Noise-Conditioned Evaluation of a CASA-Inspired IBM–MFCC–SVM Pipeline for Small-Scale Speaker Identification

Uddhav Shivaji Shid[1], Dr. Suresh Shirbahudurkar[2]

[1]Zeal College of Engineering and Research, Pune, SPPU, Pune, Maharashtra India

[2]Dr. D.Y. Patil Technical Campus, SPPU, Pune Maharashtra India

DOI: https://doi.org/10.47772/IJRISS.2026.101000092

Received: 15 October 2026; Accepted: 21 October 2026; Published: 30 October 2026

ABSTRACT

Environmental noise can alter the spectral cues used for speaker identification and reduce the reliability of acoustic classifiers. This study evaluates a compact pipeline that combines a Computational Auditory Scene Analysis (CASA)-inspired Ideal Binary Mask (IBM) enhancement stage, Mel-Frequency Cepstral Coefficient (MFCC) statistics, and a multiclass Support Vector Machine (SVM) implemented through Error-Correcting Output Codes (ECOC). A locally collected database contains 77 recordings from four speaker classes; 54 recordings were used for training and 23 for testing. The evaluation considered babble noise, children-cheering noise, and random noise. The reported IBM configuration used a 32-ms frame, 4-ms frame shift, Hamming window, and local criterion of 1. Segmental signal-to-noise ratio (SegSNR) increased from −4.17 dB for noisy speech to 5.93 dB after enhancement, an improvement of 10.10 dB. Identification accuracy was 95.65% under babble noise, 73.91% under children-cheering noise, and 56.52% under random noise. These results indicate that performance depends strongly on noise type and that improved SegSNR does not necessarily ensure equivalent preservation of speaker-discriminative information. The study is a small-scale empirical evaluation; its results should be interpreted in light of the limited dataset and single hold-out test partition.

Keywords: speaker identification; computational auditory scene analysis; ideal binary mask; MFCC; support vector machine; environmental noise; SegSNR

INTRODUCTION

Speaker identification determines which enrolled speaker produced a given speech signal. It differs from speech recognition, which primarily estimates linguistic content. Speaker identification is used in access control, voice-enabled interfaces, and other applications where speaker identity is relevant. In practice, the acoustic signal may be corrupted by competing talkers, crowd sounds, or other environmental interference. Such interference can mask speaker-specific spectral patterns and degrade classification.

Speech enhancement is commonly used as a front end to reduce interference before feature extraction. Computational Auditory Scene Analysis (CASA) draws inspiration from the auditory system's organization of complex acoustic scenes. The Ideal Binary Mask (IBM) represents a time-frequency selection rule in which each time-frequency unit is retained or suppressed according to a criterion associated with target speech and interference. Binary masks are conceptually simple, although the exact mask-generation procedure and its dependence on reference signals must be specified for reproducibility.

Mel-Frequency Cepstral Coefficients (MFCCs) provide a compact representation of the short-time spectral envelope. Statistical summaries of MFCC trajectories can reduce variable-length recordings to fixed-length vectors. Support Vector Machines (SVMs) are a conventional margin-based classifier, and ECOC can extend binary learners to a multiclass problem. This paper investigates the combined pipeline under three noise conditions using a small, primary speech dataset.

The contributions are limited to (i) documenting a CASA-inspired IBM–MFCC–SVM pipeline, (ii) reporting the measured enhancement and identification outcomes for three noise types, and (iii) examining class-wise

errors and the difference between signal-quality improvement and speaker-identification accuracy. No claim is made that the pipeline outperforms alternative systems because matched baseline experiments were not reported.

Related Work

Auditory scene analysis describes how perceptual systems organize sound mixtures into meaningful sources [1], [2]. CASA methods translate some of these grouping principles into computational approaches to speech segregation. IBM-based methods use binary decisions over time-frequency units and have been studied as targets for speech separation [4].

MFCCs became a widely used representation in speech analysis through their perceptually motivated frequency warping, logarithmic filter-bank energies, and cepstral transformation [3]. Classical speaker classification methods include vector quantization, Gaussian mixture models, hidden Markov models, and discriminative classifiers. SVMs were introduced as margin-based learning systems and remain useful when the available training set is relatively small [6], [7]. Robust feature alternatives and speech-processing methods have also been investigated in the literature [5], [8]–[10].

The present study is not a comprehensive benchmark of these alternatives. It evaluates one configured processing chain and reports results under three noise conditions. A fair comparative claim would require running baseline systems on the same recordings, splits, and noise realizations.

PROPOSED METHODOLOGY

System overview

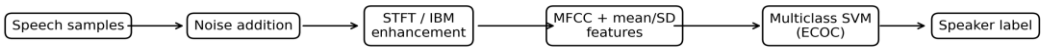

Fig. 1. Processing flow of the evaluated IBM–MFCC–SVM speaker-identification pipeline.

The processing sequence comprises speech acquisition, noise addition, time-frequency analysis, IBM-based enhancement, MFCC extraction, calculation of coefficient statistics, and multiclass SVM classification. The IBM stage is described as CASA-inspired because the supplied implementation details do not establish a full auditory grouping model.

Speech database and noise conditions

The primary dataset contains 77 speech recordings from four speaker classes, with 20, 20, 18, and 19 samples per class. Recordings were sampled at 16 kHz and stored in WAV format. The reported hold-out partition contains 54 training samples and 23 testing samples. Three noise conditions were considered: babble noise, children-cheering noise, and random noise.

A noisy signal can be represented as $y[n] = x[n] + v[n]$, where $x[n]$ is the clean speech signal and $v[n]$ is the noise component. The exact noise scaling, input signal-to-noise ratio (SNR), noise-file provenance, and randomization procedure were not specified in the supplied manuscript text and should be reported in the final submission if available.

Time-frequency analysis and IBM enhancement

The noisy waveform is transformed into the time-frequency domain using the short-time Fourier transform (STFT):

$Y(m,k) = \Sigma_n\, y[n]\, w[n - mH] \exp(-j2\pi kn/N)$

where w[n] is the analysis window, m is the frame index, H is the hop size, k is the frequency-bin index, and N is the transform length. The reported configuration uses a 32-ms frame length, 4-ms frame shift, Hamming window, and 16-kHz sampling frequency.

An IBM applies a binary mask M(m,k) to the time-frequency representation, giving the enhanced spectrum $\hat{Y}(m,k) = M(m,k)Y(m,k)$. A conventional ideal mask is commonly defined using a target-to-interference criterion at each time-frequency unit. In the supplied results, the local criterion is reported as 1; however, the exact mask equation and whether clean-speech/noise reference signals were used to construct the mask must be stated explicitly in the final methods section. If the mask requires clean target information, the method should be described as an oracle/reference-assisted enhancement experiment rather than a fully deployable real-time enhancer. The enhanced waveform is reconstructed by inverse STFT.

MFCC feature representation

MFCC extraction involves framing, windowing, spectral analysis, Mel filter-bank energy computation, logarithmic compression, and a discrete cosine transform. The Mel-frequency mapping can be written as:

$m(f) = 2595 \log_{10}(1 + f/700)$

For filter-bank energies E_i, cepstral coefficients can be expressed as $c_n = \Sigma_{i=1}^{M} \log(E_i) \cos[\pi n(i - 0.5)/M]$, subject to the convention used by the MFCC implementation. The MATLAB built-in MFCC function was used in the reported experiment. For each coefficient, the mean and standard deviation across frames were calculated, producing a fixed-length statistical feature vector. The number of retained coefficients and any normalization applied should be added from the MATLAB configuration before submission.

Multiclass SVM classifier

The resulting feature vectors were classified using a multiclass SVM with MATLAB's ECOC framework. For a binary SVM, the soft-margin objective is:

$\text{minimize } \tfrac{1}{2}\|w\|^2 + C\, \Sigma_i\, \xi_i$

$\text{subject to } y_i(w^T x_i + b) \geq 1 - \xi_i,\ \ \xi_i \geq 0.$

Here x_i is a feature vector, y_i is its binary label, w and b define the decision boundary, C is the regularization parameter, and ξ_i are slack variables. The multiclass ECOC framework combines binary learners to predict one of the four speaker classes. The SVM kernel, hyperparameters, and feature standardization procedure were not specified in the supplied text and should be documented for reproducibility.

Evaluation measures

Enhancement was evaluated using segmental signal-to-noise ratio (SegSNR), reported in decibels. Speaker classification was evaluated using accuracy, macro-averaged precision, macro-averaged recall, macro F1-score, and class-wise confusion patterns. Accuracy is the proportion of test samples assigned to the correct speaker. Precision and recall are computed per class and then macro-averaged across classes, as reported in the source results.

Experimental Setup and Results

Configuration

Parameter	Reported setting
Sampling frequency	16 kHz

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume X Issue X October 2026

Frame length	32 ms
Frame shift	4 ms
Window	Hamming
IBM local criterion	1
Dataset	77 recordings; 4 speaker classes
Train/test partition	54 training; 23 testing samples
Noise conditions	Babble; children cheering; random

Speech-enhancement performance

Speech condition	**SegSNR (dB)**
Clean speech	35.00
Noisy speech	−4.17
IBM-enhanced speech	5.93

The reported SegSNR increases by 10.10 dB from noisy speech (−4.17 dB) to IBM-enhanced speech (5.93 dB). This is evidence of improvement in the reported segmental signal-quality measure under the evaluated setup. It does not by itself establish that all speaker-discriminative cues were preserved.

Overall speaker-identification performance

Noise condition	**Accuracy (%)**	**Macro precision (%)**	**Macro recall (%)**	**Macro F1 (%)**
Babble	95.65	96.43	95.83	95.81
Children cheering	73.91	70.83	71.67	70.54
Random	56.52	63.75	55.36	57.99

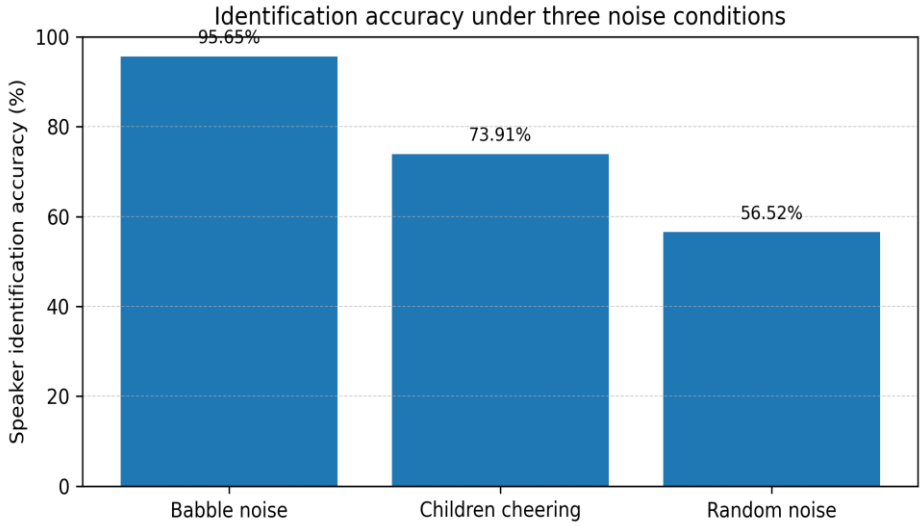

Fig. 2. Reported speaker-identification accuracy for the three noise conditions.

Babble noise produced the highest reported accuracy (95.65%), followed by children-cheering noise (73.91%) and random noise (56.52%). Since the test set contains 23 samples, one test sample corresponds to approximately 4.35 percentage points in accuracy. The results should therefore be interpreted as preliminary estimates rather than precise population-level performance claims.

Class-wise observations

Noise condition	Class label	Precision (%)	Recall (%)	F1-score (%)
Babble	Speaker10	85.71	100.00	92.31
Babble	Speaker11	100.00	83.33	90.91
Babble	Speaker12	100.00	100.00	100.00
Babble	Speaker13B	100.00	100.00	100.00
Children cheering	Speaker10K	100.00	100.00	100.00
Children cheering	Speaker11K	50.00	66.67	57.14
Children cheering	Speaker12K	33.33	20.00	25.00
Children cheering	Speaker13K	100.00	100.00	100.00
Random	Speaker10R	100.00	71.43	83.33
Random	Speaker11R	30.00	50.00	37.50
Random	Speaker12R	25.00	20.00	22.22
Random	Speaker13R	100.00	80.00	88.89

Under babble noise, the reported confusion matrix indicates that 22 of 23 test samples were correctly classified, with one Speaker11 sample assigned to Speaker10. Under children-cheering noise, the reported class-wise metrics indicate substantial difficulty for Speaker12K (20.00% recall and 25.00% F1-score). Under random noise, Speaker11R and Speaker12R show low F1-scores (37.50% and 22.22%, respectively). These observations suggest that errors are not evenly distributed across speaker classes.

DISCUSSION

Three findings emerge from the reported experiment. First, IBM enhancement increased SegSNR by 10.10 dB. Second, classification accuracy varied substantially with noise type, from 95.65% under babble noise to 56.52% under random noise. Third, class-wise metrics indicate that some speaker classes remain difficult to distinguish under children-cheering and random noise.

The contrast between SegSNR and identification performance is important: an enhancement stage can improve a signal-level metric while suppressing or distorting cues that contribute to speaker discrimination. The present results do not identify the causal acoustic features responsible for the errors, so explanations in terms of spectral overlap should be treated as hypotheses requiring additional analysis.

Several limitations affect interpretation. The dataset contains only 77 recordings and the test partition contains 23 samples. The reported single hold-out split may be sensitive to the composition of the test set. The supplied manuscript does not specify the exact SNR at which noise was added, noise source provenance, the random seed, the SVM kernel and hyperparameters, the MFCC coefficient count, or the precise IBM mask construction. These details should be added from the original MATLAB code and experiment log. A comparison against an

unenhanced MFCC–SVM baseline is also needed to quantify whether IBM enhancement improves identification accuracy rather than only SegSNR.

CONCLUSION

This paper reported a small-scale evaluation of a CASA-inspired IBM enhancement stage combined with statistical MFCC features and multiclass SVM classification for speaker identification. On a primary dataset of four speakers and 77 recordings, the reported SegSNR improved from −4.17 dB to 5.93 dB after enhancement. Identification accuracy was 95.65% under babble noise, 73.91% under children-cheering noise, and 56.52% under random noise. The results demonstrate that the measured performance depends on noise condition and that enhancement quality measured by SegSNR does not necessarily predict identification accuracy. Because of the limited dataset, single hold-out partition, and methodological details that remain to be documented, the findings should be considered preliminary. Larger datasets, repeated or speaker-independent validation, explicit SNR control, and matched baseline comparisons are required before broader generalization.

Future Work

Future experiments should (i) increase the number and diversity of speakers and recordings; (ii) report results across controlled input SNR levels and multiple noise realizations; (iii) repeat experiments with stratified cross-validation or speaker-independent splits where appropriate; (iv) compare noisy MFCC–SVM, IBM-enhanced MFCC–SVM, and other matched baselines; (v) evaluate alternative masking criteria and ratio masks; (vi) test complementary features such as GFCC and dynamic MFCC coefficients; and (vii) compare the compact pipeline with modern speaker embeddings using the same evaluation protocol.

REFERENCES

1. D. Wang and G. J. Brown, Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. Hoboken, NJ, USA: Wiley-IEEE Press, 2006.
2. A. S. Bregman, Auditory Scene Analysis: The Perceptual Organization of Sound. Cambridge, MA, USA: MIT Press, 1990.
3. S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357–366, 1980.
4. D. Wang, “On ideal binary mask as the computational goal of CASA,” in Speech Separation by Humans and Machines, New York, NY, USA: Springer, 2005, pp. 181–197.
5. C. Kim et al., “Power-normalized cepstral coefficients (PNCC) for robust speech recognition,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016. [Verify complete volume, issue, pages, and DOI before submission.]
6. C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
7. B. Schölkopf and A. J. Smola, Learning with Kernels. Cambridge, MA, USA: MIT Press, 2002.
8. J. Benesty, M. M. Sondhi, and Y. Huang, eds., Springer Handbook of Speech Processing. Berlin, Germany: Springer, 2008.
9. H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” Journal of the Acoustical Society of America, vol. 87, no. 4, pp. 1738–1752, 1990.
10. T. Virtanen, R. Singh, and B. Raj, eds., Techniques for Noise Robustness in Automatic Speech Recognition. Hoboken, NJ, USA: Wiley, 2012.