نشریه تحقیقاتی فنی ـ مهندسی استقلال، دانشگاه صنعتی اصفهان، شماره ۱۳ شهریورماه ۱۳۷۲

کاربرد مدلهای پنهانی مارکف برای درک مستقل از گویندهٔ کلمات منفرد فارسی

بابک عظیمی سجادی* ـ وحید طباطبائی* ـ سید بهرام ظهیر اعظمی** ـ کارولوکس***

چکیده

دراین مقاله نحوه تحقق یک سیستم درک صحبت به روش مدل پنهانی مارکف با مشاهدهٔ پیوسته تشریح می گردد. از این مدل جهت تشخیص تعداد محدودی از کلمات (اعداد «صفر» تا «ده») که به زبان فارسی بیان شده اند به صورت مستقل از گوینده استفاده شده است. دراین روش برای هرکلمه یک مدل چپ به راست در نظر گرفته شده است. پارامترهای مدل با تخمین بیشترین شباهت از روی داده های یادگیری تعیین می شوند. این تخمین از روش k مقدار متوسط درهرقسمت به دست می آید. برای تعیین مقادیر اولیهٔ پارامترهای مدل، روشی جدید بر مبنای چندی کردن برداری ارائه می شود که با استفاده از آن تمامی مراحل طراحی سیستم درک کنندهٔ صحبت به صورت خودکار انجام می شود. کارآیی سیستم محقق شده ۹۷/۰۰٪ است.

---

* فارغ التحصیلان کارشناسی ارشد دانشکده فنی دانشگاه تهران و کارشناسان مرکز تحقیقات الکترونیک دانشگاه صنعتی شریف

** کارشناس مرکز تحقیقات الکترونیک دانشگاه صنعتی شریف

*** دانشیار دانشکده فنی دانشگاه تهران

مقدمه

ارتباط گفتاری بین انسان و ماشین نسبت به سایر موارد، مزایای عمده ای از قبیل سهولت کاربرد و عدم نیاز به آموزش‌های پیچیده دارد. از این رو توجه بسیاری از مراکز تحقیقاتی و صنعتی جهان به این مسأله جلب شده است. در ارتباط گفتاری، کاربر دستورات و یا اطلاعات را به وسیله گفتار به کامپیوتر انتقال می دهد، کامپیوتر هم با بیان کلمات و یا جملاتی با استفاده کننده ارتباط برقرار می نماید. برای ایجاد چنین امکانی کامپیوتر ابتدا باید صحبت گوینده را درک کند و سپس بتواند منظور خود را توسط کلام بیان نماید.

در این مقاله روشی برای ارتباط انسان با ماشین یا درک صحبت توسط کامپیوتر پیشنهاد می شود.[1] برای طراحی یک سیستم درک صحبت ابتدا باید بتوان نحوهٔ تولید صحبت را مدل کرد. تولید صحبت در انسان عملی است هماهنگ که در آن اندامهای گفتاری دخالت دارند. در زمان ادای هر واج[2] این اندامها در وضعیت خاصی قرار می گیرند به طوری که امواج صوتی حاصل از این عمل خواص معینی به دست می آورند. ولی تفاوت هایی در محل قرارگرفتن اندامها، محل وقوع هر واج در بین واجهای دیگر و تفاوتهای فیزیکی و هندسی اندامهای گفتاری افراد مختلف خواص صوتی متفاوتی را در بین واجهای بیان شده ایجاد می کند[۲] این تفاوتها به شکل وسیعتری بین هجاها[3]، واژه ها و جملات نیز پدید می آید ([۱ و ۲]).

مسأله اصلی در درک صحبت، استخراج خصوصیات مشترک در اشکال مختلف ادای هر واج (واژه) است. این خصوصیات باید دارای بیشترین شباهت در اداهای مختلف یک واج (واژه) و بیشترین اختلاف در ادای واجهای (واژه های) گوناگون باشد.

درک صحبت در دو سطح معنایی[4] و ساختاری[5] مطرح می شود. سطح ساختاری که در این مقاله مورد نظر است، از سه جنبهٔ تعداد کلمات قابل تشخیص، محدودیت های گفتاری در بیان کلمات و

---

1 . Speech Recognition

۲ ـ حرف قابل گویش در یک زبان (Phoneme)

۳ ـ رشته ای پیوسته و بدون مکث از واجها (Syllable)

4 . Semantic

5 . Syntactic

تنوع استفاده کنندگان تقسیم بندی می شود. هر اندازه محدودیتهای کمتری برای استفاده کننده قائل شویم به سیستم پیچیده تری نیاز خواهیم داشت. به طوری که هنوز سیستم قابل اعتمادی که بدون هیچ محدودیتی درک صحبت را انجام دهد ارائه نشده است. با وجود پیشرفت های قابل ملاحظه حتی طراحی سیستم هایی که محدودیتهایی برای استفاده کننده قائل می شوند، هنوز هم از مسایل در درست تحقیق به شمار می رود[۲].

سه روش عمده ای که در حال حاضر برای درک صحبت به کار می روند عبارتند از:

۱- مقایسه پارامترهای سیگنال صحبت با یک سری پارامترهای نمونه[1] ([۳ تا ۵])

۲- استفاده از شبکه های عصبی ([۶ و ۷])

۳- تشخیص صحبت براساس خصوصیات آماری سیگنال صحبت ([۸ تا ۱۰])

در این مقاله تحقق روش مدل پنهانی مارکف[2] به عنوان یک روش آماری موفق ارائه خواهد شد. با استفاده از این مدل یک سیستم درک صحبت مستقل از گوینده[3] برای تشخیص اعداد "صفر" تا "ده" که به صورت تک کلمه ای[4] به زبان فارسی ادا می شوند طراحی می شود.

در بخش بعدی مدل پنهانی مارکف معرفی شده، مسائل و مشکلات آن توضیح داده می شود. سپس به نحوه استخراج مشخصه های سیگنال صحبت خواهیم پرداخت. آنگاه روش انتخاب مقادیر اولیه و مسائل عملی بررسی می شود و پس از آن نحوهٔ آزمایش و نتایج آن ارائه خواهد شد. در انتها کارآیی روش ارائه شده مورد ارزیابی قرار می گیرد و علیرغم وجود تفاوت اصولی در نتایج این آزمایشها با نتایج به دست آمده توسط دیگر محققان برای زبان انگلیسی، جهت تکمیل مطلب مقایسه ای صورت می گیرد.

معرفی مدل پنهانی مارکف ([۱۰ تا ۱۲])

در این روش فرض می شود که مشخصه های سیگنال صحبت یک فرایند مارکف مرتبهٔ اول با

---

1 . Dynamic Time Warping (DTW)

2 . Hidden Markov Model(HMM)

3 . Speaker Independent

4 . Isolated Word

مشاهده ناقص[1] است. منظور از فرایند مارکف مرتبهٔ اول فرایندی است که در آن:

$$p\ (x_n \mid x_{n-1}, x_{n-2}, \ldots, x_0) = p\ (x_n \mid x_{n-1}) \qquad (۱)$$

به عبارت دیگر با داشتن حالت در لحظه n-1 ، حالت در لحظه n از لحظات قبل از n-1 مستقل است. مثالی خوب برای فرایند مارکف یک ماشین حالت[2] است که گذار حالت در آن به صورت تصادفی با احتمال های مشخص انجام می شود. حال اگر در فرایند مارکف به جای حالت، یک متغیر تصادفی وابسته به حالت مشاهده شود، به آن فرایند مارکف با مشاهده ناقص می گویند. به مدل چنین فرایندی، مدل پنهانی مارکف گفته می شود. مدل پنهانی مارکف مدل خوبی برای بیان رفتار سیگنال صحبت است.

برای تشخیص صحبت به روش فوق به ازای هر یک از کلمات مقطع، یک مدل مارکف در نظر گرفته می شود. در مرحله یادگیری ابتدا پارامترهای مدلها تخمین زده شده، سپس در مرحله تشخیص از مدلهای تخمین زده شده استفاده می گرددو کلمه ای که مدل متناظر با آن دارای بیشترین احتمال باشد، انتخاب می شود.

پارامترهای مدل پنهانی مارکف در حالت مشاهده گسسته عبارتند از:

۱ - N تعداد حالتها، حالتهای ممکن $S_1, \ldots, S_N$

۲ - M تعداد الفبای مشاهده، الفبای مشاهده $V_1, \ldots, V_M$

۳ - A ماتریس گذار حالت،

$$A = \{a_{ij}\};\ a_{ij} = p[q_{t+1} = S_j \mid q_t = S_i] \qquad (۲)$$

دراین رابطه $1 \le i,j \le N$ و t مشخص کنندهٔ زمان است.

---

1 . Incomplete Observation 2 . State Machine

۴ - B ماتریس احتمال صدور الفبای k در حالت j،

$$B = \{b_{jk}\} \; ; \; b_{jk} = p[V_k \text{at } t | \; q_t = S_j] \quad ; \quad 1 \le j \le N \qquad (۳)$$

در این جا $1 \le k \le M$ و t مشخص کنندهٔ زمان است.

۵ - π توزیع آغازین حالت،

$$\pi = \{\pi_i\} \; ; \; \pi_i = p[q_1 = s_i] \quad ; \quad 1 \le i \le N \qquad (۴)$$

سلسله[1] مشاهدات را با O و سلسله حالات طی شده را با Q نمایش می دهیم:

$$O = o_1 o_2, \ldots, o_T \qquad (۵)$$

$$Q = q_1 q_2, \ldots, q_T \qquad (۶)$$

در این رابطه T کل زمان مشاهده است. در ضمن جهت اختصار، مدل پنهانی مارکف با λ نمایش داده می شود:

$$\lambda = (A, B, \pi) \qquad (۷)$$

در مراحل مختلف یادگیری و تشخیص صحبت توسط این مدل سه مسأله وجود دارد که عبارتند از:

---

1 . Sequence

مسأله اول : در مرحله تشخیص باید احتمال وقوع یک مشاهده را به شرط انتخاب هر مدل محاسبه کرد. به عبارت دیگر به ازای هر کلمه (در این جا ۱۱ کلمه) یک مدل λ_i وجود دارد و صحبت ادا شده به سلسله مشاهده O تبدیل شده است. حال، محاسبه $p(O|\lambda_i), i = 0, \ldots, 10$ مد نظر خواهد بود. پس از محاسبه، مقدار ماگزیمم مشخص شده و کلمه متناظر با مدل دارای بیشترین احتمال، به عنوان کلمه برگزیده معرفی می گردد.

در صورت استفاده از روش مستقیم در محاسبه $P(O|\lambda_i)$ یعنی محاسبه

$$P(O|\lambda) = \sum_{Q} P(O|Q, \lambda).P(Q|\lambda) \qquad (۸)$$

$$= \sum_{q_1,q_2,\ldots,q_T}^{allQ} \Pi_{q_1}\, b_{q_1}(o_1) a_{q_1q_2}\, b_{q_2}(o_2) \ldots a_{q_{T-1}q_T}\, b_{q_T}(o_T)$$

مرتبه محاسبات $2T.N^T$ خواهد بود که برای $T = ۴۰$ و $N = ۶$ برابر $۱۰^{۳۳}$ می شود. این میزان محاسبات از توان کامپیوترهای موجود خارج است.

با استفاده از روش بازگشتی[1] می توان در هر مرحله احتمال $P(o_1, \ldots, o_t, q_t = S_i|\lambda)$ را از روی اطلاعات یک مرحله قبل به صورت زیر محاسبه کرد:

$$P(o_1, \ldots, o_t, q_t = S_i|\lambda) = \qquad (۹)$$

$$\sum_{j=1}^{N} a_{ji}\, P(o_1, \ldots, o_{t-1}, q_{t-1} = S_j|\lambda).b_i(o_t) \; ; \; i=1, \ldots, N$$

و نهایتاً با داشتن احتمالات در زمان T، $P(O|\lambda)$ را محاسبه نمود.

---

1 . Recursive

$$P(O|\lambda) = \sum_{i=1}^{N} P(o_1 , \dots , o_T, q_T = S_i|\lambda) \qquad (۱۰)$$

مرتبه محاسبات در این روش $T.N^2$ است که برای مثال قبل $(N = 6 , T = 40)$ برابر ۱۴۴۰ می شود. در نتیجه محاسبه $P(O|\lambda)$ از این روش توسط کامپیوتر عملی خواهد بود.

مسأله دوم: در مرحله یادگیری لازم است محتمل ترین مسیر حالت، با داشتن مشاهده O و مدل λ محاسبه شود. یعنی سلسله Q را به نحوی انتخاب کنیم که $P(Q|O,\lambda)$ ماگزیمم شود. این کار معادل است با محاسبه

$$Q = \underset{q_1 q_2 \dots q_T}{\text{Arg Max}} \; P[Q, O|\lambda] \qquad (۱۱)$$

بر اساس اصل بهینگی[1] می دانیم اگر مسیر منتهی به هر حالت در زمان t بهترین مسیر ممکن (مسیر بهینه) باشد، با انتخاب یک سیاست بهینه در فاصله زمانی (t+1, T) در نهایت به یک سری جواب بهینه برای مسیرهای منتهی به هر حالت در زمان T خواهیم رسید [۱۳]. بهترین مسیر، مسیری است که تابع ارزش مربوط به آن حداکثر باشد. در این مسأله خاص تابع ارزش، احتمال هر مسیر بوده و هدف ماگزیمم کردن آن است. روش ویتربی[2]، روشی تکراری برای یافتن مسیر بهینه است [۱۳]. شکل ۱ این روش را به طور خلاصه نشان می دهد.

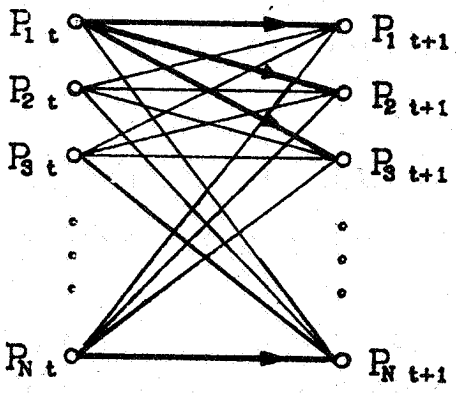
مسیر انتخاب شده درگذار از زمان t به t+1 برای هر حالت i ، بر اساس تابع ارزش در زمان t از رابطهٔ زیر محاسبه می شود:

$$P_i(t+1) = \underset{j}{\text{Max}} \, (P_j(t). \, a_{ji}) \, . \, b_i(o_{t+1}) \qquad (۱۲)$$

---

1 . Principle of Optimality

2 . Viterbi Algorithm



شکل ۱ - نمایش انتخاب مسیرهای بهینه با روش ویتربی، تا لحظه t+1 با استفاده از مسیرهای بهینه در لحظه t

به غیر از مسیری که رابطه بالا را برای هر حالت ماگزیمم می کند، بقیه مسیرها حذف می شوند. (مسیرهایی که در شکل با خطوط نازکتر رسم شده ، حذف گردیده اند.) بنابراین در هر گذار زمان فقط N مسیر حفظ خواهد شد، در خاتمه در زمان T، N مسیر بهینه که هر کدام به یکی از حالت های $S_1, \ldots, S_n$ ختم شده اند باقی می مانند و مسیر بهینه مسیری است که احتمال آن بیشترین مقدار را داشته باشد.

مسأله سوم : در مرحله یادگیری لازم است مدلی که بیشترین شباهت را به مشاهده O دارد، تخمین زده شود. برای حل این مسأله روشهای متفاوتی وجود، که هیچکدام رسیدن به جواب بهینه را تضمین نمی کند. روشی که برای طراحی سیستم حاضر در نظر گرفته شده عبارت است از تخمین درستنمایی بیشینهٔ[1] مدل λ به فرض آنکه مسیر Q طوری انتخاب شده باشد که $P(O,Q|\lambda)$ حداکثر شود([۱۰ تا ۱۲]):

$$\lambda = \underset{\lambda}{\text{ArgMax}} \{\underset{Q}{\text{Max}}\, P[Q,\ O|\lambda]\} \qquad (۱۳)$$

---

1 . Maximum Likelihood

جهت حل این مسأله ابتدا از یک روش مناسب حدس اولیه ای برای مدل λزده می شود، سپس با توجه به سلسله های مشاهدات با یک روش تکراری تخمین بهتری ($\bar{\lambda}$) از روی مقدار قبلی λ، به دست می آید[۱۴]. این عمل تا رسیدن به λ ی بهینه تکرار می شود. در به دست آوردن روابط تکراری تخمین λ ، مشاهده می شود که تخمین ضرائب ماتریس B مستقل از تخمین ضرائب ماتریس A است. این امر باعث می شود که روش به دست آمده تا اندازه ای ساده شود. به طور خلاصه این روش که "روش k مقدار متوسط برای هر قسمت[1]" نامیده می شود، به صورت زیراست [۱۴]:

۱- به دست آوردن مسیر حالت Q^* برای مشاهده O با ماگزیمم کردن تابع $P(O, Q|\lambda)$ (حل مساله دوم).

۲- بدست آوردن مجموعه های T_{ij} ; $1 \le i, j \le N$:

$$T_{ij} = \{t : \overset{*}{q}_{t-1} = i \ , \ \overset{*}{q}_t = j\} \qquad (۱۴)$$

۳- به دست آوردن a_{ij} از روی مجموعه های T_{ij} به صورت زیر:

$$\bar{a}_{ij} = \frac{\| T_{ij} \|}{\sum_{k=1}^{N} \| T_{ik} \|} \ ; \ 1 \le i, j \le N \qquad (۱۵)$$

که در آن $\| \ . \ \|$ برای نمایش تعداد اعضای یک مجموعه بکار می رود.

تعبیر رابطهٔ بالا بسیار ساده است، در واقع تعداد دفعات گذار از حالت i به j محاسبه شده و به تعداد دفعات حضور در حالت i تقسیم می شود.

۴- مجموعه های O_i ; $i = 1, \ldots, N$ به صورت زیر بدست می آید:

---

1 . Segmented K-Means

$$O_i = \{o = o_t \; ; \; q_t = S_i\} \qquad (۱۶)$$

مجموعه O_i شامل کلیه مشاهداتی است که در حالت S_i اتفاق افتاده اند.

۵- تخمین ماتریس B از روی مجموعه های O_i به دست می آید. اگر مشاهدات به صورت گسسته باشند، b_{ik} معادل است با تکرار V_k در مجموعهٔ O_i، تقسیم بر کل تعداد اعضای مجموعه O_i.

تا کنون فرض بر این بود که مشاهدات، کمیات گسسته ای هستند. در حالی که می توان مشاهدات را به صورت بردارهایی پیوسته نیز در نظر گرفت. در این صورت سلسله مشاهدات $O = \underline{o}_1 \; \underline{o}_2 \ldots \underline{o}_T$ است که $\underline{o}_i$ یک بردار پیوسته است. در نتیجه به جای تخمین تابع توزیع احتمال گسسته، (ماتریس B) باید تابع چگالی احتمال پیوسته بردار $\underline{o}$ برای هر حالت تخمین زده شود. به منظور تسهیل مسأله تخمین تابع چگالی احتمال برای هر حالت، مجموع M تابع توزیع نرمال فرض می شود.

$$b_j(\underline{o}) = \sum_{m=1}^{M} c_{jm} . N(\underline{o} \; , \underline{\mu}_{jm}, \sigma_{jm}) \qquad (۱۷)$$

در رابطهٔ ۱۷، M تعداد خوشه ها[1] (ترکیب ها) و c_{jm} ضریب تأثیر خوشهٔ m در حالت j است که به صورت $P(m|j)$ احتمال شرطی خوشهٔ m در حالت j نیز در نظر گرفته می شود. $\underline{\mu}_{jm}$ بردار متوسط تابع توزیع خوشهٔ m در حالت j، σ_{jm} ماتریس کواریانس خوشهٔ m در حالت j، و N نمایانگر تابع توزیع نرمال برداری است. تابع توزیع پیوستهٔ $b_j(\underline{o})$ بدین علت به صورت ترکیبی از چند تابع توزیع نرمال در نظر گرفته شده است که معمولاً توابع توزیع واقعی برای مشاهدهٔ $\underline{o}$ پیچیده تر از آن است که با یک تابع توزیع نرمال قابل توصیف باشد. برای نمونه ممکن است چندین قله در شکل تابع توزیع واقعی وجود داشته باشد که بطور وضوح با یک تابع توزیع نرمال قابل تقریب زدن نخواهد بود.

---

1. Clusters

برای برقراری رابطهٔ $\int_{-\infty}^{\infty} b_j(\underline{o})d\underline{o} = 1$ در رابطهٔ ۱۷ باید شرط $\sum_{m=1}^{M} c_{jm} = 1$ و $c_{jm} \geq 0$ برقرار باشد.

تخمین زدن $\underline{\mu}_{jk}^{\tau+1}$ و $\sigma_{jk}^{\tau+1}$ و $c_{jk}^{\tau+1}$ با فرض موجود بودن این مقادیر در زمان τ و اینکه رابطهٔ

$$b_j(\underline{o}) = \sum_{K=1}^{M} c_{jk}^{\tau+1} N(\underline{o}, \underline{\mu}_{jk}^{\tau+1}, \sigma_{jk}^{\tau+1}) \quad 1 \leq j \leq 6 \qquad (۱۸)$$

تخمینی از توابع چگالی احتمال اعضای مجموعهٔ A_j است، صورت می گیرد. مجموعهٔ A_j به صورت زیر به دست می آید [۱۴].

$$A_j = \{\underline{O}_{tk} : q_{tk} = S_j, 1 \leq k \leq L\} \qquad (۱۹)$$

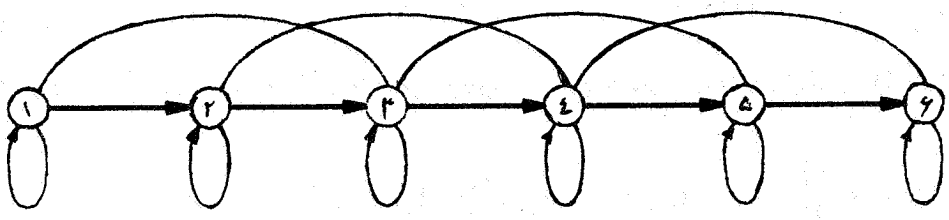
Lکل تکرار کلمهٔ مورد نظر است.

در روابط بالا j نشاندهندهٔ حالت و M تعداد توابع توزیع نرمال به ازای هر حالت است.

روش تخمین مقادیر جدید $\underline{\mu}_{jk}^{\tau+1}$ و $\sigma_{jk}^{\tau+1}$ و $c_{jk}^{\tau+1}$ از روی مجموعه های $A_j, j = 1, \ldots, 6$، مشابه به تخمین در روش k مقدار متوسط در حالت مشاهدات گسسته است. در این روش ابتدا احتمال شرطی هر مشاهده $O_t \in A_j$، نسبت به M خوشه محاسبه می شود، بیشترین احتمال معین می کند که O_t به کدام خوشه مربوط می شود. از این طریق اعضای مجموعهٔ A_j نیز به M دسته (دراینجا $M = 5$) تقسیم می شوند. با استفاده از این تقسیم بندی تخمین پارامترهای مورد نظر به صورت زیر به دست می آید:

$$c_{jk}^{\tau+1} = \frac{\text{تعداد اعضای خوشهٔ k در تکرار } \tau}{\text{کل تعداد اعضای مجموعهٔ } A_j \text{ در تکرار } \tau} \qquad (۲۰)$$

$$\underline{\mu}_{jk}^{\tau+1} = \text{متوسط داده های موجود در خوشهٔ k در حالت j در تکرار } \tau \qquad (۲۱)$$



شکل ۲- مدل مارکف چپ به راست شش حالته

$$\sigma_{jk}^{\tau+1} = \text{واریانس داده های موجود در خوشهٔ k در حالت j در تکرار} \tau \qquad (۲۲)$$

با استفاده از مشخصات سیگنال صحبت، می توان فرضیاتی بر ساختار مدل پنهانی مارکف اضافه کرد که علاوه بر ساده کردن یادگیری، همگرایی به نقطه بهتری را موجب گردد.
مدل مارکف در نظر گرفته شده، یک مدل چپ به راست است (شکل ۲). در نتیجه انتخاب چنین مدلی، به تعریف بردار π نیازی نیست و حالت آغازین همواره حالت یک (۱) است. به علاوه بسیاری از عناصر ماتریس گذار حالت (A) نیز صفر خواهند بود.

$$a_{ij} = 0 \;\; ; \;\; j<i \text{ or } j>i+2 \qquad (۲۳)$$

در پژوهش حاضر از آن جا که هدف، شناسایی اعداد "صفر" تا "ده" است، تعداد حالات، N ، شش در نظر گرفته شده است که برابر بیشترین واجگونه های[1] موجود در بین این اعداد است. چرا که در کلمه "چهار /čâhâr/" در بیان های مختلف پنج یا شش واجگونه وجود دارد. تعداد خوشه ها، M، نیز به طور تجربی پنج خوشه برای هر حالت در نظر گرفته شده است.

---

1 . Allophone

استخراج مشخصه های سیگنال صحبت

به منظور استخراج مشخصه های سیگنال صحبت ابتدا باید محدوده کلمات ادا شده، از محدوده سکوت جدا شود.عمل تفکیک صحبت از سکوت با استفاده از مقدار آستانه انرژی و پاره ای فرضیات انجام می گیرد. به این ترتیب که ابتدا در حالت سکوت حداکثرانرژی محاسبه شده و مقادیرآستانه از روی آن محاسبه می گردد. آنگاه یک ماشین حالت چهار وضعیتی ابتدا و انتهای هر کلمه را تشخیص می دهد ([۱۵ و ۱۶]).

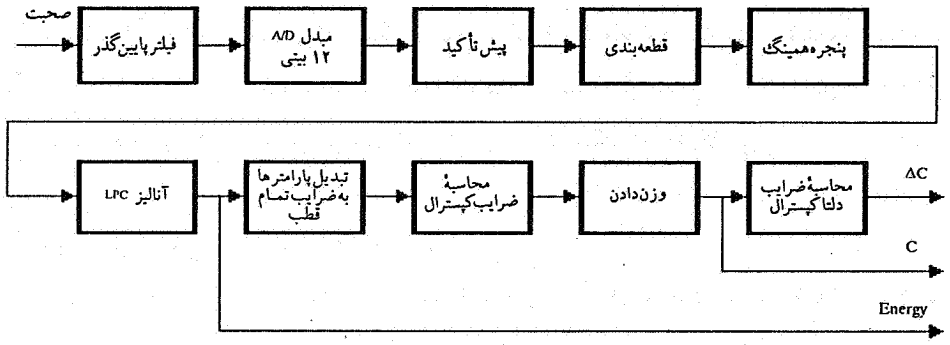
پس از تشخیص ابتدا و انتهای کلمه، مشخصه های سیگنال صحبت در طی زمان ادای کلمه محاسبه می شوند. یک دسته از مشخصات مفید، ضرایب کپسترال سیگنال صحبت هستند ([۱۷ تا۱۹]) انرژی، ضرایب کپسترال و ضرایب دلتا کپسترال پس از انجام پیش پردازش بر روی سیگنال محاسبه می شوند (شکل۳). دلیل استفاده از ضرائب دلتا کپسترال،وجود اطلاعات مربوط به تغییرات بین قطعه های مجاور در این دسته ضرائب است [۱۷].

جزئیات مراحل پیش پردازش و پردازش به قرار زیر است:

* فیلتر آنالوگ پایین گذر
* مبدل A/D ۱۲ بیتی، با نرخ نمونه برداری ۸۰۰۰ هرتز
* پیش تاکید[1] با یک صفر در $H(z) = 1 - 0.9\, z^{-1}$, $z = 0.9$
* تقسیم بندی به قطعه[2]های ۳۷/۵ میلی ثانیه ای (این قطعه ها هر بار ۱۲/۵ میلی ثانیه به جلو برده می شوند. به عبارت دیگر هر قطعه شامل ۳۰۰ نمونه و میزان جابجایی ۱۰۰ نمونه است و هر دو قطعه مجاور در ۲۰۰ نمونه مشترک هستند.)
* پنجره همینگ[3][۲۰]
* آنالیز LPC[4] با مرتبه هشت که از روی آن پارامترهای فیلتر تمام قطب محاسبه می شوند [۲۱].
* محاسبه دوازده پارامتر اول کپسترال از روی پارامترهای فیلتر تمام قطب [۲۲].

---

1. Preemphasis
2. Frame
3. Hamming Window
4. Linear Predictive Coding



شکل ۳ - دیاگرام بلوکی استخراج مشخصه های صحبت

* وزن دادن[1] پارامترهای کپسترال محاسبه شده توسط رابطه زیر [۱۹]:

$$\hat{c}_l(m) = [1+Q \sin(\pi.m/Q)/2].c_l(m) \quad ; \quad 1 \le m \le Q \quad Q = 12 \qquad (۲۴)$$

(در این رابطه $c_l(m)$ ضریب کپسترال m در زمان l است.)

* محاسبه ضرایب دلتا کپسترال[2] که توسط رگرسیون خطی مرتبه دو از روی پارامترهای کپسترال قطعه های مجاور آن به صورت زیر محاسبه می شوند [۱۷]:

$$\Delta\hat{c}_l(m) = [\sum_{k=-2}^{2} k\, \hat{c}_{l-k}(m)] \times 0.375 \quad ; \quad 1 \le m \le Q \quad Q = 12 \qquad (۲۵)$$

(ضریب ۰/۳۷۵ برای متعادل نمودن واریانس این ضرایب با ضرایب کپسترال بکار می رود.)

بردارمشاهده‌مربوط‌به‌هرقطعه شامل ضرایب‌کپسترال، ضرایب دلتاکپسترال ولگاریتم‌انرژی است.

---

1. Weighting
2. Delta Cepstral

تحقق مدل پنهانی مارکف

همانطورکه قبلاً توضیح داده شد، یادگیری به "روش k مقدار متوسط برای هر قسمت" که یک روش جستجوی محلی[1] است انجام می گیرد. با توجه به این موضوع که مسأله دارای چندین ماگزیمم محلی است، الگوریتم به ماگزیممی همگرا می شود که شرایط اولیه در بستر جذب آن قرار داشته اند. در نتیجه انتخاب شرایط اولیه درکارآیی نهایی سیستم اثر زیادی خواهد داشت. البته مسلم است که با توجه به روش بکار رفته هیچگاه نمی توان مطمئن بود که نقطه بهتری برای همگرا شدن وجود نداشته است.از میان پارامترهای مختلف مدل λ اهمیت بردارهای متوسط توابع چگالی نرمال، μ_{jm}، بیشتر است. در مرحله اول مقادیر اولیه پارامترهای μ_{jm} به صورت تصادفی انتخاب شدند. در این حال به علت وجود خطای ته ریز[2] اصولاً الگوریتم با مشکل روبرو شد و به جوابی نرسید. دلیل این امر مکان نامناسب نقاط آغازین است. توجه کنید μ_{jm} یک بردار ۲۵ بعدی است. انتخاب تصادفی مقدار آغازین μ_{jm} در واقع انتخاب تصادفی یک نقطه در فضای ۲۵ بعدی است. لذا امکان آن که هیچ یک از پنج بردار متوسط مربوط به پنج خوشه در نزدیکی بردارهای مشاهده نباشند زیاد است. از طرف دیگر تابع چگالی احتمال مربوط به هر خوشه بصورت نرمال است. به دلیل وجود فرم نمائی در تابع چگالی نرمال با زیاد شدن فاصله بردار متوسط از بردار مشاهدات احتمال بردار مشاهده به شدت کاهش می یابد و در عمل خطای ته ریز پیش می آید. برای اجتناب از این امر باید در نزدیکی هر یک از بردارهای مشاهده حداقل یکی از بردارهای متوسط قرار داشته باشد.

برای اینکه جواب نهایی روش محقق شده ماگزیمم نسبتاً بزرگی باشد و همچنین برای اینکه مشکل ته ریز باعث بروز خطا نشود، روشی برای تعیین مقادیر اولیه μ ابداع شده است که در ادامه به توضیح آن خواهیم پرداخت ([۱۵ و ۲۳]).

ایدهٔ اصلی این روش این است که هر حالت را می توان به یک واج (واجگونه) درون کلمه مورد نظر نسبت داد. این واجها اغلب پشت سر هم بیان می شوند (به غیر از مواردی که واجها خورده می شوند و بیان نمی شوند). بردارهای مشاهده مربوط به هر واج در یک محدوده کوچک در فضای

---

1. Local Search

2 . Underflow

۲۵بعدی‌قراردارند. اگر در این محدوده های نسبتاً کوچک بردار متوسط محاسبه شود، شانس نسبت داده شدن‌بردارهای مشاهده هرواج به‌ناحیه مربوطه زیاد شده وعلاوه‌براینکه‌این‌موضوع به درستی عمل‌تشخیص‌کمک‌می‌کند، جلوی‌کوچک شدن بیش‌ازاندازهٔ‌احتمالهای محاسبه شده را نیز می گیرد.

با تقسیم بندی کلمات ادا شده به تکه های مجزا، که هر تکه یک واج (واجگونه) را شامل می شود، می توان پایگاه داده ها را به N دسته (N تعداد حالات است) تقسیم‌کرد که هر دسته به یک حالت مربوط می شود. این تقسیم بندی می تواند خیلی دقیق و یا با تقریب (با در نظر گرفتن N قسمت برابر در هر کلمه) صورت پذیرد.

برای محاسبه مقادیر μاز روی این N دسته در حالت مشاهدهٔ پیوسته کافی است با استفاده از یک روش چندی کردن برداری [۲۴] با M بردار نمونه، داده های هر دسته را به M خوشه تقسیم کرده و متوسط هر خوشه را به عنوان مقادیر اولیه μانتخاب نمود. با این روش این تضمین بوجود می آید که بردارهای متوسط $k=1,...,M$ و μ_{jk}; $j=1,...,N$ در محدودهٔ بردارهای مشاهده قرار بگیرند. همچنین به‌کمک این روش مرحلهٔ یادگیری در مدل پنهانی مارکف به صورت کاملا" خودکار انجام می گیردوعوض کردن مجموعهٔ کلمات و مجموعهٔ صداها وقتی از شخص نخواهد گرفت. جهت انجام چندی کردن برداری از روش LBG[1] استفاده شد.

از مسائل عملی که در تحقق مدل مورد نظر پیش می آید، خطای ته ریز است. برای مثال در حل مسأله اول به روش بازگشتی مقادیر $p(\underline{o}_1,...\underline{o}_t, q_t=S_i|\lambda)$ با زیاد شدن زمان بسیار کوچک می شوند، به طوری که از حدود دقت کامپیوتر خارج می گردند. جهت اجتناب از این مشکل ، در هر مرحله $p(\underline{o}_1,...\underline{o}_t, q_t=S_i|\lambda)$ را در عدد مناسبی ضرب می کنیم به طوری که حاصل در محدوده دقت کامپیوتر باقی بماند. اگر عدد نرمالیزاسیون (c_t) به صورت زیر انتخاب شود:

$$c_t = \frac{1}{\sum_{i=1}^{N} P(\underline{o}_1 ... \underline{o}_t, q_t = S_i|\lambda)} \quad ; t = 1, ... , T \qquad (۲۶)$$

---

1 . Linde , Buzo , Gray

$p(O|\lambda)$ نیز نهایتاً توسط رابطهٔ زیر بدست می آید:

$$P(O|\lambda) = \frac{1}{\prod_{t=1}^{T} c_t} \qquad (۲۷)$$

ویا

$$\text{.Log } P(O|\lambda) = -\sum_{t=1}^{T} \text{Log } c_t \qquad (۲۸)$$

مشکل ته ریز در محاسبه تابع چگالی احتمال پیوسته ($b_j(o_t)$) نیز وجود دارد. برای حل این مشکل ابتدا در هر حالت احتمال خوشه ای که حداکثر احتمال را دارد در نظر گرفته می شود. سپس حداقل این مقادیر در بین حالات مختلف، بدست آمده و تمام احتمالات محاسبه شده به این مقدار نرمالیزه می شود. در این صورت می توان امیدوار بود که مقادیر به دست آمده در محدوده دقت کامپیوتر باقی بمانند . به سادگی می توان نشان داد که با این نرمالیزاسیون روش حل مساله دوم (یافتن مسیر حالت بهینه) تغییری نمی کند. حل مسأله اول (محاسبه $p(o|\lambda)$) نیز با انجام عمل عکس نرمالیزاسیون[1] تصحیح می شود.

روش آزمایش

جهت تخمین پارامترهای مدلها و آزمایش سیستم محقق شده از سه مجموعه سیگنال صحبت به نامهای ZX، YW، UV استفاده شد. هرکدام از این مجموعه ها شامل صدای ۲۵ زن و ۲۵ مرد است که هر یک از افراد اعداد صفر تا ده را دوبار بیان کرده اند. ضبط صدا در محیط آزمایشگاه و توسط میکروفون و تقویت کننده صورت گرفته است. عمل ضبط به صورت ۱۲ بیتی بر روی کامپیتر مجهز به A/D صورت پذیرفته است. از دو مجموعهٔ ZX و YW برای تخمین پارامترهای مدل و از مجموعهٔ UV برای آزمون کارآیی سیستم استفاده شد.نتایج این آزمایش در جدول ۱ آمده است. همان گونه که در این جدول مشاهده می شود، میزان خطای کل سیستم ۳/۰۰٪ برآورد شده است.

---

1 . Denormalization

لازم به تذکر است که در آزمایش صدای UV نیز عمل شناسایی و ضبط صدا در دو مرحلهٔ مختلف صورت گرفته است.

استفاده از مدل پنهانی مارکف به صورت چپ به راست با این فرض همراه است که هر یک از حالات مدل با یکی از ساختارهای صحبت موجود در کلمه (واج) متناظر است. البته این تناظر در کلیه موارد به طور دقیق به چشم نمی خورد. به هر حال شکل ۴ تقسیم بندی یک سیگنال نمونه به حالتهای تشکیل دهنده اش توسط این مدل را نشان می دهد. همان طوری که در شکل ۴ ملاحظه می شود هر یک از حالتها به یکی از واج ها (یا واجگونه ها) نسبت پیدا می کند.

یکی‌ازمسائل مهم درمدل پنهانی مارکف پیوسته تخمین‌توابع چگالی احتمال مربوط به هر حالت است. همان طوری که گفته شد برای هر حالت تابع چگالی احتمال به صورت مجموع پنج تابع چگالی نرمال در نظر گرفته شده است.استفاده از چندتابع نرمال می‌تواند به تخمین توابع پیچیده‌تر کمک کند.در شکل ۵ نحوه‌تخمین چند نمودارستونی[1] توسط این‌روش نشان داده شده است. برای نمونه درشکل ۴-۵ وجود دوماگزیمم درنمودار ستونی نیزبه خوبی‌تخمین زده‌شده است.

نتیجه‌ای که درپایان ازآزمایش مدل محقق شده‌روی مجموعه های S1 (داده‌های یادگیری) و S2 (داده های‌آزمایش) گرفته شد، درجدول ۱ آمده است:

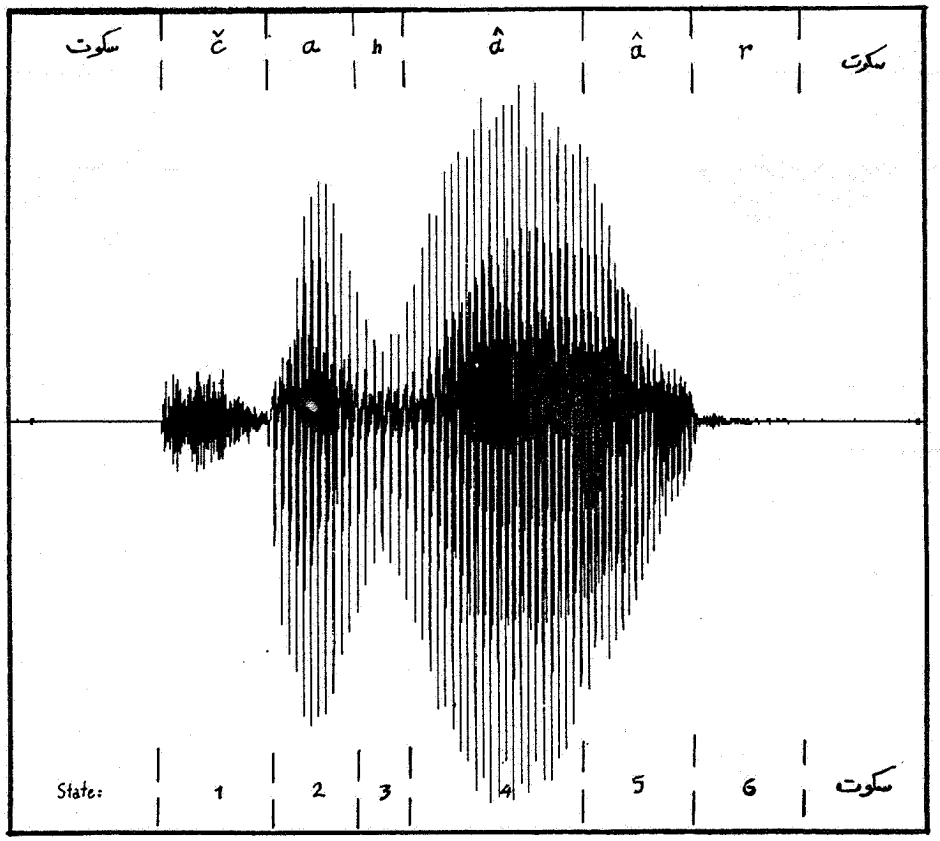
همان طورکه مشاهده می‌شودبیشترین خطامربوط به کلمات ("صفر/sefr/" و "سه /se/") است و این به خاطر شباهت این دو کلمه به یکدیگر است. خصوصاً که بسیاری از افراد واج آخر کلمه "صفر" را به درستی بیان نمی کنند. همین طور دو کلمه ("دو/do/" و"نه/noh/") نیز گاهی با هم اشتباه می شوندوکلمات ("هفت/haft/" با "هشت/hašt/") نیزدیگر موارد عمدهٔ خطا راتشکیل می دهند.

نتیجه گیری

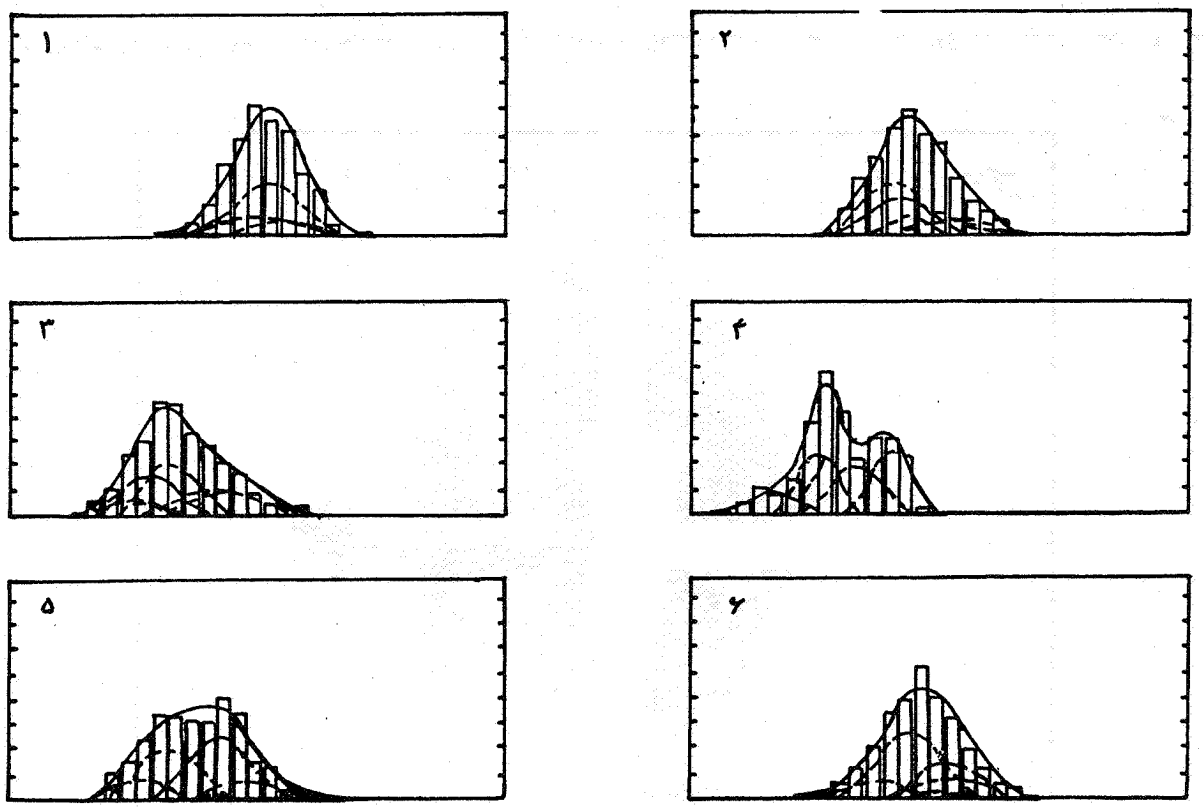
در این مقاله تحقق یک سیستم درک صحبت به روش مدل پنهانی مارکف ارائه شد. برای انتخاب مقادیر اولیه از چندی کردن برداری داده های یادگیری تقسیم شده بین حالتها در هر کلمه استفاده

---

1 . Histogram



شکل ۴ - نحوهٔ تقسیم بندی کلام به حالت های مختلف (متناظر با واجها) توسط HMM ،نمونهٔ مورد استفاده کلمهٔ چهار می باشد.



شکل ۵ - هیستوگرام وتابع چگالی تخمینی برای یک متغیر (C5) درکلمهٔ "چهار"، درحالتهای ششگانه. هیستوگرام نمایش دهنده توزیع تجربی، منحنی پررنگ تخمین تابع چگالی احتمال ومنحنی های کم رنگ حاصل ضرب تابع چگالی احتمال هرخوشه دراحتمال آن خوشه می باشند.

جدول ۱- میزان خطا برای کلمات و مجموعه های مختلف

مجموعه کلمه	ZX آموزش			YW آموزش			UV آزمون		
	زن	مرد	زن‌ومرد	زن	مرد	زن‌ومرد	زن	مرد	زن‌ومرد
۰	۳/۵۰	۶/۵۰	۹/۱۰۰	۴/۵۰	۲/۵۰	۶/۱۰۰	۰	۱/۵۰	۱/۱۰۰
۱	۰	۰	۰	۰	۰	۰	۰	۰	۰
۲	۰	۰	۰	۰	۰	۰	۱/۵۰	۶/۵۰	۷/۱۰۰
۳	۲/۵۰	۰	۲/۱۰۰	۲/۵۰	۱/۵۰	۳/۱۰۰	۸/۵۰	۴/۵۰	۱۲/۱۰۰
۴	۰	۰	۰	۰	۰	۰	۰	۰	۰
۵	۰	۰	۰	۰	۰	۰	۰	۰	۰
۶	۰	۰	۰	۰	۰	۰	۱/۵۰	۰	۱/۱۰۰
۷	۱/۵۰	۰	۱/۱۰۰	۰	۱/۵۰	۱/۱۰۰	۰	۳/۵۰	۳/۱۰۰
۸	۰	۰	۰	۰	۰	۰	۰	۰	۰
۹	۰	۰	۰	۲/۵۰	۲/۵۰	۴/۱۰۰	۰	۵/۵۰	۵/۱۰۰
۱۰	۰	۰	۰	۰	۰	۰	۲/۵۰	۲/۵۰	۴/۱۰۰
۰-۱۰	٪۱/۰۹	٪۱/۰۹	٪۱/۰۹	٪۱/۴۵	٪۱/۰۹	٪۱/۲۷	٪۲/۱۸	٪۳/۸۱	٪۳/۰

مقادیر اولیه از چندی کردن برداری داده های یادگیری تقسیم شده بین حالتها در هر کلمه استفاده گردید. با آزمایش سیستم محقق شده خطای سیستم ۳/۰۰ درصد برآورد شد. مقایسه کارآیی سیستم با سایر سیستمهای درک صحبت به دلیل عدم دسترسی به اطلاعات سیستم های مشابه برای زبان فارسی میسر نشد. مقایسه این سیستم با سیستم هایی که برای زبانهای بیگانه طراحی شده است نیز خالی از اشکال نیست. با اینحال اگر با تسامح بخواهیم مقایسه ای انجام بدهیم، دیده می شود که در سیستمی که مشابه سیستم حاضر در زبان انگلیسی می باشد و برای تشخیص اعداد "صفر" تا "نه" به کار می رود [۱۰] و تنها مرحلهٔ مقدار دهی اولیهٔ پارامترهای آن با سیستم حاضر متفاوت است (در واقع در مورد روش مقدار دهی اولیهٔ آن سیستم اطلاعی در دست نیست)، مشاهده می شود که

کارآیی سیستم فارسی حدود یک درصد کمتر است، همان طور که مشاهده می شود در مجموعه کلمات فارسی زوج های ("دو"، "نه") و ("صفر"، "سه") همچنین‌کلمات ("هفت"، "هشت"، "پنج") شباهت زیادی دارند. چنین شباهتی برای ارقام "صفر" تا "نه" به زبان انگلیسی فقط در زوج های (Nine ، Five) و (Six ، Three) دیده می شود که درجه شباهت نیز از موارد مشابه در زبان فارسی کمتر است. چنانچه اگر یکی از موارد شباهت در فارسی مثلا" "صفر" به کلی حذف شود، میزان خطا به ۱/۸ درصد می رسد که در حدود درصد خطای سیستم انگلیسی (۱/۳٪ و ۱/۸٪ برای دو مجموعه آزمایشی مختلف) خواهد بود.

قدردانی

نویسندگان مقاله لازم می دانند که از مسئولین محترم مرکز تحقیقات الکترونیک دانشگاه صنعتی شریف بویژه آقایان دکتر محمود تبیانی و مهندس همایون برهانی و همچنین از معاونت پژوهشی دانشگاه صنعتی شریف و مسئولین محترم امور پژوهشی دانشگاه تهران به جهت در اختیار گذاردن شرایط و امکانات انجام این پروژه تشکر و قدردانی به عمل آورند.

مراجع

۱- ی . ثمره، *آواشناسی زبان فارسی، آواها و ساخت آوائی هجا*، مرکز نشر دانشگاهی، ۱۳۶۴.

2. Levinson, E. and Roe, D.B. ,"A Perspective on Speech Recognition", *IEEE Comm. Mag.*,Vol. 28, No. 1, Jan. 1990.

3. Slutsker, G.," Non - Linear Method of Analysis of Speech Signal ", *Trudy N.I.I.R.*,1968.

4. Venlichko, V. and Zagoruyko, N., "Automatic Recognition of 200 words", *Int. J. Man - Machine Studies,* 2, pp. 223-234, 1970.

5. Vintsyuk, T.K., "Speech Recognition by Dynamic Programming", *Kybernetika,* 4, 1, pp. 81-88, 1968.

۶. و. طباطبایی، ب. عظیمی سجادی، ب.ظهیر اعظمی و ک.لوکس، بررسی و مقایسهٔ دو روش تشخیص کلمات منفرد فارسی، *کارنامهٔ اولین کنفرانس آموزش، پژوهش و کاربرد کامپیوتر در ایران*، تهران، بهمن ۱۳۷۱.

7. Waibel, A. et al, "Phoneme Recognition Using Time Delay Neural Networks", *IEEE Trans. on ASSP*, Vol. ASSP-37, No. 3, March 1989.

8. Avarbuch, A. et al, "An IBM - PC Based Large - Vocabulary Isolated Utterance Speech Recognizer", in *Proceedings IEEE ICASSP-86*, Tokyo, Japan, April 1986.

9. Lee, K.F., *Large - Vocabulary Speaker Independent Continuous Speech Recognition:The SPHINX System,* Ph.D. Dessertation, Computer Science Department, Carnegie Mellon University, 1988.

10. Rabiner, L.R., "A Tutorial on Hidden Markov Models and Selected

Applications in Speech Recognition" ,*IEEE Proc.,* Vol. 77, No. 2, Feb. 1989.

11. Baum, L.E. and Petrie, T., "Statistical Inference for Probabilistic Functions of Finite State Markov Chains", *Ann. Math. Stat.,* Vol.37,1966.
12. Baum, L.E., Petrie, T. , Soules, G. and Weiss, N., "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains", *Ann. Math. Stat.* ,Vol. 41, No. 1, 1970.
13. Forney, G.D ,"The Viterbi Algorithm" , *Proc. of the IEEE,* Vol. 61, No.3 ,Mar. 1973.
14. Juang, B.H. and Rabiner, L.R., "The Segmented K - Means Algorithm for Estimating Parameters of Hidden Markov Models", *IEEE Trans. on ASSP,* Vol. ASSP-38, No. 9, Sep. 1990.

۱۵. و. طباطبایی ، *استخراج مشخصه ها و پیش پردازش سیگنال صحبت به منظور درک کلمات*، پایان نامه کارشناسی ارشد، گروه برق دانشکده فنی دانشگاه تهران، ۱۳۷۱.

16. Rabiner, L.R.and Sambur, M.R., "An Algorithm for Determining the End Point of Isolated Utterances", *B.S.T.J.*, Vol 54, No. 2, Feb. 1975.
17. Furui, S., "Speaker Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum", *IEEE Trans. on ASSP,* Vol. ASSP-34, No.1,Feb. 1986.
18. Furui, S., *Digital Speech Processing, Synthesis and Recognition,* Copyright 1989 by Marcel Decker, Inc.
19. Juang, B.H., Rabiner, L.R. and Wilpon, J.G.,"On the Use of Bandpass Liftering in Speech Recognition", *IEEE Trans. on ASSP,* Vol. ASSP-35, No. 7, July 1987.
20. Oppenheim, A.V. and Schafer, R.W., *Discrete - Time Signal Processing* , Prentice Hall, INC. Englewood Cliffs, NJ.,1989.

21. Parsons, T.W., *Voice and Speech Processing,* McGraw-Hill, 1986.

22. Viswanathan, R.and Makhoul, J.,"Quantization Properties of Transmission Parameters in Linear Predictive Systems", *IEEE Trans.on ASSP,* Vol. ASSP-23, June 1975.

۲۳. ب. عظیمی سجادی، *درک صحبت مستقل از گوینده برای کلمات منفرد فارسی با تعداد محدود*، پایان نامه کارشناسی ارشد، گروه برق دانشکده فنی دانشگاه تهران، ۱۳۷۱.

24. Makhoul, J.,Roucos S. and Gish, H.,"Vector Quantization in Speech Coding", *IEEE Proc.*, Vol. 73, No. 11, Nov. 1985.