سومین کنفرانس فناوری اطلاعات و دانش
۶ تا ۸ آذرماه ۱۳۸۶

بهبود بازشناسی گفتار توسط شبکه‌های عصبی دربرگیرندهٔ الگوهای زمانی

یاسر شکفته و فرشاد الماس‌گنج

دانشکده مهندسی پزشکی دانشگاه صنعتی امیرکبیر

E-mail: yshekofteh@bme.aut.ac.ir

چکیده – الگوهای زمانی یا *"Temporal Patterns"* از جمله ویژگی‌های موثر در بهبود بازشناسی گفتار می‌باشند. از آنجا که شبکه‌های عصبی *MLP*، جزو طبقه‌بندی‌کننده‌های قدرتمند استاتیک می‌باشند، اگر بخواهیم از این نوع ساختار شبکه‌ای در بازشناسی گفتار استفاده کنیم، بهتر است به هر صورت ممکن، شبکه را وادار سازیم، الگوهای زمانی را از دنبالهٔ ویژگی‌های ورودی یادگیری نموده و آن را در فرآیند دسته‌بندی قاب‌های ورودی گفتار دخالت دهد. در این مقاله، با بررسی و آزمایش این شبکه و مدل‌های دیگر شبکه‌ایِ که به صورتی قوی‌تر به پردازش و یادگیری ویژگی‌های مبتنی بر الگوهای زمانی می‌پردازند، سعی در افزایش کارایی سیستم‌های بازشناس گفتار پیوسته و مستقل از گوینده خواهیم داشت. همچنین به تحلیل نوع عملکرد این روش‌ها و مدل‌های مبتنی بر آن‌ها خواهیم پرداخت. با در نظر گرفتن نتایج یک مدل پایه بازشناس *"TDNN"* که بازدهی حدود ۸۳٫۷٪ بر روی دادهٔ تست انتخابی ما از دادگان فارس‌دات داشته است، با استفاده از یک مدل ترکیبی ارائه شده در تحقیق حاضر، میزان بازشناسی ۸۸٫۱٪ در دادگان آزمون تمیز حاصل شده است. در حالی که مدل ترکیبی پیشنهادی، نسبت به مدل پایه بازشناس *"TDNN"* در شرایط نویز شدید، بین ۷ تا ۲۰ درصد مقاوم‌تر از شبکهٔ *"TDNN"* عمل می‌نماید. همچنین مدل ترکیبی دیگری معرفی می‌گردد که به علت استفادهٔ متمرکزتر از ویژگی الگوهای زمانی، مقاوم‌تر از مدل ترکیبی اولیه، در شرایط نویز شدید عمل می‌نماید.

کلید واژه‌ها: استخراج ویژگی، الگوهای زمانی، بازشناسی گفتار، مقاوم‌سازی ویژگی‌ها، شبکه عصبی.

۱- مقدمه

در طی دهه‌های گذشته محققان حوزه پردازش و بازشناسی گفتار، تلاش‌های زیادی برای ساخت سیستم‌های بازشناسی گفتار مقاوم در برابر تنوعات (مانند تنوعات گوینده، نویز محیط، کانال انتقال، پژواک صدا و ...) انجام داده‌اند؛ اما با این وجود، نتایج بازشناسی هنوز به مطلوب واقعی خود نرسیده است [۱].

در سال ۱۹۹۴ میلادی، آلن (Alen) با الهام از آزمایشات شنیداری فلچر (Feltcher) نشان داد که بازشناسی آواها توسط انسان در هجاهای بدون معنی، نرخ خطایی در حدود ۱٫۵٪ دارد [2]. در صورتی که بازشناسی همین آواها با استفاده از مدل‌های بازشناس خودکار گفتار، به مراتب ضعیف‌تر می‌باشد. نکته قابل توجه در آزمایشات آلن عدم توانایی انسان در استفاده از اطلاعات معنایی کلمه، در جهت بازشناسی بود؛ چون هجاهای مورد استفاده بدون معنا بودند. نتیجه این آزمایش موید این مطلب بود که انسان در بازشناسی گفتار پردازش شده در مغز، از اطلاعات واحدهای بزرگتر زمانی (حداقل ۲۰۰ میلی ثانیه) برای بازشناسی آوا استفاده می‌کند [3]. این در حالی است که در روش‌های متداول بازشناسی گفتار، فرایند استخراج ویژگی که بیشتر مبتنی بر تبدیل فوریه زمان کوتاه (مانند ضرایب کپسترال در مقیاس مِل (MFCC)، لگاریتم هنینگ باندهای بحرانی (LHCB) و پارامترهای پیشگویی خطی ادراکی (PLP)) است، از اطلاعات قاب‌های زمانی کوتاه مدت (بین ۱۰ تا ۲۵ میلی ثانیه) سیگنال گفتار به دست می‌آید [۱]. البته با محاسبه پارامترهای دلتا و یا دلتادلتا بازنمایی، که به نوعی مشتقات زمانی ویژگی‌های استخراج شده محسوب می‌شوند،