اين شبكه در برابر شبكه‌هاي عصبي تك‌سويه است. در ادامه، به كمك متوالي كردن مدل واژگاني مذكور با مدل صوتي، متغيرهاي بازنمايي بر اساس روش‌هاي معكوس‌سازي شبكه‌هاي عصبي اصلاح گرديد. بهسازي گفتار با اين روش نتايج قابل ملاحظه‌اي در كاهش عدم تطابق دادگان آزمون و آموزش در پي داشت. در اين مقاله كارايي مدل واژگاني و بهسازي گفتار در قالب افزايش صحت بازشناسي واج به ميزان ۱۸٪ نسبت به مدل صوتي نشان داده‌شده‌است.

واژه‌هاي كليدي: بازشناسي گفتار؛ بهسازي گفتار؛ معكوس سازي شبكه‌هاي عصبي؛ شبكه‌هاي عصبي دوسويه؛ مدلسازي واژگاني

* عهده‌دار مكاتبات

نشاني: اصفهان، خ هزارجريب، دانشگاه اصفهان، دانشكده فني مهندسي، گروه مهندسي پزشكي

تلفن: ۰۳۱۱۲۶۴۶۹۹۷، **دورنگار:** ۰۹۱۳۳۱۵۲۰۴۵، **پيام نگار:** yazdchi@eng.ui.ac.ir

© Copyright 2008 ISBME, http://www.ijbme.org

۱- مقدمه

تحقيقات چهار دهه گذشته در حوزه بازشناسي گفتار به ايجاد دسته وسيعي از سيستم‌هاي بازشناسي خودكار گفتار[1] تجاري كنوني منجر شده‌است. اين سيستم‌ها به منظور بازشناسي گفتار كلمات مجزا[2]، بازشناسي گفتار پيوسته[3] يا گفتار محاوره‌اي[4] با حجم واژه‌نامه ۱۰ تا بيش از ۶۵۰۰۰ كلمه، در هر دو حالت وابسته به گوينده يا مستقل از آن[5] توسعه يافته‌اند [۱].

مقايسة انجام شده به وسيله ليپمن[6] [۲] بين بازشناسي گفتار ماشيني و شنوندگان انساني در يك تكليف[7] مستقل از گوينده نشان مي‌دهد، بازشناسي گفتار ماشيني فقط در حالت‌هاي ساده مثل بازشناسي اعداد پيوسته مناسب است. نرخ خطاي كلمه با زيادشدن دشواري تكليف، حتي در شرايطي كه هيچ‌گونه تأثير محيطي وجود ندارد، افزايش مي‌يابد[۲]. دشواري تكليف به روش‌هاي متعددي از جمله سرگشتگي[8] (تعداد كلماتي كه مي‌تواند يك كلمه مفروض را دنبال كند) قابل اندازه‌گيري است. اين درحالي است كه شنوندگان انساني در شرايط متنوع وسيعي، بخوبي عمل مي‌كنند. هنگام نبود تطابق[9] بين شرايط آزمون و آموزش در سيستم‌هاي بازشناسي، بازشناسي دشوارتر مي‌شود. كارايي سيستم‌هاي بازشناسي گفتار كه در شرايط كنترل‌شده/تميز[10] آموزش ديده‌اند به‌طور چشمگيري با آزمايش در شرايط اعوجاجي[11] كاهش مي‌يابد. به‌منظور نيل به سيستمي با كارايي انسان نياز است كه سيستم‌هاي بازشناسي خودكار گفتار به اندازه كافي مقاوم[12] شوند، تا بتوانند تنوعات[13] زيادي ازجمله اعوجاج پس‌زمينه[14]، طنين[15] و تنوعات بين گوينده‌اي[16] را تحمل كنند[۳].

انسان‌ها حتي در شرايط آغشتگي شديد گفتار با اعوجاج ايستا و غيرايستا قادر به بازشناسي گفتارند [۲، ۴]. نتايج تحقيقات نشان مي‌دهد شنوندگان انساني حتي در شرايطي كه بخش قابل‌ملاحظه‌اي از طيف گفتار از دست‌رفته‌است، نيز قادر به‌درك گفتارند. به‌عنوان نمونه، گفتاري كه فيلتر بالاگذر يا پايين‌گذر با فركانس قطع ۱۸۰۰ هرتز به‌آن اعمال‌شده‌است، براي انسان قابل درك است [۵]. اين مشاهدات حاكي از آن است كه در سيگنال گفتار به اندازه كافي افزونگي[17] وجود دارد و امكان بازشناسي قابل‌قبول به وسيله شنوندگان انساني، در حالي كه تنها جزيي از اطلاعات زماني-طيفي در آن ارائه‌شده است، را فراهم‌مي‌آورد. روش‌هاي مرسوم براي بهبود بازشناسي در شرايط عدم تطابق، عبارت‌اند از [۶]:

۱-۱- آموزش در شرايط چندگانه[18]

در اين روش ساده، سيستم ASR با انواع مختلف سيگنال گفتار (به عنوان مثال گفتار تميز و اعوجاجي با نرخ‌هاي مختلف) آموزش‌داده مي‌شود. به اين ترتيب سيستم مي‌تواند در طول بازشناسي براي شرايط عدم تطابق ناديده، تعميم يابد. عيب روش آموزش در شرايط چندگانه آن‌است كه خطاي بازشناسي در شرايطي كه دادگان آزمون و آموزش هر دو بدون اعوجاجند، افزايش مي‌يابد. اين مشكل به خاطر افزايش پراكندگي توزيع طبقه‌هاي صوتي مختلف به‌وجود مي‌آيد.

۲-۱-بهسازي گفتار[19]

هدف اين روش تميزكردن گفتار اعوجاجي با كسر اعوجاج تخمين‌زده شده ازآن است. تفاضل خطي، مقادير منفي انرژي را در محل‌هايي كه اعوجاج تخمين‌زده‌شده از اعوجاج واقعي بيشتر است، به‌وجود مي‌آورد. جايگزيني اين مقادير با مقدار صفر، سبب بروز مشكلاتي مي‌شود كه مي‌توان آنها را با تفاضل غيرخطي برطرف نمود. شبكه‌هاي‌عصبي به‌خاطر توانايي مدل‌كردن مؤثر توابع غيرخطي، نتايج مناسبي در بهسازي گفتار در حوزه‌هاي غيرخطي داشته‌اند [۷].

۳-۱-روش‌هاي تطبيقي

در اين روش‌ها مشكل عدم تطابق بين شرايط آزمون و آموزش، با اصلاح متغير[20]هاي مدل براي گويش‌وران يا شرايط اعوجاجي جديد، كاهش مي‌يابد.

در مقالــه حاضــر، تلاشــي در جهــت ايجــاد مــدل‌هــاي واژگاني بر اساس روش‌هاي پيوندگرا، انجام شده‌اسـت. در اين روش سعي مي‌شود تا با مدل‌سازي دنبالــه واج مربـوط به واژگان مختلف در يك شبكه‌عصبي دوسويه، از يـك‌سـو دنبالــه واج گونــه‌هــاي تلفــظ‌هــاي اشــتقاقي[21] بــه وسـيله

[1] **A**utomatic **S**peech **R**ecognition
[2] Isolated Word
[3] Continuous Speech
[4] Spontaneous
[5] Speaker Dependent/Independent
[6] Lippmann
[7] Task
[8] Perplexity
[9] Mismatch
[10] Clean/Controlled
[11] Noise
[12] Robust
[13] Variations
[14] Background Noise
[15] Reverberation
[16] Inter-Speaker
[17] Redundancy
[18] Multicondition Training
[19] Speech Enhancement
[20] Parameter
[21] Alternative Pronunciation

© Copyright 2008 ISBME, http://www.ijbme.org

شبكه‌هاي‌عصبي به دنباله واج تلفظ رسمي[۲۲] و از سوي ديگر دنباله واج بازشناسي شده به وسيله مدل صوتي به دنباله واج برچسب‌شده تصحيح شود.

تلفظ رسمي و اشتقاقي عبارت است از:

۱. تلفظ رسمي

دنباله واج‌هاي استاندارد است كه در گفتار خواندني تلفظ مي‌شود و تنوعاتي نظير تفاوت گوينده، لهجه و هم‌توليدي آواها در آن در نظر گرفته نمي‌شود.

۲. تلفظ اشتقاقي

دنباله واج تلفظ شده در گفتار است كه تلفظ‌هاي مختلف مربوط به تنوع گوينده و محاوره‌اي بودن گفتار را شامل مي‌شود.

به اين ترتيب مدل واژگاني، قادر خواهد بود تا تلفظ‌هاي چندگانه يك كلمه را به تلفظ رسمي آن كلمه تصحيح نمايد. در اين روش به دادگان آموزشي با حجم بزرگ نياز نيست زيرا شبكه‌هاي‌عصبي قادرند تا براي ورودي آزمون كه هرگز در ميان دادگان آموزش نبوده است، بر طبق آن‌چه كه از دادگان آموزش فراگرفته‌اند، تصميم مناسبي اتخاذ نمايند [۸].

حال در صورتي‌كه مدل واژگاني با مدل صوتي متوالي شود، مي‌تواند خطاهاي بازشناسي مدل صوتي را بر اساس يادگيري دنباله‌هاي واج مجاز، برطرف نمايد. مدل‌هاي واژگاني مذكور، نوعي فيلتر پيوندگرا را اجرا مي‌كنند. براي اين منظور، از قابليت‌هاي شبكه‌هاي‌عصبي در حذف اعوجاج و رفع تنوعات از ورودي استفاده مي‌شود. تمام عوامل مذكور در علت ايجاد تنوع در تلفظ، با تأثير بر روي دادگان صوتي سبب ايجاد تلفظ‌هاي چندگانه در يك كلمه واحد مي‌شوند. تصحيح دادگان صوتي مي‌تواند در حذف تلفظ‌هاي چندگانه مؤثر باشد زيرا علت عمده را بايد در تنوع دادگان صوتي جستجو نمود. تصحيح دنباله واج با استفاده از مدل‌سازي واژگاني، اين امكان را فراهم مي‌سازد تا با دوسويه كردن پردازش از سطح واژگان تا سطح دادگان صوتي (متغيرهاي بازنمايي) و اصلاح اين متغيرها، صحت بازشناسي را دوباره افزايش داد. اين روش، تلفيق دانش پيشين واژگاني در اطلاعات صوتي است. مدل‌هاي واژگاني، دانش سطح بالا را استخراج مي‌كنند و روش‌هاي معكوس‌سازي شبكه‌هاي عصبي تلفيق اين دانش را براي افزايش كارايي سطوح پايين‌تر فراهم مي‌آورند.

در بخش دوم مقاله حاضر روش‌هاي الهام گرفته‌شده از سيستم ادراک گفتار انسان مرور مي‌شود. در بخش سوم به روش‌هاي معكوس‌سازي شبكه‌هاي عصبي پرداخته مي‌شود. در بخش چهارم دادگان مورد استفاده و نحوه استخراج ويژگي از دادگان، بيان مي‌شود. در بخش پنجم مدل صوتي استفاده شده تشريح مي‌شود. در بخش ششم مدل واژگاني طراحي شده بيان مي‌شود. در اين بخش الگوريتم كاملي براي آموزش و نحوه عملكرد شبكه عصبي ارائه شده بيان شده است. در بخش هفتم به بيان دو روش به منظور بهسازي متغيرهاي بازنمايي با استفاده از روش‌هاي معكوس‌سازي شبكه‌هاي عصبي، و در بخش بعدي به جمعبندي نتايج حاصل مي‌پردازيم.

۲-روش‌هاي ASR مقاوم الهام‌گرفته از سيستم ادراك گفتار انسان

مشاهده سيستم شنوايي انسان، دو روش جالب در بازشناسي مقاوم را پيشنهاد مي‌دهد: روش داده‌هاي مفقود[۲۳] و بازشناسي چندباندي[۲۴]. اين راهكارها برخلاف ساير روش‌هاي ASR مقاوم، قابليت‌هاي زيستي شنوندگان انساني، در رويارويي با شرايط عدم تطابق غيرقابل پيش‌بيني را، تقليد مي‌نمايند. با استفاده از اين روش‌ها، سيگنال گفتار كه جزئي از اطلاعات آن قابل اتكاست[۲۵]، بازشناسي مي‌شود. اين روش‌ها نياز به دانش پيشين[۲۶] از توزيع اعوجاج ندارند و بنابراين مي‌توانند با اعوجاج‌هاي ايستا و غيرايستا يا عدم تطابق‌هاي ناديده مقابله نمايند.

۲-۱-داده‌هاي مفقود

در بازشناسي گفتار، فرض بر آن است كه برخي از نواحي زماني-فركانسي گفتار، به‌گونه‌اي به وسيله ديگر منابع پوشانده شده‌است. شايان ذكر است بازيابي اطلاعات مفيد

[22] Canonical Pronunciation [23] **M**issing **D**ata [24] Multiband Recognition [25] Reliable
[26] Prior Knowledge

© Copyright 2008 ISBME, http://www.ijbme.org

گفتاري از اين نواحي غيرممکن است [۹]. در اين روش، با انجام يك پردازش اوليه، ميان نواحي قابل اتکا و غيرقابل اتکا[27] تصميم‌گيري مي‌شود. سيستم بازشناسي با استفاده از دادگان تميز، آموزش داده مي‌شود و الگوريتم رمزگشايي[28] نيز هنگام بازشناسي، به گونه‌اي اصلاح مي‌شود که گفتار با بازنمايي قابل اتکاي کاهش‌يافته نيز بتواند بازشناسي شود.

۲-۲-روش چندباندي

ايده اصلي اين روش آن است که باندهاي بسامدي[29] مختلف، با مقادير مختلفي آغشته مي‌شوند. با اين فرض، حوزه کلي سيگنال گفتار به زيرباندهايي[30] تقسيم مي‌شود و بازشناس‌هاي جداگانه‌اي براي هر باند آموزش داده مي‌شوند. بازشناسي براساس باندها يا ترکيب باندهايي که کمتر از اعوجاج متأثر شده‌اند، انجام مي‌گيرد.

۳-روش‌هاي معکوس‌سازي شبکه‌هاي عصبي در تشکيل شبکه‌هاي عصبي دوسويه

در اين قسمت دو روش براي معکوس‌سازي شبکه‌هاي عصبي و به دست آوردن ورودي از خروجي، معرفي مي‌کنيم:

در روش اول پس از آموزش شبکه با دادگان آموزش، اين امکان وجود دارد که با اعمال يك ورودي دلخواه به شبکه و به دست آوردن خروجي آن، بردار خطاي ما بين خروجي مطلوب و خروجي فعلي محاسبه شود؛ همچنين بدون اصلاح وزن‌ها، خطا به لايه‌هاي قبل پس‌انتشار شود تا در لايه ورودي، ورودي اصلاح شود [۱۰]. در اين روش وزن‌ها ثابت و ورودي همانند وزن‌ها براي حصول ورودي بهينه از ديدگاه شبکه اصلاح مي‌شود.

روش دوم بر مبناي آموزش دو شبکه معکوس يکديگر بر روي دادگان آموزش است، به اين معنا که براي شبکه جلوسو (مستقيم)، ورودي و خروجي به‌ترتيب ورودي و خروجي دادگان آموزش و در شبکه‌معکوس، عکس آن است [۱۱]. در ادامه با بررسي‌هاي تحليلي نشان خواهيم داد که هر دو روش قادر به به‌دست آوردن ورودي از خروجي‌اند.

۳-۱-معکوس‌سازي شبکه با اصلاح ورودي به روش گراديان

در اين روش، از پس‌انتشار خطا به منظور اصلاح ورودي استفاده مي‌شود. يك شبکه جلوسو مطابق شکل۱، با دو لايه پنهان HH_1 و HH_2، لايه‌هاي ورودي و خروجي II و OO مفروض است. x_i مؤلفه iام بردار ورودي است که در روش گراديان، مطابق روابط زير اصلاح مي‌شود.

در اين روش، تغييرات کوچکي در ورودي در جهتي که متناظر با منفي گراديان خطاي اندازه‌گيري شده‌است، ايجاد مي‌شود.

$$\Delta x_i = -\gamma \frac{\partial E}{\partial x_i} \qquad (۱)$$

$$\frac{\partial E}{\partial x_i} = \frac{\partial E}{\partial z_{1m}} \frac{\partial z_{1m}}{\partial net_{z_1}(m)} \frac{\partial net_{z_1}(m)}{\partial x_i} = \dot{z}_{1m} w_{im}^{I} \qquad (۲)$$

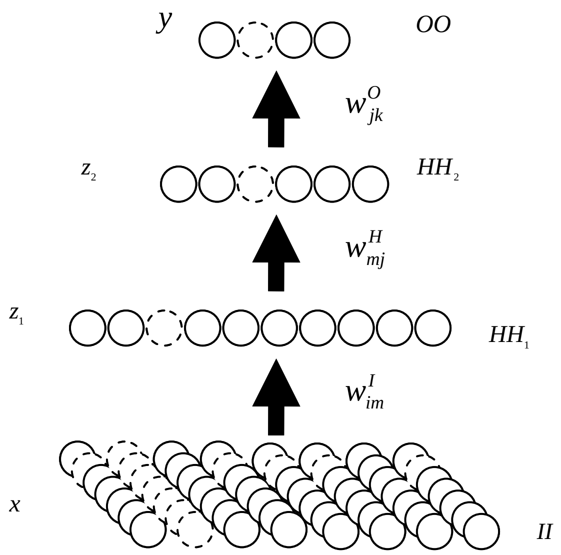
در حالت استفاده از SSE[31]،

$$\dot{y}_k = (\hat{y}_k - d_{y_k}) f'(net_y(k)) \qquad 1 \le k \le OO \qquad (۳)$$

$$\dot{z}_{3j} = \left(\sum_{k=1}^{OO} w_{jk}^{O} \dot{y}_k \right) f'(net_{z_3}(j)) \qquad 1 \le j \le HH_3 \qquad (٤)$$

$$\dot{z}_{2m} = \left(\sum_{j=1}^{HH_3} w_{mj}^{H_2} \dot{z}_{3j} \right) f'(net_{z_2}(m)) \qquad 1 \le m \le HH_2 \qquad (۵)$$

$$\dot{z}_{1l} = \left(\sum_{m=1}^{HH_2} w_{lm}^{H_1} \dot{z}_{2m} \right) f'(net_{z_1}(l)) \qquad 1 \le l \le HH_1 \qquad (٦)$$



شکل۱- يك شبکه جلوسوي نمونه

[27] Unreliable
[28] Decoding Algorithm
[29] Frequency
[30] Subband
[31] **S**um **S**quare **E**rror

© Copyright 2008 ISBME, http://www.ijbme.org

ورودي‌هاي شبكه، بعد از محاسبه گراديان براي تمام فريم‌هاي زماني در يك نمونه آموزشي، اصلاح مي‌شود.

$$X_i = X_i + \sum_t \Delta_t x_i \qquad 1 \le i \le II \qquad (۷)$$

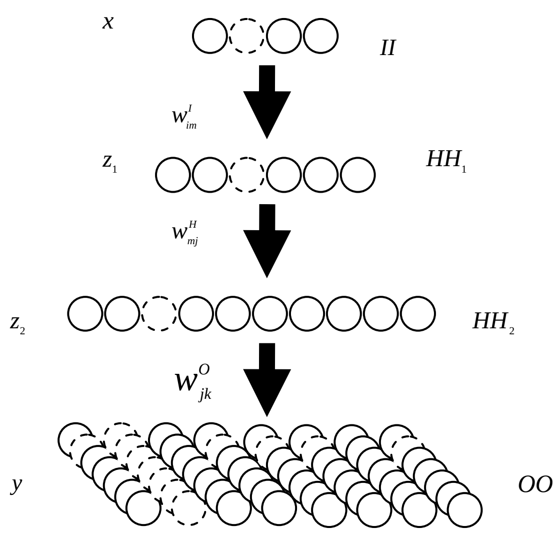
در اصلاح دسته‌اي[32]، ورودي‌ها بعد از همه دادگان تعليم، با N نمونه، اصلاح مي‌شود.

$$X_i = X_i + \sum_N \sum_t \Delta_t x_i \qquad 1 \le i \le II \qquad (۸)$$

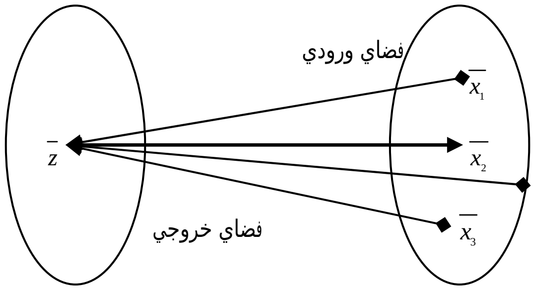
اين روش مانند ساير روش‌هاي مبتني بر گراديان، احتمال گيرافتادن در كمينه‌هاي موضعي را دارد.

۳-۲-معكوس‌سازي شبكه به روش آموزش شبكه معكوس

روش ديگري كه در شبكه‌هاي‌عصبي براي به‌دست آوردن ورودي از خروجي وجود دارد، استفاده از دو شبكه جلوسو به‌صورت معكوس يكديگر است. آموزش اين شبكه‌ها در صورتي كه دادگان آموزش يك به يك باشند، همگرا خواهد شد. به عنوان نمونه اگر دادگان چند به يك باشند، شبكه مستقيم همگرا مي‌شود و شبكه معكوس ميانگين ورودي‌ها را بازسازي مي‌كند [۱۲]. اين شبكه‌ها مي‌توانند بر اساس الگوريتم پس‌انتشارخطا، به صورت جداگانه، يا در قالب يك الگوريتم دوسويه آموزش داده شود [۱۲]. ساختار كلي شبكه معكوس مدل صوتي در شكل۲ و نگاشت بين فضاي ورودي و خروجي اين شبكه‌ها در شكل۳ نمايش داده شده است.



شكل۲- شبكه معكوس مدل صوتي



شكل۳- نگاشت بين فضاي ورودي و خروجي در دو شبكه مستقيم و معكوس

۴-دادگان گفتاري و استخراج بازنمايي

۴-۱-دادگان گفتاري

به علت نياز به وجود تنوعات بالا در دادگان اين تحقيق، دادگان گفتاري كه مورد استفاده قرار مي‌گيرد، دادگان فارس دات تلفني[33] است. اين دادگان از ۵ بخش تشكيل شده است كه به وسيله ۶۴ گويش‌ور بيان شده است. از ميان تمام كلمات بيان شده در دادگان، حداكثر۴۰۰ كلمه وجود دارد كه حداقل بيش از يكبار ادا شده‌است، به همين جهت، اين ۴۰۰ كلمه از ۶۴ گويش‌ور انتخاب و در آموزش و آزمون شبكه‌ها مورد استفاده قرار مي‌گيرد [۱۳].

۷۵% از دادگان انتخابي، به عنوان دادگان آموزش و ۲۵% از آن به عنوان دادگان آزمون، در آموزش و آزمون تمام شبكه‌ها استفاده مي‌شود. بنابراين در كلماتي كه به وسيله همه گويش‌وران ادا شده است، دادگان مربوط به ۴۸ گويش‌ور به جهت آموزش و ۱۶ گويش‌ور در آزمون مدل‌ها استفاده شده است. در آزمون بعضي از مدل‌ها، دادگان به‌دست آمده از مدل صوتي در آزمون مدل، مورد استفاده قرار مي‌گيرد. نرخ نمونه برداري دادگان 11025 Hz است كه به 8 kHz (نرخ نمونه‌برداري گفتار تلفني) تغيير داده مي‌شود. پنجره‌گذاري به جهت استخراج ويژگي از دادگان، با طول پنجره‌ي ۱۲۸ و با گام پيشروي ۶۴ نمونه، انجام مي‌شود كه از نظر زماني حدود ۱۶ ميلي‌ثانيه با همپوشاني ۵۰٪ است.

[32] Batch

[33] Actual Telephone Database (TFARSDAT Database)

© Copyright 2008 ISBME, http://www.ijbme.org

۴-۲-استخراج متغيرهاي بازنمايي

در بيشتر روش‌هاي استخراج ويژگي طيفي، مبتني بر تبديل فوريه، پس از محاسبه طيف به وسيله تبديل فوريه گسسته زمان كوتاه، از يك بانك فيلتر جهت محاسبه ميزان انرژي طيف در باندهاي فركانسي متفاوت استفاده مي‌شود. در اين ميان روش‌هاي الهام گيرنده از سيستم‌هاي زيستي، عملكرد بهتري از خود نشان مي‌دهند [۱۴]. مهم‌ترين روشي كه از سيستم شنوايي انسان در استخراج ويژگي اقتباس‌شده است، انتخاب مقياس غيرخطي براي محور فركانس است. در ميان روش‌هاي متداولي كه بر اساس دو نوع مقياس غيرخطي بارك[۳۴] و مل[۳۵] ارائه شده است، ضرايب كپستروم مقياس مل(MFCC)[۳۶] از متداول‌ترين روش‌هاي موجود در استخراج متغيرهاي بازنمايي است كه در آن پس از حصول مقادير لگاريتم انرژي بانك فيلترها با اعمال بانك فيلتر هنينگ در مقياس مل، اين مقادير به حوزه كپستروم انتقال مي‌يابند. با توجه به محدوده ۱۲۵ تا ۳۵۰۰ هرتزي براي دادگان تلفني از بين ۱۸ بانك فيلتر طراحي شده براي گفتار غيرتلفني، فيلترهاي دوم تا چهاردهم انتخاب و با محاسبه لگاريتم انرژي گفتار در هريك از اين فيلترها، ۱۳ متغير LFBE[۳۷] به‌دست مي‌آيد. متغيرهاي MFCC با اعمال تبديل كسينوسي بر روي متغيرهاي LFBE در تعداد دلخواه به‌دست مي‌آيند. به طور معمول به‌صورت رايج در بازشناسي گفتار، ۱۲ متغير MFCC استخراج مي‌گردد كه با اضافه كردن لگاريتم انرژي كل فريم، در مجموع ۱۳ متغير مي‌شود. بنابراين براي استخراج بردارهاي بازنمايي ۱۳تايي MFCC، تبديل كسينوسي ۱۲تايي بر روي بردارهاي ۱۳تايي LFBE اعمال شده‌است. با اضافه‌شدن لگاريتم انرژي كل فريم به آنها، بردار بازنمايي ۱۳تايي MFCC به‌دست مي‌آيد. اين محاسبات در روابط (۹) تا (۱۱) نشان داده‌شده‌است. در اين روابط L_i متغيرهاي LFBE و C_i متغيرهاي MFCC را تشكيل مي‌دهد.

$$L_i = \log(fbank_i) \qquad i = 2,\ldots,14 \tag{۹}$$

$$C_i = \sum_{j=2}^{14} A_{ij} \log(fbank_j) \qquad i = 1,\ldots,12 \tag{۱۰}$$

$$C_{13} = \log(Energy) \tag{۱۱}$$

به‌منظور هنجارسازي[۳۸] متغيرهاي بازنمايي، هر متغير به تغييراتي كه در كل دادگان مربوط به هر گوينده دارد بهنجار مي‌شود (هنجارسازي طولي متغيرهاي بازنمايي، جداگانه براي هر گوينده). مزيت اين روش نسبت به هنجارسازي طولي در كل دادگان اين است كه متغيرها تا حدي نسبت به گوينده هنجارسازي مي‌شوند. به اين روش، هنجارسازي به ميانگين و واريانس نيز گفته مي‌شود. در ادامه روابط اين روش هنجارسازي آورده شده است.

$$\bar{k}_1 = \frac{1}{M}\sum_{i=1}^{M}\bar{x}_i \tag{۱۲}$$

$$\bar{q}_i = \bar{x}_i - \bar{k}_1 \tag{۱۳}$$

$$\bar{k}_2 = \frac{1}{M}\sum_{i=1}^{M}\bar{q}_i^{\,2} \tag{۱۴}$$

$$\bar{x}_i^{normalized} = \left\{ \frac{q_{in}}{\sqrt{k_{2n}}}, n = 1,2,\ldots,k,\ldots,N \right\} \tag{۱۵}$$

در روابط فوق M تعداد بردارهاي بازنمايي در كل دادگان مربوط به يك گوينده و N تعداد متغيرهاي هر بردار بازنمايي است. با اضافه كردن مشتقات مرتبه اول و دوم (ضرايب مرتبه اول و دوم دلتا)، به منظور هنجارسازي بهتر نسبت به گوينده و نيز به جهت افزودن اطلاعات در رابطه با ويژگي‌هاي پوياي سيگنال، تعداد متغيرهاي هر بردار بازنمايي(N) به ۳۹ متغير مي‌رسد.

ضريب مرتبه iام دلتا به‌صورت زير به‌دست مي‌آيد.

$$\Delta^i\{\bar{u}_t\} = \Delta^{i-1}\{\bar{u}_{t+1}\} - \Delta^{i-1}\{\bar{u}_{t-1}\} \quad , \quad \Delta^0\{\bar{u}_t\} = \bar{u}_t \tag{۱۶}$$

۵-مدل صوتي[۳۹]

مدل صوتي، استخراج‌كننده آواي متناظر هر فريم براساس متغيرهاي بازنمايي است؛ اين مدل يك شبكه پرسپترون چندلايه[۴۰] با دو لايه پنهان، (شكل ۴) و با ساختار نوروني ۳۹*۹-۲۵۶-۱۲۸-۳۴ است. تعداد نورون‌هاي بهينة لايه پنهان، با چندين آزمايش مشخص مي‌شود. تعداد اين نورون‌ها، ابعاد لايه پنهاني است كه در بين آزمايش‌هاي مختلف (در اين آزمايش‌ها تمام شرايط يكسان و تنها تعداد نورون‌هاي لايه پنهان متفاوت است)، بهترين نتيجه را در برداشته است. در ورودي ۹ فريم متوالي، ۴ فريم مجاور

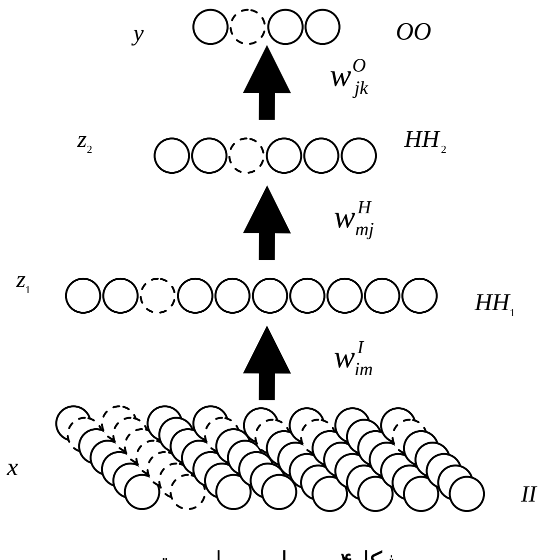
[34] Bark Scale
[35] Mel Scale
[36] **M**el **F**requency **C**epstral **C**oefficients
[37] **L**ogarithmic **F**ilter**b**ank **E**nergy
[38] Normalization
[39] Acoustic Model
[40] **M**ulti**L**ayer **P**erceptron

© Copyright 2008 ISBME, http://www.ijbme.org

راست و ۴ فريم مجاور چپ قرار داده مي‌شوند، تا در خروجي فريم مياني به وسيله ۳۴ نورون (به تعداد آواها) بيان شود. توابع غيرخطي لايه‌هاي پنهان و نيز خروجي به جهت تسريع بيشتر در همگرايي شبكه، تانژانت هيپربوليك، با مقادير خروجي بين ۱و ۱- و وزن‌هاي اوليه بر طبق الگوريتم گوين-ويدرو[41] انتخاب گرديد [۱۵]. شبكه با ضريب يادگيري، ضريب ممنتوم متغير تا همگرايي كامل و بر اساس الگوريتم پس‌انتشار استاندارد آموزش مي‌بينند. اين ضرايب در ابتدا بزرگ و سپس به مرور و با كاهش خطا كوچك‌تر مي‌شوند.

در آموزش شبكه مذكور، از آموزش دسته‌اي استفاده مي‌شود. از دادگان ۴۸ گويش‌ور در مرحله آموزش و ۱۶ گويش‌ور باقيمانده در آزمون مدل استفاده مي‌شود. پس از آموزش مدل، در مرحله آزمون، صحت بازشناسي فريم، بر روي فريم‌هاي غيرسكوت ۷۱/۳۴% به‌دست مي‌آيد.

در تبديل فريم به واج متوسط طول واج‌ها بر حسب فريم مدنظر قرار مي‌گيرد. چون به‌طورمعمول طول واج‌ها در دادگان از دو فريم بيشتر است، تبديل به‌صورت زير انجام خواهدشد. در دنباله فريم به‌دست آمده از خروجي مدل‌صوتي، اگر تعداد فريم‌هاي متوالي از يك نوع، برابر ۲ يا بيشتر از آن باشد به عنوان واج در نظر گرفته مي‌شود. بنابراين از فريم‌هاي تكي صرف‌نظر مي‌گردد. ميزان صحت[42] و دقت[43] واج بازشناسي شده مدل صوتي از دادگان آزمون، به همراه ميزان حذف[44]، درج[45]، جانشيني[46] در جدول ۱ نشان داده شده است.



شكل۴- معماري مدل صوتي

جدول۱-نتايج بازشناسي واج دادگان آزمون مدل صوتي

دقت	صحت	حذف	درج	جانشيني
%۶۱/۱	%۷۱/۶	%۱۳/۷	%۱۰/۴	%۱۴/۷

۶-مدل واژگاني

تاكنون شبكه‌هاي‌عصبي جلوسوي چندلايه، به‌طوروسيعي جهت بازشناسي الگوها مورد استفاده قرار گرفته‌اند. با وجود اين ساختارهاي رايج شبكه‌هاي‌عصبي جلوسو، در بازشناسي الگوها، بخصوص مواردي كه الگوها تحت تأثير اعوجاج قرار گرفته‌اند يا در اثر عوامل مختلف تغيير يافته‌اند، كه به آن تنوعات در الگوهاي ورودي يا عدم تطابق دادگان[47] آموزش با آزمون گفته مي‌شود، نارسايند [۱۲]. از سوي ديگر شواهد موجود حاكي از پردازش دوسويه الگوهاي مورد بازشناسي در مغز است [۱۶-۱۸]. براي افزايش كارايي، رويكرد دوسويه‌سازي پردازش در شبكه‌هاي‌عصبي، رويكرد مناسبي به نظر مي‌رسد. از سوي ديگر، شيوه پردازش مقاوم الگوها در مغز انسان به ويژه در شرايطي كه در الگوهاي ورودي تنوعات وجود داشته باشد، توانايي مناسب جداسازي از طريق مؤلفه‌هاي اساسي غيرخطي سيگنال‌ها در مغز را نشان مي‌دهد [۱۲].

يافته‌هاي تحقيقات علوم اعصاب نيز حاكي از وجود اتصالات دوسويه و تعامل دوطرفه مابين نواحي اوليه و نواحي بالاتر پردازش‌هاي حسي در قشر مغز است. به نظر مي‌رسد قشر مخ در پاسخ به ورودي‌هاي حسي، علاوه بر پردازش‌هاي از پايين به بالا از طريق اعصاب آوران، از طريق پردازش‌هاي بالا به پايين به وسيله اعصاب وابران به سيگنال‌هاي حسي ورودي، دانش اضافه مي‌كند تا فرضيه‌ها و تصورهاي بازسازي شده به واقعيت اشيا در جهان واقعي تا حد ممكن نزديك‌تر شوند [۱۶].

شكل‌گيري جاذب‌ها[48] در قعر بستر جذب‌ها[49] از خصوصيات شبكه‌هاي عصبي با اتصالات دوسويه است، كه تركيب اين خواص با توانايي‌هاي شبكه‌هاي جلوسو، امكان طراحي مدل‌هاي مفيدتر را فراهم مي‌سازد [۱۲].

تشكيل جاذب در شبكه‌هاي عصبي به‌منظور تكميل

[41] Nguyen-Widrow
[42] Correction Ratio
[43] Accuracy Ratio
[44] Deletion Ratio
[45] Insertion Ratio
[46] Substitution Ratio
[47] Mismatch
[48] Attractor
[49] Basin of Attraction

© Copyright 2008 ISBME, http://www.ijbme.org

الگو[50]، در کاربردهاي متعددي از جمله تکميل اعداد دست‌نوشته استفاده شده است [۱۲]. در اين روش نشان داده شده است که با آموزش مناسب يك شبکه‌عصبي جلوسو و داراي اتصالات بالا به پايين جهت اصلاح ورودي، بلوك‌هاي مفقود در الگوهاي ناقص به نحو مؤثري بازسازي مي‌شوند.

با توجه به توانايي مطلوب شبکه‌هاي‌عصبي دوسويه نسبت به شبکه‌هاي عصبي تك‌سويه در تشکيل جاذب و تکميل الگو در ادامه با طراحي يك شبکه‌عصبي دوسويه به‌عنوان يك مدل پيوندگراي واژگاني، توانايي‌هاي شبکه‌هاي‌عصبي به منظور رفع تنوعات واژگاني در يك سيستم بازشناسي گفتار نشان داده مي‌شود. اين شبکه‌عصبي با نحوه عملکردي منحصر به فرد به وسيله الگوريتم خاصي قابل آموزش است، واج‌هاي مفقود در دنباله واج نظير کلمات مجزاي استخراج‌شده از مدل صوتي را به نحو مناسبي بازسازي مي‌کند. به منظور بررسي قابليت‌هاي شبکه مذکور، نتايج با شبکه عصبي جلوسو مقايسه شده است. دو هدف کلي از ايجاد اين مدل واژگاني دنبال مي‌شود:

- اصلاح دنباله واج خروجي مدل صوتي در قياس با دنباله واج برچسب‌شده،
- تصحيح دنباله واج تلفظ‌هاي چندگانه در خروجي مدل صوتي به دنباله واج تلفظ رسمي.

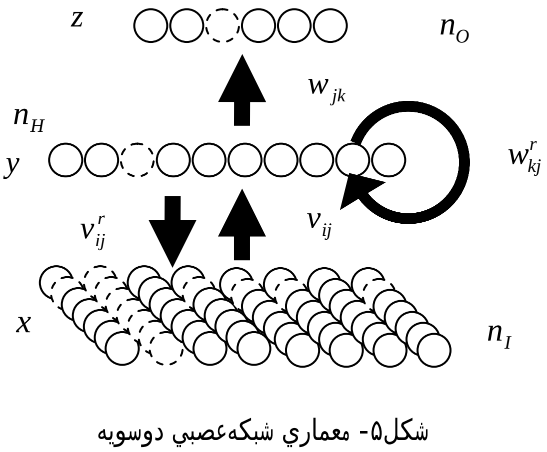
به‌منظور نيل به اهداف مذکور، ساختار مدل واژگاني و ابعاد آن در هر دو حالت کاملاً يکسان است و تنها دو دسته دادگان آموزش مختلف در آموزش اين شبکه، به‌ترتيب به منظور اصلاح دنباله واج در قياس با دنباله واج تلفظ رسمي و برچسب‌شده، مورد استفاده قرار مي‌گيرد. در جدول ۲ به‌عنوان نمونه، برخي از تلفظ‌هاي اشتقاقي و تلفظ رسمي کلمه "چهارده" نشان داده شده‌است.

جدول ۲- تلفظ رسمي و برخي از تلفظ‌هاي اشتقاقي رايج کلمه "چهارده"

تلفظ رسمي	برخي از تلفظ‌هاي اشتقاقي
h/r=dah/'$	r=dah/'$
	r=da/'$
	ah/r=dah'$
	h/rdah/'$
	h/r=deh/'$
	h/r=da/'$
$:بست چ، ':چ، /:آ، =:بست د است.	

۶-۱- معماري شبکه

اين شبکه به طور اساسي يك شبکه‌عصبي MLP با دو اتصال بازگشتي، يك لايه پنهان و يك لايه خروجي با توابع غيرخطي تانژانت هيپربوليك، ولايه ورودي با تابع خطي است. اتصال بازگشتي از لايه پنهان به خودش، به منظور ايجاد حافظه بلندمدت[51] و اتصال موجود ميان لايه پنهان و ورودي جهت اصلاح ورودي در گام بعدي است. اين شبکه قادر است طي چندين گام (تعداد تکرارها به دادگان بستگي دارد و در آموزش شبکه به‌صورت تجربي مشخص مي‌گردد) الگوي ناقص را تصحيح کند. تکرارهاي متوالي با فراهم نمودن سيگنال‌هاي آموزشي بيشتر، مشکل محدوديت نمونه‌هاي تعليم را کم‌رنگ‌تر، مي‌کند. اين مدل قادر است تا نواحي مفقود ورودي را با مقاديري که در مرحله تعليم فراگرفته است، جايگزين نمايد. بنابراين، يکبار به منظور اصلاح دنباله واج در تلفظ‌هاي چندگانه يك کلمه به دنباله واج تلفظ رسمي آن کلمه و بار ديگر جهت تصحيح دنباله واج در تلفظ‌هاي چندگانه يك کلمه به دنباله واج برچسب‌شده به‌کار گرفته مي‌شود. در ورودي اين مدل هر واج با ۳۴ بيت بيان مي‌شود. اين نحوه بيان، قابليت اتصال مستقيم مدل مذکور به لايه‌هاي پايين‌تر را فراهم مي‌سازد. با توجه به مشخص بودن مرز کلمات در اين مدل (کلمات مجزا)، ورودي اين شبکه متناظر با دنباله واج طويل‌ترين کلمه در نظر گرفته مي‌شود. براي کلمات کوچك‌تر از اين طول، در انتهاي کلمه واج سکوت قرارداده مي‌شود. خروجي اين شبکه نمايش باينري کلمه متناظر با دنباله واج ورودي است. با توجه به دادگان مورد استفاده در آموزش و آزمون اين مدل، ابعاد شبکه ۳۴*۱۶-۲۰-۹ در نظر گرفته مي‌شود. شکل ۵ ساختار اين شبکه را نشان مي‌دهد.



شکل۵- معماري شبکه‌عصبي دوسويه

[50] Pattern Completion
[51] Long Term

© Copyright 2008 ISBME, http://www.ijbme.org

۲-۶- عملكرد شبكه

پس از اعمال ورودي به شبكه (ورودي مي‌تواند شامل بلوك‌هاي مفقود باشد)، مطابق رابطه (۱۷)، در تكرار nام خروجي محاسبه و در تكرار $n+1$ام ورودي اصلاح مي‌شود. اين عمل N_0 مرتبه تكرار مي‌شود تا الگوي ناقص به سمت آنچه كه در تعليم شبكه، آموزش ديده است تكميل گردد. N_0 (تعداد تكرارها) به دادگان ورودي وابسته است و به‌صورت تجربي تعيين و در آموزش شبكه مورد استفاده قرار مي‌گيرد. خروجي شبكه متوسط خروجي‌هاي محاسبه شده در هر تكرار است.

$$
\begin{aligned}
& y(j,n) = f\Big(\sum_{i=0}^{n_I} x(i,n)v_{ij} + \sum_{k=1}^{n_H} y(k,n-1)w^r_{kj}\Big) \qquad j=1,\dots,n_H \\
& z(k,n) = f\Big(\sum_{j=1}^{n_H} y(j,n)w_{jk}\Big) \qquad k=1,\dots,n_O \\
& x(i,n+1) = (1-\gamma)x(i,n) + \gamma f\Big(\sum_{j=1}^{n_H} y(j,n)v^r_{ji}\Big) \qquad i=1,\dots,n_I
\end{aligned}
\tag{۱۷}
$$

در رابطه فوق، $n=1,\dots,N_0, 0\le\gamma\le 1$ است. γ كه سرعت تكميل الگوي ناقص را مشخص مي‌نمايد، به‌صورت تجربي تعيين و در آموزش و آزمون مدل استفاده مي‌شود (در اينجا $\gamma=0.7$). انتخاب نامناسب اين ضريب مي‌تواند به واگرايي منجر شود. اين ضريب ميزان اختلاط ورودي و لايه پنهان، در جايگزين‌كردن مقادير مفقود را مشخص مي‌نمايد و مي‌تواند به صورت تطبيقي نيز تغيير كند.

۳-۶-آموزش شبكه

شبكه پس از مقداردهي اوليه وزن‌ها، مطابق رابطه (۱۸) و براساس روش خاصي كه از اصول روش‌هاي مبتني بر گراديان خطا و الگوريتم پس‌انتشار خطا مشتق شده‌است، آموزش داده مي‌شود.

$$
\begin{aligned}
& \hat{z}(k,n) = \frac{1}{n}\sum_{m=1}^{n} z(k,m) \qquad k=1,\dots,n_O \\
& \dot{z}(k,n) = (\hat{z}(k,n) - d_z(k))\, f'(z(k,n)) \qquad k=1,\dots,n_O \\
& w_{jk} = w_{jk} + \eta \dot{z}(k,n) y(j,n) \qquad k=1,\dots,n_O,\ j=1,\dots,n_H \\
& \dot{y}(j,n) = f'(Y(j,n))\Big(\sum_{k=1}^{n_I} \dot{x}(k,n+1)v^r_{jk} + \\
& \quad \sum_{k=1}^{n_O} \dot{z}(k,n)w_{jk} + \sum_{i=1}^{n_H} \dot{y}(i,n+1)w^r_{ij}\Big) \qquad j=1,\dots,n_H \\
& w^r_{ij} = w^r_{ij} + \eta \dot{y}(j,n) y(i,n-1) \qquad i=1,\dots,n_H,\ j=1,\dots,n_H \\
& v_{ij} = v_{ij} + \eta \dot{y}(j,n) x(i,n-1) \qquad i=1,\dots,n_I,\ j=1,\dots,n_H \\
& \dot{x}(i,n) = (1-\gamma)\dot{x}(i,n+1) + \\
& \quad \gamma\Big(\sum_{j=1}^{n_H} \dot{y}(j,n)v_{ij}\Big) f'(x(i,n+1)) \qquad i=1,\dots,n_I \\
& v^r_{ij} = v^r_{ij} + \eta \dot{x}(k,n) y(j,n) \qquad k=1,\dots,n_I,\ j=1,\dots,n_H
\end{aligned}
\tag{۱۸}
$$

در رابطه فوق، $n=N_0-1,\dots,1$ است.

رابطه (۱۹) محاسبه خطا در آخرين تكرار را نشان مي‌دهد. خطا در تكرارهاي قبلي، با توجه به رابطه (۱۸) به‌صورت بازگشتي محاسبه مي‌شود.

$$
\begin{aligned}
& \hat{x}(k,N_0) = \frac{1}{N_0}\sum_{m=1}^{N_0} x(k,m) \qquad k=1,\dots,n_I \\
& \dot{x}(k,N_0) = (\hat{x}(k,N_0) - d_x(k))\, f'(X(k,N_0)) \qquad k=1,\dots,n_I
\end{aligned}
\tag{۱۹}
$$

در آموزش شبكه از كلمات ادا شده به وسيله ۴۸ گويش‌ور استفاده مي‌شود. در آزمون مدل، دنباله واج استخراج شده از مدل صوتي از دادگان آزمون، به ورودي مدل اعمال مي‌شود. نتايج حاصل، در جدول ۳ و ۴ مشاهده مي‌شود. لازم به ذكر است كه نتايج اين جداول به‌ترتيب با در نظر گرفتن دنباله واج تلفظ رسمي كلمه و دنباله واج برچسب‌شده در دادگان، به عنوان خروجي مطلوب محاسبه شده‌است. نتايج به‌دست آمده، بهبود نتايج مدل صوتي در قياس با تلفظ رسمي و نيز تلفظ‌هاي برچسب‌شده را نشان مي‌دهد.

جدول ۳- نتايج اصلاح واج دادگان آزمون مدل واژگاني دوسويه با در نظر گرفتن دنباله واج برچسب‌شده به عنوان خروجي مطلوب

مدل واژگاني	دقت	صحت	حذف	درج	جانشيني
شبكه عصبي تك‌سويه	٪۶۷	٪۷۴/۵	٪۱۲	٪۷/۵	٪۱۳/۵
شبكه عصبي دوسويه	٪۸۲/۵	٪۸۸/۵	٪۴/۷	٪۶	٪۶/۸

© Copyright 2008 ISBME, http://www.ijbme.org

جدول ۴- نتايج اصلاح واج دادگان آزمون مدل واژگاني دوسويه با در نظر گرفتن دنباله واج تلفظ رسمي به عنوان خروجي مطلوب

مدل واژگاني	دقت	صحت	حذف	درج	جانشيني
مدل صوتي و بدون مدل واژگاني	%۵۰/۲	%۶۵	%۱۶/۶	%۱۴/۸	%۱۸/۴
شبكه‌عصبي تك‌سويه	%۵۷/۸	%۶۹/۲	%۱۴/۶	%۱۱/۴	%۱۶/۲
شبكه‌عصبي دوسويه	%۷۱/۵	%۸۰/۴	%۹/۳	%۸/۹	%۱۰/۳

۷- بهسازي گفتار با استفاده از متوالي‌كردن مدل واژگاني با مدل صوتي

بهسازي گفتار راهكاري براي افزايش كارايي سيستم‌هاي بازشناسي گفتار در شرايط عدم تطابق الگوهاي آزمون و آموزش است. اين رويكرد بخصوص در شرايط اعوجاجي به منظور حذف اعوجاج و تميزشدن دادگان صوتي استفاده مي‌شود. تأثير اين روش در افزايش كارايي بازشناسي غيرقابل انكار است. رايج‌ترين روش در بهسازي گفتار تفاضل طيفي، روشهاي مبتني بر HMM و فيلتر كالمن[۵۲] است. تفاضل طيفي ساده‌ترين روش از نظر پياده‌سازي است، اما از چندين مشكل رنج مي‌برد.

۱. اين روش، به فرض پيشيني از تخمين اعوجاج احتياج دارد. به‌عبارت ديگر تخمين اعوجاج مناسب، پيش‌نياز اين روش است كه براي اعوجاج غيرايستا قابل دسترسي نيست.

۲. مشكل جدي ديگر اين روش از فراتخمين[۵۳] اعوجاج ناشي مي‌شود. اين حالت سبب مي‌شود نواحي زماني-طيفي كم انرژي، پس از كسر اعوجاج، منفي شوند. به‌طور سنتي اين مقادير منفي با صفر جايگزين مي‌شوند كه گفتار بازسازي‌شده اثر اضافه‌شدن اعوجاج موزيكي را دارد.

روش مرسوم در غلبه بر اعوجاج موزيكي، جايگزين كردن مقادير منفي با كف اعوجاج است. مشكل روش‌هاي مبتني بر مدل و فيلتر، ساختار پيچيده آنهاست [۲۰]. شبكه‌هاي‌عصبي، بخش ديگري را علاوه بر روش‌هاي سنتي بهسازي گفتار در هر دو حوزه زمان و فركانس فراهم مي‌آورند [۲۰]. مزيت استفاده از شبكه‌هاي‌عصبي در بهسازي حوزه طيف، در بسياري از موارد حذف اعوجاج موزيكي است. شبكه‌هاي‌عصبي همچنين در مواردي كه اختلاط اعوجاج با گفتار به‌صورت غيرخطي است، مناسب به‌نظرمي‌رسند. روش‌هاي بهسازي گفتار با شبكه‌هاي‌عصبي به بخش‌هاي فيلتركردن حوزه-زمان، نگاشت حوزه-تبديل، راه‌گزيني مدل وابسته به حالت و روش‌هاي تكراري برخط طبقه‌بندي مي‌شود.

در اينجا به منظور بهسازي گفتار، مدل واژگاني مطرح‌شده در بخش قبل با مدل صوتي متوالي مي‌شود. سپس با استفاده از روش‌هاي معكوس سازي شبكه‌هاي عصبي كه در بخش سوم بيان‌شد، اصلاح دنباله واج در بهسازي متغيرهاي بازنمايي به‌كار گرفته مي‌شود. به‌اين ترتيب روش‌هاي به‌كاربرده‌شده قادرند تا با تلفيق دانش سطوح بالا با سطوح پايين، سبب بهسازي گفتار شوند. در ادامه با‌به‌كارگيري دو روش معكوس‌سازي بخش سوم، يعني معكوس‌سازي به‌روش گراديان و شبكه معكوس كه در اين بخش به‌ترتيب معكوس‌سازي با اصلاح ورودي به‌روش گراديان و معكوس‌سازي شبكه به‌روش آموزش شبكه‌معكوس نامگذاري شده است، نتايج قابل ملاحظه‌اي در افزايش صحت بازشناسي واج حاصل مي‌شود.

۷-۱-بهسازي گفتار با استفاده از معكوس سازي با اصلاح ورودي به روش گراديان

در اين روش پس از استخراج دنباله واج از دنباله فريم خروجي مدل صوتي، با استفاده از شبكه‌عصبي مطرح‌شده در بخش ششم، از مدل واژگاني شبكه‌عصبي دوسويه مبتني بر داده‌هاي مفقود براي تصحيح دنباله واج به‌دست آمده به دنباله واج برچسب‌شده، استفاده مي‌شود. با مقايسه دنباله واج خروجي مدل واژگاني و دنباله واج ورودي آن (خروجي مدل صوتي) و پس‌انتشار خطاي اين دو به سطح دادگان صوتي (به اين ترتيب كه دنباله واج به‌دست آمده از مدل واژگاني به عنوان خروجي مطلوب در نظر گرفته مي‌شود و با پس‌انتشار خطاي مابين خروجي مطلوب و خروجي فعلي به سطح ورودي) ورودي به روش گراديان خطا اصلاح مي‌شود. به اين ترتيب تمام دادگان آزمون (متغيرهاي بازنمايي) اصلاح مي‌شوند.

---

[52] Kalman Filtering
[53] Overestimation

© Copyright 2008 ISBME, http://www.ijbme.org

حال اگر مدل صوتي جديدي كه بر روي دادگان تعليم اصلاح شده، آموزش ديده است به همراه مدل واژگاني مذكور بر روي دادگان صوتي آزمون اصلاح شده، به‌كار گرفته شود، بهبود نتايج بازشناسي واج طبق جدول ۵ مشاهده مي‌شود. شكل ۶ نمودار مربوط به اين روش را نشان مي‌دهد.

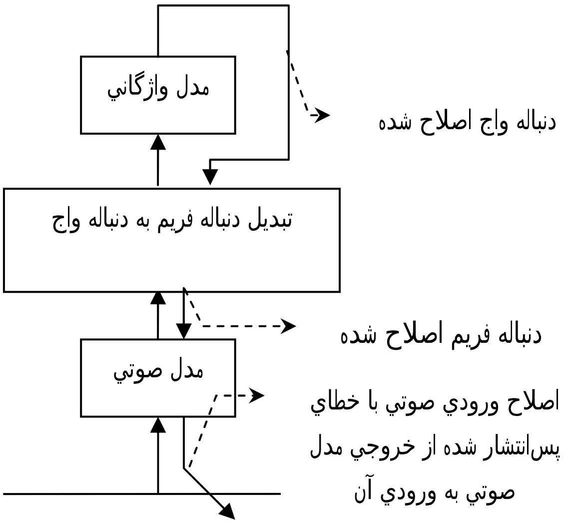
۲-۷- تلفيق با استفاده از معكوس سازي شبكه به روش آموزش شبكه معكوس

در اين روش به علت نياز به شبكه معكوس مدل صوتي هنگام آموزش مدل صوتي، معكوس اين مدل نيز آموزش داده مي‌شود (ورودي مدل معكوس، دنباله فريم و خروجي آن متغيرهاي بازنمايي است). به اين ترتيب مدل معكوس قادر خواهد بود تا از خروجي مدل صوتي، تقريبي از ورودي آن يعني دادگان صوتي را به‌دست آورد.

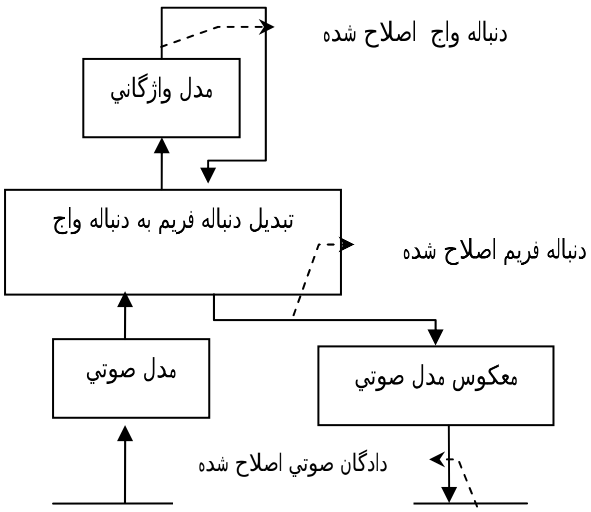
مانند روش مطرح شده در بخش قبلي، پس از به‌دست آوردن دنباله واج مدل واژگاني شبكه‌عصبي دوسويه مبتني بر داده‌هاي مفقود، دنباله فريم (خروجي مدل صوتي) مطابق اين خروجي اصلاح مي‌شود. سپس با استفاده از شبكه معكوس مدل صوتي، دادگان صوتي جديدي از روي دنباله فريم اصلاح شده به‌دست مي‌آيد. اعمال متغيرهاي بازنمايي به‌دست‌آمده از معكوس مدل صوتي، به مدل صوتي جديدي كه مطابق روش قبلي بر روي دادگان آموزش اصلاح شده، تعليم ديده است و متوالي‌شدن مدل واژگاني با آن، افزايش قابل ملاحظه‌اي در صحت بازشناسي واج دادگان آزمون اصلاح شده مطابق نتايج جدول ۵ از خود نشان مي‌دهد. شكل ۷ نمودار كلي طرح را نشان مي‌دهد.

جدول۵- نتايج اصلاح واج دادگان آزمون با روش‌هاي تلفيق دانش زباني با دادگان صوتي (تشخيص مرز واج از برچسب دادگان)

روش تلفيق	دقت	صحت	حذف	درج	جانشيني
روش گراديان	٪۸۳/۴	٪۸۹/۱	٪۴/۳	٪۵/۷	٪۶/۶
شبكه معكوس	٪۸۳/۹	٪۸۹/۶	٪۴/۱	٪۵/۷	٪۶/۳



شكل ۶- نمودار كلي تلفيق دانش زباني با اطلاعات صوتي به روش معكوس‌سازي با اصلاح ورودي به روش گراديان



شكل ۷- نمودار كلي تلفيق دانش زباني با اطلاعات صوتي به روش معكوس‌سازي شبكه به روش آموزش شبكه معكوس

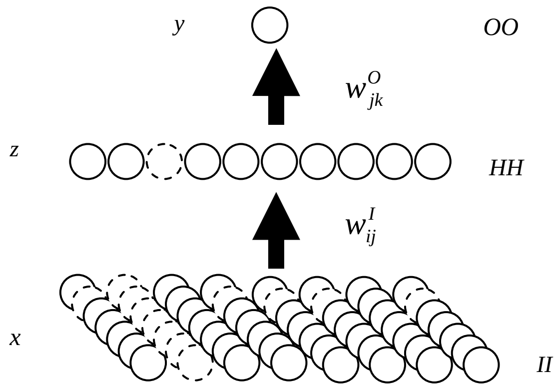
۳-۷-شناسايي مرز واج با استفاده از شبكه‌عصبي (شبكه مرزشناس[۵۴])

در روش‌هاي مطرح‌شده، در اصلاح دنباله فريم از دنباله واج اصلاح شده، اطلاع از موقعيت مرز واج كاملاً ضروري است. نتايج به‌دست آمده قبلي با در نظرگرفتن مرز واج از برچسب دادگان، محاسبه شده‌اند. به منظور عملي‌تر كردن روش‌هاي فوق، مرز واج به وسيله شبكه مرزشناس و از روي متغيرهاي بازنمايي(دادگان صوتي) شناسايي مي‌شود. به اين منظور از

[54] Border Detector

© Copyright 2008 ISBME, http://www.ijbme.org

يك شبكه‌عصبي MLP با يك لايه پنهان و با ساختار نوروني ۹*۳۹-۵۰-۱ استفاده مي‌شود. در ورودي اين شبكه ۹ فريم متوالي (٤ فريم مجاور راست و ۴ فريم مجاور چپ) قرار داده مي‌شوند، تا وجود يا نبود مرز دو واج در فريم مياني مشخص شود. در خروجي شبكه يك نورون با تابع غيرخطي سيگمويید وجود دارد كه نورون مرزشناس ناميده مي‌شود. نحوه برچسب‌دهي نورون مرزشناس چنين است: زماني‌كه مرز بين دو واج دقيقاً در وسط ورودي شبكه قرارگيرد خروجي "۱"، يك فريم جلوتر و عقب‌تر خروجي "۰/۷۵"، دو فريم جلوتر و عقب‌تر خروجي "۰/۲۵" و در غير از اين موارد خروجي "۰" خواهد داشت. شكل ۸ ساختار اين شبكه و جدول ۶ نتايج بازشناسي واج حاصل از روش‌هاي تلفيق، با مرز واج به‌دست آمده از شبكه مرزشناس را نشان مي‌دهد. در نتايج مذكور نيز از مدل شبكه‌عصبي دوسويه مبتني بر داده‌هاي مفقود، كه بهترين نتايج را دربرداشته است، به‌عنوان مدل واژگاني استفاده شده‌است.



شكل۸- ساختار شبكه مرزشناس

جدول۶- نتايج اصلاح واج دادگان آزمون با روش‌هاي تلفيق دانش زباني با دادگان صوتي (تشخيص مرز واج با شبكه مرزشناس)

روش تلفيق	دقت	صحت	حذف	درج	جانشيني
روش گراديان	%۸۲/۵	%۸۸/۵	%۴/۷	%۶	%۶/۸
شبكه معكوس	%۸۳/۱	%۸۹/۱	%۴/۶	%۶	%۶/۳

۸-بحث و نتيجه‌گيري

بررسي سيستم درك و بازشناسي در انسان نشان مي‌دهد كه اين سيستم يك ساختار سلسله مراتبي و دوسويه است [۱۵، ۱۶، ۲۰، ۲۱]. استفاده از دانش پيشين در سيستم ASR، به‌منظور مقابله با اثرات هم‌توليدي آواها، نتايج مناسبي به همراه داشته‌است [۲۲]. استفاده از اطلاعات سطوح فوقاني سيستم بازشناسي تصوير در تفسير و پردازش دادگان ورودي، كارايي و انعطاف‌پذيري اين سيستم را به نحو بسيار مطلوبي افزايش داده است [۱۶]. دوسويه كردن شبكه‌هاي عصبي جلوسو، افزايش قابل ملاحظه‌اي در كارايي اين شبكه‌ها و تشكيل بسترهاي جذب پويا در آنها داشته‌است [۲۳-۲۵]. در ادامه، تلفيق دانش واژگاني به‌عنوان دانش پيشين و سطح بالا با متغيرهاي بازنمايي، مي‌تواند افزايش مجدد صحت بازشناسي را نشان دهد. دوسويه بودن پردازش، اين امكان را فراهم مي‌سازد تا تنوعات يا عدم تطابق الگوهاي ورودي، تا حد ممكن براساس دانش سطوح فوقاني برطرف شود.

نتايج به‌دست آمده از مدل‌هاي واژگاني مطرح شده، از دو جهت مختلف مي‌تواند سبب بهبود كارايي يك سيستم بازشناسي گفتار شود. از يك‌سو با افزايش صحت بازشناسي واج در مقايسه با تلفظ رسمي، حجم واژه‌نامه مورد نياز براي در برداشتن تلفظ‌هاي چندگانه كلمات كاهش مي‌يابد. لذا با واژه‌نامه كوچك‌تر، صحت بازشناسي كلمه همچنان در حد قابل قبولي باقي مي‌ماند. از سوي ديگر مي‌تواند با تصحيح دنباله واج سبب ارتقاي صحت بازشناسي واج مدل صوتي شود. زيرا مقايسه دنباله واج توليدشده در خروجي مدل واژگاني با دنباله واج برچسب‌شده در دادگان نيز حكايت از افزايش صحت دارد. با توجه به موضوع اخير، اكنون اين سؤال مطرح مي‌شود كه با وجود آموزش كافي مدل صوتي با استفاده از واج برچسب‌شده در دادگان، علت افزايش صحت واج اخير چيست؟ بخشي از پاسخ اين سؤال را بايد در تنوعات دادگان صوتي جستجو نمود. اين تنوعات كه به علل مختلف در سطح دادگان صوتي ظاهر مي‌شوند، با وجود آموزش كافي مدل صوتي، مي‌تواند سبب خطا در بازشناسي

© Copyright 2008 ISBME, http://www.ijbme.org

واج مدل صوتي و عدم تعميم‌دهي بهينه شود. حال اگر اختلاف ايجادشده ميان دنباله واج خروجي مدل واژگاني و ورودي آن (خروجي مدل صوتي)، تنها در دادگان صوتي جستجو شود، تلفيق اطلاعات نهفته در اين اختلاف، با دادگان صوتي در ورودي مدل صوتي، مي‌تواند مفيد باشد. به اين منظور، اطلاعات سطوح بالاتر با اطلاعات سطوح پايين تلفيق مي‌شود. در ادامه با به‌كارگيري دو روش، يعني معكوس‌سازي به روش گراديان و شبكه معكوس كه در اين بخش به‌ترتيب معكوس‌سازي با اصلاح ورودي به روش گراديان و معكوس‌سازي شبكه به‌روش آموزش شبكه معكوس ناميده مي‌شود، نتايج قابل ملاحظه‌اي در افزايش صحت بازشناسي واج حاصل مي‌شود.

مراجع

[1] Bimbot F., Chollet G., Paoloni A., Assessment methodology for speaker identification and verification systems: An overview of SAM-A Esprit project 6819 - Task 2500. *ESCA Workshop on Automatic Speaker Recognition Identification and Verification* 1994; 75-82.

[2] Lippmann R.P., Speech recognition by machines and humans; Speech Communication 1997; 22:1-15.

[3] Gong, Y.; Speech recognition in noise environments: a survey; Speech Communication 1995; 16:261-291.

[4] Miller G.A., Licklider J.C.R., The intelligibility of interrupted speech; Journal of the Acoustic Society of America 1950; 22:167-173.

[5] Fletcher H., Speech and Hearing in Communication. Journal of the Acoustic Society of America 1953; 28:164-172.

[6] Furui S., Recent advances in robust speech recognition; Speech Communication 1997; 22:27-39.

[7] Lockwood P., Boudy J., Experiments with non-linear Spectral Subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars; Speech Communication 1992; 11: 215-228.

[8] Diamantaras K.I., Neural networks and principal component analysis, In: Handbook of neural network signal processing; CRC Press, 2002.

[9] Cooke M., Morris A., Green P., Recognizing Occluded Speech; ESCA Tutorial and Workshop on the Auditory Basis of Speech Perception 1996, Keele University, 15-19.

[10] Jensen C.A., Reed R.D., Marks R.J., Inversion of Neural Networks: Algorithms and Applications; IEEE ,Neural Networks 1999; 87:1536-1549.

[11] Williams R.J., Inverting a Connectionist Network Mapping by Backpropagation of Error; Proc. 8th Annu. Conf. Cognitive Science.

[۱۲] سيد صالحي سيدعلي؛ افزايش كارايي بازشناخت الگوي شبكه‌هاي‌عصبي جلوسو از طريق توسعه روش‌هايي براي دوسويه كردن عملكرد آنها؛ گزارش طرح پژوهشي، دانشگاه صنعتي اميركبير، دانشكده مهندسي پزشكي، مهر ۱۳۸۳.

[13] Bijankhan M., Seikhzadeghan J., Roohani M.R., Samareh Y., Lucas K. Tebyani M., FARSDAT-the speech Database of Farsi Spoken Language; Proc. of SST94 1994;826-831.

[۱۴] ولي منصور، سيدصالحي سيدعلي؛ ارزيابي كارايي دو بازنمايي MFCC و LHCB در بازشناسي مقاوم به تنوعات گفتار مستقيم و تلفني، مجموعه مقالات دهمين كنفرانس سالانه انجمن كامپيوتر ايران، ۱۳۸۳، ۳۰۵-۳۱۲.

[15] Nguyen D., Widrow B., Neural Networks for Selflearning Control Systems; IEEE Control Systems Magazine 1990; 10:18-23.

[16] Koerner E., Gewaltig M.O., Koerner U., Richter A., Rodemann U., A Model of Computation in Neocortical Architecture; Neural Networks Elsevier Science 1999;12: 989-1005.

[17] Koerner E., Tsujino H., Masutani T., A cortical type modular neural network for hypothetical reasoning; Neural Networks , Elsevier Science 1997; 10: 791-814.

[18] Koerner E., Matsumoto G., Cortical architecture and self-referential control for brain-like computation; Engineering in Medicine and biology Magazine, IEEE 2002; 21:121-133

[19] Wan E, Nelson AT; Networks for Speech Enhancement. In: Handbook of Neural Networks for Speech Processing, 1998:541-541.

[20] Ghosen J., Bengio Y., Bias Learning, Knowledge sharing; IEEE Trans. On Neural Networks 2003; 14:84-108.

[21] Mesulam M.M., From Sensation to Cognition; Brain, Oxford Univ. Press 1998, 121:1013- 1052.

[۲۲] انصاري ليلا؛ مدل‌سازي اثرات هم‌توليدي آواها در يك مدل شبكه‌عصبي بازشناخت گفتار؛ پايان‌نامه كارشناسي ارشد بيوالكتريك، دانشكده مهندسي پزشكي، دانشگاه صنعتي اميركبير، ۱۳۸۲.

[23] Saul L.K., Jordan M.I.; Attractor dynamics in feed forward neural networks; Neural Computation,Massachusetts Institute of Technology 2000; 12: 1313-1335.

[24] Trappenberg, T., Continuous attractor neural networks. In L. N. de Castro & F. J. V. Zuben (Eds.), Recent developments in biologically inspired computing. Hershey, PA: Idee Group; 2003.

[25] Wu Y., Pados D.A.; A feedforward bidirectional associative memory; IEEE Trans. On Neural Networks 2000;11:42.

© Copyright 2008 ISBME, http://www.ijbme.org