NHẬN DẠNG TIẾNG NÓI BỀN VỮNG SỬ DỤNG KỸ THUẬT THỪA SỐ HÓA MA TRẬN KHÔNG ÂM KẾT HỢP VỚI KỸ THUẬT VỀ ĐỘ KHÔNG ĐẢM BẢO CỦA CÁC ĐẶC TRƯNG ÂM HỌC

Nguyễn Hữu Bình[1], Phạm Thị Ngọc Yến[1,2], Nguyễn Quốc Cường[1,2*]

Tóm tắt: *Trong hệ thống nhận dạng tiếng nói kỹ thuật thừa số hóa ma trận không âm có thể được sử dụng trong khâu tiền xử lý để loại bỏ nhiễu, nâng cao chất lượng tiếng nói cần nhận dạng và do đó có thể tăng chất lượng của hệ thống nhận dạng tiếng nói trong môi trường nhiễu. Tuy nhiên, tín hiệu sau khi nâng cao thường vẫn còn chứa một phần nhiễu. Thông tin sai khác giữa tín hiệu nâng cao và tín hiệu sạch, hay gọi là độ không đảm bảo, có thể là thông tin hữu ích cho quá trình giải mã của hệ thống nhận dạng tiếng nói. Trong bài báo này, chúng tôi trình bày một phương pháp nâng cao chất lượng hệ thống nhận dạng tiếng nói dựa trên kỹ thuật thừa số hóa ma trận không âm kết hợp với kỹ thuật giải mã sử dụng thông tin về độ không đảm bảo của vec-tơ đặc trưng. Chúng tôi đã đánh giá phương pháp kết hợp này trong hệ thống nhận dạng tiếng nói tiếng Việt. Các kết quả cho thấy phương pháp kết hợp đã nâng cao độ chính xác của hệ thống nhận dạng hơn so với việc chỉ sử dụng kỹ thuật thừa số hóa ma trận không âm trong hệ thống nhận dạng tiếng nói.*

Từ khóa: Nhận dạng tiếng nói, Thừa số hóa ma trận không âm, Ước lượng độ không đảm bảo.

1. ĐẶT VẤN ĐỀ

Nhận dạng tiếng nói tự động (ASR: Automatic Speech Recognition) là lĩnh vực thu hút sự quan tâm của nhiều nhà nghiên cứu trong các ứng dụng về tương tác người máy và dịch tiếng nói tự động. Kết quả nhận dạng trong môi trường không nhiễu đạt tỷ lệ khá cao, hơn 90%. Tuy nhiên, trong môi trường ứng dụng thực tế có nhiễu thì chất lượng nhận dạng giảm đáng kể do có sự sai khác giữa cơ sở dữ liệu dùng để huấn luyện mô hình nhận dạng, thường được thu âm trong môi trường không nhiễu, và dữ liệu cần nhận dạng, thường có nhiễu môi trường. Để nâng cao tính bền vững của hệ thống ASR trong môi trường nhiễu, một thuật toán nâng cao chất lượng tiếng nói thường được tích hợp vào trong hệ thống ASR với mục đích loại bỏ nhiễu để tín hiệu tiếng nói đưa vào nhận dạng giống với tín hiệu tiếng nói dùng để huấn luyện mô hình. Các thuật toán nâng cao chất lượng tiếng nói có thể phân thành hai nhóm: nhóm các thuật toán sử dụng một microphone và nhóm các thuật toán sử dụng nhiều microhone [1]. Nhóm các thuật toán sử dụng nhiều microphone thường cho chất lượng tốt hơn trong môi trường nhiễu do có nhiều thông tin về tiếng nói cần nhận dạng cũng như thông tin về nhiễu cần loại bỏ. Tuy nhiên, việc triển khai hệ thống nhiều microphone thường khó khăn trong các ứng dụng thực tế do vấn đề về chi phí cũng như kích thước của mảng microphone. Trong khi đó, nhóm thuật toán sử dụng một microphone cho phép triển khai đơn giản và chi phí thấp trong tất cả các hệ thống nhận dạng tiếng nói. Các phương pháp sử dụng một microphone có thể kể đến là: phương pháp trừ phổ [2], phương pháp lọc Wiener [3]. Điểm chung là các phương pháp này thường yêu cầu tín hiệu nhiễu cần loại bỏ là tín hiệu ngẫu nhiên dừng [4]. Điều kiện này trong thực tế thường khó đảm bảo.

Một trong các hướng nghiên cứu nâng cao chất lượng tiếng nói sử dụng một microphone đang được các nhà nghiên cứu quan tâm là kỹ thuật tách nguồn sử dụng phương pháp thừa số hóa ma trận không âm (NMF: Nonnegative Matrix Factorization). Phương pháp NMF sau khi được Lee và Seung đề xuất trong [5] đã được áp dụng rộng rãi cho nhiều lớp bài toán ứng dụng: phân tích văn bản, xử lý ảnh, xử lý âm thanh. Trong nâng cao chất lượng tiếng nói ưu điểm của NMF so với các phương pháp trừ phổ hoặc phương pháp lọc Wiener là có thể áp dụng cho trường hợp nhiễu không dừng [4], [6].

Phương pháp NMF có thể được sử dụng như là bước tiền xử lý tín hiệu tiếng nói có nhiễu tại đầu vào của hệ thống ASR trước khi đưa vào giải mã. Tuy nhiên, tín hiệu tiếng nói ước lượng vẫn có sự sai khác so với tín hiệu tiếng nói sạch. Để nâng cao chất lượng nhận dạng một kỹ thuật giải mã có thể được áp dụng trong ASR đó là kỹ thuật giải mã dựa trên thông tin về độ không đảm bảo (UD: Uncertainty decoding) hay phương sai của tín hiệu tiếng nói [7-8]. Trong hệ thống ASR sử dụng HMM phương sai hay độ không đảm bảo được sử dụng để thích nghi mô hình âm học tại mỗi khung thời gian trong quá trình giải mã của ASR.

Đóng góp chính của bài báo này là giới thiệu một hệ thống ASR trong đó NMF được áp dụng để loại bỏ nhiễu trong tiếng nói cần nhận dạng. Độ không đảm bảo của tiếng nói được ước lượng và được lan truyền đến miền đặc trưng MFCC (Mel Frequency Cepstral Coeficient). Sau đó, một bộ giải mã có sử dụng thông tin về độ không đảm bảo dựa trên mô hình Markov ẩn HMM (Hidden Markov Model) được sử dụng. Hệ thống ASR được đánh giá trên một cơ sở dữ liệu tiếng nói mô phỏng với các tỷ số tín hiệu trên nhiễu khác nhau.

Bài báo này được tổ chức như sau: Mục 2 sẽ trình bày cơ sở thuật toán NMF. Ước lượng độ không đảm bảo và thủ tục lan truyền độ không đảm bảo từ miền phổ sang miền đặc trưng MFCC sẽ được mô tả trong mục 3. Kết quả thí nghiệm về nhận dạng của hệ thống ASR được trình bày ở mục 4 và kết luận ở mục 5.

2. NÂNG CAO CHẤT LƯỢNG TIẾNG NÓI SỬ DỤNG NMF

2.1. Thuật toán thừa số hóa ma trận không âm NMF

Cho một ma trận không âm **V**, cần tìm các ma trận không âm **W** và **H** sao cho:

$$\mathbf{V} \approx \mathbf{W}\,\mathbf{H} \tag{1}$$

Để tìm được một xấp xỉ tốt cho (1), trước tiên cần định nghĩa hàm mục tiêu $D(\mathbf{V} \,||\, \mathbf{WH})$ thể hiện sự sai khác giữa **V** và **WH**. Ma trận **W** và **H** được lựa chọn sao cho hàm mục tiêu là cực tiểu

$$\min_{\mathbf{W}\geq 0,\mathbf{H}\geq 0} D(\mathbf{V} \,||\, \mathbf{WH}) \tag{2}$$

Các hàm mục tiêu thường được sử dụng là sai khác Ơ-clit[5], sai khác Kullback-Leibler[5] hay sai khác Itakuar-Saito[9].

Do hàm mục tiêu (2) thường không phải là hàm lồi cho cả hai biến **W** và **H**, do đó việc tìm cực tiểu toàn cục là không khả thi. Thay vào đó có nhiều kỹ thuật tối ưu hóa bằng phương pháp số để tìm **W** và **H** tại cực tiểu địa phương [10].

2.2. Nâng cao chất lượng tiếng nói sử dụng NMF

Cho tín hiệu tiếng nói có nhiễu v , hay còn gọi là tín hiệu trộn, gồm tín hiệu tiếng

nói sạch *x* và nhiễu *y*. Ký hiệu $\mathbf{V} \in \mathbb{C}^{F\times N}$, $\mathbf{X} \in \mathbb{C}^{F\times N}$ và $\mathbf{Y} \in \mathbb{C}^{F\times N}$ là ma trận các hệ số biến đổi Fourier thời gian ngắn (STFT) của tín hiệu trộn, tín hiệu tiếng nói và nhiễu, với *F* là số điểm tần số và *N* là số khung cửa sổ thời gian.

Giả thiết phổ biên độ của tín hiệu trộn:

$$|\mathbf{V}| \approx |\mathbf{X}| + |\mathbf{Y}| \tag{3}$$

Mục tiêu bài toán nâng cao chất lượng tiếng nói là khôi phục $\mathbf{X}$ từ $\mathbf{V}$.

Sử dụng NMF, $|\mathbf{V}| \in \mathbb{R}_{+}^{F\times N}$ có thể được phân tách thành hai ma trận không âm $\mathbf{W} \in \mathbb{R}_{+}^{F\times K}$ và $\mathbf{H} \in \mathbb{R}_{+}^{K\times N}$. Trong đó, $\mathbf{W}$ được xem là các đặc trưng phổ của $|\mathbf{V}|$ với *K* là số vec-tơ phổ cơ sở. Ma trận $\mathbf{H}$ thể hiện thời điểm kích hoạt các đặc trưng phổ cơ sở trong tín hiệu. Kỹ thuật NMF cho nâng cao chất lượng tiếng nói thường dựa trên NMF có giám sát [6], [11], bao gồm hai pha: pha huấn luyện và pha đánh giá. Trong pha huấn luyện, từ phổ biên độ tín hiệu tiếng nói $|\mathbf{V}_X|$ và nhiễu $|\mathbf{V}_Y|$ cho trước, tiến hành ước lượng ma trận $\mathbf{W}_X$ và $\mathbf{W}_Y$ dựa trên việc tối ưu hóa theo tiêu chuẩn (2). Ma trận $\mathbf{W}$ của $|\mathbf{V}|$ nhận được bởi:

$$\mathbf{W} = [\mathbf{W}_X \ \mathbf{W}_Y] \tag{4}$$

Trong pha đánh giá (quá trình nâng cao chất lượng tiếng nói), dựa trên $\mathbf{W}$ nhận được từ pha huấn luyện, phổ biên độ $|\mathbf{V}|$ được phân tích theo NMF thành:

$$|\mathbf{V}| = [\mathbf{W}_X \ \mathbf{W}_Y] \ \mathbf{H} \tag{5}$$

Với ma trận $\mathbf{H}$ sau đó cũng được phân tách thành hai khối:

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_X \\ \mathbf{H}_Y \end{bmatrix} \tag{6}$$

Trong đó, $\mathbf{H}_X$ là ma trận kích hoạt ứng với tiếng nói $|\mathbf{X}|$ còn $\mathbf{H}_Y$ ứng với nhiễu $|\mathbf{Y}|$.

Ước lượng của phổ tín hiệu tiếng nói không nhiễu $|\hat{\mathbf{X}}|$ được tính theo dạng lọc Wiener [12]

$$|\hat{\mathbf{X}}| = |\mathbf{V}| \odot \frac{\mathbf{W}_X\mathbf{H}_X}{\mathbf{W}_X\mathbf{H}_X + \mathbf{W}_Y\mathbf{H}_Y} \tag{7}$$

với $\odot$ ký hiệu cho phép nhân từng phần tử.

Tín hiệu tiếng nói trong miền thời gian có thể được khôi phục lại sử dụng biến đổi FFT ngược và phương pháp OLA (Overlap-Add) [13].

3. ƯỚC LƯỢNG VÀ LAN TRUYỀN ĐỘ KHÔNG ĐẢM BẢO

Trong phần này, chúng tôi sẽ mô tả ước lượng độ không đảm bảo của phổ tín hiệu thu được từ thuật toán nâng cao chất lượng tiếng nói và sự lan truyền từ miền phổ qua miền vec-tơ đặc trưng MFCC.

3.1. Ước lượng độ không đảm bảo

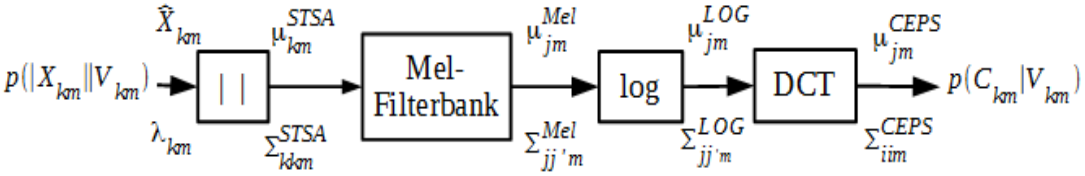
Ký hiệu $\hat{X}_{km}$ và V_{km} là phổ của tín hiệu tiếng nói ước lượng và tín hiệu trộn tại điểm tần số *k* và khung thời gian *m*, với $0 \le k \le F$ và $1 \le m \le N$.

Độ không đảm bảo được xác định dựa trên sai lệch giữa tín hiệu được ước lượng và tín hiệu trộn [14].

$$\lambda_{km} = \left| \hat{X}_{km} - V_{km} \right|^2 \tag{8}$$

3.2. Lan truyền độ không đảm bảo

Giá trị ước lượng $\hat{X}_{km}$ và độ không đảm bảo hay phương sai λ_{km} cần được lan truyền đến miền đặc trưng và sau đó sử dụng bởi bộ giải mã của ASR. Chúng tôi sử dụng thuật toán được đề xuất trong [15] để lan truyền độ không đảm bảo của phổ biên độ vào miền MFCC.



***Hình 1**. Sơ đồ khối lan truyền độ không đảm bảo cho đặc trưng MFCC[15].*

Hình 1 trình bày các bước thực hiện việc lan truyền độ không đảm bảo từ miền phổ qua miền MFCC. Đầu vào của quá trình lan truyền là mật độ phân bố xác suất hậu nghiệm của phổ biên độ tiếng nói sạch $|X_{km}|$ khi biết V_{km} , được giả thiết là phân bố Rice với ước lượng là $\hat{X}_{km}$ và phương sai là λ_{km} . Đầu ra của quá trình lan truyền là hàm mật độ phân bố xác suất hậu nghiệm của hệ số cepstra C_{km} khi biết V_{km}, là phân bố Gauss với trung bình là μ_{im}^{CEPS} và phương sai là Σ_{iim}^{CEPS}.

3.2.1. Lan truyền qua STSA (Short Time Spectral Amplitude)

Trung bình của độ không đảm bảo STSA được tính như sau:

$$\mu_{km}^{STSA} = \Gamma(1.5)\sqrt{\lambda_{km}}\ \mathrm{L}_{\frac{1}{2}}\left(-\frac{|\hat{X}_{km}|^2}{\lambda_{km}}\right) \tag{9}$$

Với Γ là hàm gamma và $\mathrm{L}_{\frac{1}{2}}$ là đa thức Laguerre.

Nếu xem xét các hệ số Fourier của phổ tín hiệu là độc lập thống kê thì ma trận hiệp phương sai của các hệ số phổ trong một khung thời gian là ma trận đường chéo, do đó phương sai của độ không đảm bảo STSA được tính

$$\Sigma_{kkm}^{STSA} = \lambda_{km} + |\hat{X}_{km}|^2 - (\mu_{km}^{STSA})^2 \tag{10}$$

3.2.2. Lan truyền qua bộ lọc thang Mel

Khi qua J bộ lọc Mel, trung bình μ_m^{Mel} và hiệp phương sai Σ_m^{Mel} của các đặc trưng được tính:

$$\mu_{jm}^{Mel} = \sum_{k=1}^{F} M_{jk}\mu_{km}^{STSA} \tag{11}$$

$$\Sigma_{jj'm}^{Mel} = \sum_{k=1}^{F} M_{jk} M_{j'k} \Sigma_{kkm}^{STSA} \tag{12}$$

Với $1 \le j, j' \le J$ và M_{jk} là hệ số bộ lọc Mel thứ *j* tại điểm tần số *k*.

3.2.3. Lan truyền qua tính toán loga

$$\mu_{jm}^{LOG} \approx \sum_{i=1}^{2J+1} W_i \cdot \log(S_{ji}) \tag{13}$$

$$\Sigma_{jj'm}^{LOG} \approx \sum_{i=2}^{2J+1} W_i \cdot \left(\log(S_{ji}) - \mu_{jm}^{LOG}\right) \cdot \left(\log(S_{j'i}) - \mu_{j'm}^{LOG}\right) \tag{14}$$

$$\begin{cases} W_1 & = \dfrac{\kappa}{J+\kappa} \\ S_1 & = \boldsymbol{\mu}_m^{MEL} \\ S_i & = \boldsymbol{\mu}_m^{MEL} + \left(\sqrt{(J+\kappa)\cdot\boldsymbol{\Sigma}_m^{MEL}}\right)_i \\ S_{i+J} & = \boldsymbol{\mu}_m^{MEL} - \left(\sqrt{(J+\kappa)\cdot\boldsymbol{\Sigma}_m^{MEL}}\right)_i \\ W_i & = \dfrac{1}{2(J+\kappa)} \quad \text{for } i \in \{2\cdots J+1\} \end{cases} \tag{15}$$

Với ký hiệu $(\,)_i$ tương ứng với hàng thứ *i* của ma trận và $\kappa = 3 - J$.

3.2.4. Lan truyền qua biến đổi Cosine rời rạc

$$\mu_{im}^{CEPS} = \sum_{j=1}^{J} T_{ij} \mu_{jm}^{LOG} \tag{16}$$

$$\Sigma_{iim}^{CEPS} = \sum_{j=1}^{J} \sum_{j'=1}^{J} T_{ij} T_{ij'} \Sigma_{jj'm}^{LOG} \tag{17}$$

với **T** là ma trận biến đổi cosine rời rạc DCT.

3.3. Giải mã sử dụng độ không đảm bảo

Trong hệ thống ASR sử dụng HMM, mỗi trạng thái *q* được mô hình bởi một hàm mật độ phân bố xác suất $p(\mathbf{c}_m|q) \sim \mathrm{N}(\mathbf{c}_m; \mu, \Sigma)$, hàm này nhận được từ pha huấn luyện dựa trên các vec-tơ đặc trưng $\mathbf{c}_m$ trích ra từ cơ sở dữ liệu huấn luyện cho trước. Trong pha nhận dạng, cho một quan sát $\mathbf{z}_m$, cần tính likelihood $p(\mathbf{z}_m|q)$. Khi tiếng nói cần nhận dạng sai khác với tiếng nói huấn luyện do nhiễu, likelihood $p(\mathbf{z}_m|q)$ có thể được tính như sau[7]:

$$p(\mathbf{z}_m|q) = \int_{\mathbf{c}_m} \frac{p(\mathbf{c}_m|\mathbf{z}_m)\,p(\mathbf{z}_m)}{p(\mathbf{c}_m)} p(\mathbf{c}_m|q)\,d\mathbf{c}_m \tag{18}$$

Nếu sự sai khác giữa $\mathbf{c}_m$ và $\mathbf{z}_m$ là không lớn thì có thể gần đúng:

$$p(\mathbf{z}_m|q) \approx \int_{\mathbf{c}_m} p(\mathbf{c}_m|\mathbf{z}_m) p(\mathbf{c}_m|q)\,d\mathbf{c}_m \tag{19}$$

Nếu giả thiết $p\left(\mathbf{c}_m|\mathbf{z}_m\right) \sim \mathrm{N}\left(\mathbf{c}_m; \hat{\mu}_{\mathbf{c}_m}, \hat{\Sigma}_{\mathbf{c}_m}\right)$ cũng là phân bố Gauss thì likelihood trong (19) có thể được tính:

$$p\left(\mathbf{z}_m|q\right) \approx \mathrm{N}\left(\hat{\mu}_{\mathbf{c}_m}; \mu, \Sigma + \hat{\Sigma}_{\mathbf{c}_m}\right) \quad (20)$$

Với vec-tơ đặc trưng là MFCC, $p\left(\mathbf{c}_m|\mathbf{z}_m\right)$ được xác định từ quá trình lan truyền độ không đảm bảo của phổ tiếng nói như mô tả ở mục 3.2.

4. THỬ NGHIỆM

Để đánh giá phương pháp nhận dạng tiếng nói trong môi trường nhiễu, chúng tôi xây dựng một hệ thống ASR cho tiếng Việt. Các mô hình âm học là các mô hình âm ba (tri-phone) sử dụng HMM với 5 trạng thái, trong đó mỗi trạng thái được mô hình bởi mô hình trộn Gauss GMM (Gaussian Mixture Model) và được huấn luyện từ cơ sở dữ liệu gồm các đoạn văn, các từ rời rạc và các số rời rạc. Dữ liệu huấn luyện được thu âm trong môi trường không nhiễu, gồm tiếng nói giọng miền Bắc của 7 nam và 5 nữ. Tổng số có 16309 file thu âm các chữ số rời rạc, các từ rời rạc, các câu và các đoạn văn ngắn, với thời lượng khoảng 10 giờ tiếng nói.
Do thuật toán nâng cao chất lượng tiếng nói và giải mã sử dụng thông tin về độ không đảm bảo tác động trực tiếp vào tín hiệu và mô hình âm học, vì vậy, chúng tôi xây dựng hệ thống nhận dạng chỉ các chữ số tiếng Việt. Điều này để tránh ảnh hưởng của mô hình ngôn ngữ.

Trong thí nghiệm thứ nhất, cơ sở dữ liệu đánh giá gồm 828 file, khoảng 1 giờ tiếng nói, chứa các số điện thoại được thu âm trong môi trường không nhiễu với giọng của 7 nam và 5 nữ, cũng là các giọng được sử dụng làm cơ sở dữ liệu huấn luyện. Mục đích của thí nghiệm này là để tạo kết quả tham chiếu cho các thí nghiệm đánh giá thuật toán nâng cao chất lượng tiếng nói và thuật toán giải mã sử dụng thông tin về độ không đảm bảo.

Trong thí nghiệm thứ hai, cơ sở dữ liệu tiếng nói đánh giá được trộn với âm thanh lấy từ các bản nhạc không lời. Tỷ số tín hiệu trên nhiễu SNR được trộn với 3 mức khác nhau là +5dB, 0dB và -5dB.

Chúng tôi sử dụng công cụ của Virtanen[1] cho thuật toán NMF được mô tả trong [16] cho hệ thống nâng cao chất lượng tiếng nói. Cơ sở dữ liệu dùng để huấn luyện cho mô hình NMF nhằm xác định ma trận **W** ở biểu thức (4) gồm: tiếng nói các đoạn văn được lấy ra từ tập huấn luyện mô hình HMM và một đoạn nhạc không lời khác với đoạn nhạc dùng làm nhiễu trong cơ sở dữ liệu đánh giá trong thí nghiệm thứ hai. Các tham số được chọn cho mô hình NMF là: số điểm tần số $F = 960$; số vec-tơ cơ sở $K = 384$ trong đó số vec-tơ cơ sở cho tiếng nói là 256 và cho nhiễu là 128; cửa sổ thời gian là Hamming 60ms với thời gian trượt giữa hai cửa sổ liên tiếp là 15ms; hàm mục tiêu sử dụng sai khác Kullback-Leibler.

Cho hệ thống ASR cơ sở, chúng tôi sử dụng công cụ Hidden Markov Toolkit 3.4.1, với vec-tơ đặc trưng 39 chiều bao gồm 13 hệ số MFCC , 13 hệ số delta và 13 hệ số delta-delta MFCC. Mô hình âm học là mô hình âm ba được huấn luyện sử dụng cơ sở dữ liệu tiếng nói sạch.

[1] http://www.cs.tut.fi/~tuomasv/software.html

Bảng 1 trình bày kết quả của hệ thống ASR, trong đó: "Hệ thống cơ sở" là hệ thống nhận dạng tiếng Việt cơ sở, với mỗi trạng thái của HMM sử dụng 6 GMM; NMF+ASR là Hệ thống cơ sở với tín hiệu tiếng nói được loại bỏ nhiễu sử dụng thuật toán NMF; NMF+UP+UD_ASR là Hệ thống cơ sở với tín hiệu tiếng nói được loại bỏ nhiễu sử dụng NMF sau đó độ không đảm bảo của tín hiệu được lan truyền đến miền đặc trưng MFCC sử dụng thuật toán đề xuất trong[15] và bộ giải mã sử dụng HVite của HTK có sửa đổi do Astudillo[2] viết cài đặt phương pháp giải mã với thông tin về độ không đảm bảo của vec-tơ đặc trưng. Với tín hiệu tiếng nói sạch, độ chính xác có thể đạt 97,96%. Tuy nhiên, khi tỷ số SNR thấp, độ chính xác giảm xuống rất nhanh. Nếu sử dụng thuật toán NMF nâng cao chất lượng tiếng nói, độ chính xác có thể nâng cao từ 2% đến 16% tùy vào mức SNR. Nếu kết hợp thêm thông tin về độ không đảm bảo thì độ chính xác có thể tăng thêm gần 1% so với chỉ áp dụng NMF nâng cao chất lượng tiếng nói. Trong các Bảng 2 và Bảng 3, khi cấu hình các trạng thái của HMM với số GMM lần lượt là 8 và 10, chúng ta cũng có kết quả tương tự. Kết quả của hệ thống ASR có kết hợp sử dụng thông tin về độ không đảm bảo của vec-tơ đặc trưng luôn cao hơn so với trường hợp hệ thống ASR chỉ sử dụng nâng cao chất lượng tiếng nói.

Bảng 1. *Độ chính xác nhận dạng của hệ thống ASR sử dụng 6 GMM cho một trạng thái HMM (%).*

Hệ thống ASR	SNR			
	Sạch	5 dB	0 dB	-5 dB
Hệ thống cơ sở	97.96	87.73	70.5	45.63
NMF + ASR		90.61	80.67	62.46
NMF+UP+UD_ASR		91.21	81.59	63.39

Bảng 2. *Độ chính xác nhận dạng của hệ thống ASR sử dụng 8 GMM cho một trạng thái HMM (%).*

Hệ thống ASR	SNR			
	Sạch	5 dB	0 dB	-5 dB
Hệ thống cơ sở	98.19	86.17	69.17	43.26
NMF + ASR		91.02	80.43	62.99
NMF+UP+UD_ASR		91.55	81.62	63.8

Bảng 3. *Độ chính xác nhận dạng của hệ thống ASR sử dụng 10 GMM cho một trạng thái HMM (%).*

Hệ thống ASR	SNR			
	Sạch	5 dB	0 dB	-5 dB
Hệ thống cơ sở	98.14	86.32	69.44	43.84
NMF + ASR		90.55	79.33	61.09
NMF+UP+UD_ASR		91.36	80.75	62.66

5. KẾT LUẬN

Trong bài báo này chúng tôi đã đề xuất ước lượng độ không đảm bảo của tín hiệu nâng cao chất lượng tiếng nói sử dụng NMF, sau đó, áp dụng kỹ thuật giải mã dựa trên thông tin độ không đảm bảo và vec-tơ đặc trưng MFCC cho hệ thống

[2] http://www.astudillo.com/ramon/research/stft-up/

ASR. Kết quả cho thấy với việc áp dụng nâng cao chất lượng tiếng nói sử dụng NMF trước khi đưa vào hệ thống nhận dạng đã nâng cao độ chính xác của hệ thống lên đáng kể khi tỷ số SNR thấp. Bên cạnh đó việc áp dụng thuật toán giải mã sử dụng thông tin về độ không đảm bảo của vec-tơ đặc trưng luôn làm tăng độ chính xác nhận dạng của hệ thống. Tuy nhiên, việc tăng độ chính xác chưa thực sự tốt. Điều này có thể do phương pháp ước lượng độ không đảm bảo của vec-tơ đặc trưng chưa tốt. Trong thời gian tới chúng tôi sẽ nghiên cứu phương pháp ước lượng độ không đảm bảo tốt hơn để từ đó có thể áp dụng nâng cao chất lượng của hệ thống nhận dạng tiếng nói bền vững. Đồng thời chúng tôi sẽ xây dựng cơ sở dữ liệu tiếng nói lớn hơn để tăng độ tin cậy của các kết quả đánh giá.

TÀI LIỆU THAM KHẢO

[1]. J. Benesty, J. Chen, and E. A. P. Habets, *"Speech Enhancement in the STFT Domain."* Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[2]. S. Boll, *"Suppression of acoustic noise in speech using spectral subtraction,"* Acoust. Speech Signal Process. IEEE Trans. On, vol. 27, no. 2, pp. 113–120.

[3]. Jae Lim and A. Oppenheim, *"All-pole modeling of degraded speech,"* IEEE Trans. Acoust. Speech Signal Process., vol. 26, no. 3, Jun. 1978, pp. 197–210.

[4]. K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, *"Speech denoising using nonnegative matrix factorization with priors,"* in IEEE ICASSP, 2008, pp. 4029–4032.

[5]. D. D. Lee and H. S. Seung, *"Learning the parts of objects by non-negative matrix factorization,"* Nature, vol. 401, no. 6755, Oct. 1999, pp. 788–791.

[6]. P. Smaragdis, B. Raj, and M. Shashanka, *"Supervised and Semi-supervised Separation of Sounds from Single-Channel Mixtures,"* in Independent Component Analysis and Signal Separation, vol. 4666, M. E. Davies, C. J. James, S. A. Abdallah, and M. D. Plumbley, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 414–421.

[7]. Li Deng, J. Droppo, and A. Acero, *"Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion,"* IEEE Trans. Speech Audio Process., vol. 13, no. 3, May 2005, pp. 412–421.

[8]. R. F. Astudillo and D. Kolossa, *"Uncertainty Propagation,"* in Robust Speech Recognition of Uncertain or Missing Data, D. Kolossa and R. Häb-Umbach, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 35–64.

[9]. A. Lefèvre, F. Bach, and C. Févotte, *"Itakura-Saito nonnegative matrix factorization with group sparsity,"* in IEEE ICASSP, 2011, pp. 21–24.

[10]. D. D. Lee and H. S. Seung, *"Algorithms for Non-negative Matrix Factorization,"* in Advances in Neural Information Processing Systems 13, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds. MIT Press, 2001, pp. 556–562.

[11]. D. L. Sun and G. J. Mysore, *"Universal speech models for speaker independent single channel source separation,"* in IEEE ICASSP, 2013, pp. 141–145.

[12]. B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, *"Non-negative matrix factorization based compensation of music for automatic speech recognition,"* in Proc. of Interspeech. Makuhari, 2010.

[13]. J. B. Allen and L. R. Rabiner, *"A unified approach to short-time Fourier analysis and synthesis,"* Proc. IEEE, vol. 65, no. 11, pp. 1558–1564, 1977.

[14]. D. Kolossa, R. Fernandez Astudillo, E. Hoffmann, and R. Orglmeister, *"Independent Component Analysis and Time-Frequency Masking for Speech Recognition in Multitalker Conditions,"* EURASIP J. Audio Speech Music Process., vol. 2010, pp. 1–13, 2010.

[15]. Fernández Astudillo, Ramón, *"Integration of Short-Time Fourier Domain Speech Enhancement and Observation Uncertainty Techniques for Robust Automatic Speech Recognition,"* 2010.

[16].T. Virtanen, B. Raj, J. F. Gemmeke, and H. V. hamme, *"Active-set newton algorithm for non-negative sparse coding of audio,"* in IEEE ICASSP, 2014, pp. 3092–3096.

ABSTRACT

ROBUST ASR BASED ON THE NONNEGATIVE MATRIX FACTORIZATION AND THE UNCERTAINTY-OF-OBSERVATION TECHNIQUE

In this paper, an improvement for the robustness of automatic speech recognition (ASR) based on the nonnegative matrix factorization (NMF) and uncertainty-of-observation technique is presented. While NMF technique could greatly improve the robustness of ASR however the processed signal always contains some residual noise or uncertainty. The uncertainty could be estimated and it could be used usefully for decoding in ASR. We evaluated this combination in the Vietnamese ASR system. The experiment results show that our combination method can be further improved the performance of ASR compared to NMF technique only in a very high nonstationary noise.

Keywords: Automatic speech recognition, Nonnegative matrix factorization, Uncertainty estimation.

Nhận bài ngày 15 tháng 07 năm 2016
Hoàn thiện ngày 15 tháng 08 năm 2016
Chấp nhận đăng ngày 17 tháng 08 năm 2016

Địa chỉ: [1] Viện Điện – trường Đại học Bách khoa Hà Nội;
[2] Viện nghiên cứu quốc tế MICA – trường Đại học Bách khoa Hà Nội.
* Email: cuong.nguyenquoc@hust.edu.vn