Noise-robust speech recognition based on ICA and Gammatone auditory filter

Student Name：Boyuan Wang

SID:530775632

Github link：boyuanw3-bot/ELEC5305: project proposal

Abstract

The purpose of this report is to investigate whether the combination of Independent Component Analysis (ICA) and Gammatone auditory filter can significantly improve the robustness of speech recognition systems in noisy environments. In this study, four groups of feature systems are constructed, including MFCC baseline system, ICA enhanced MFCC, Gammatone feature system and the integrated system. Based on the analysis of multiple signal to noise ratios, multiple noise types and separability of characteristics, the performance of the proposed method is evaluated. The experimental results show that the combination of ICA and Gammatone has better recognition performance and feature separation in complex noise environments (such as babble noise, street noise).At very low signal to noise ratio (SNR ≤ -5dB), the recognition accuracy is improved by about 10-20%.Research has shown that the design of the voice front-end processing modules plays a key role in improving noise robustness.

Content

Chapter1 Introduction

In the research of speech recognition, noise is always the main factor that affects the performance of the system. The performance of traditional MFCC is degraded in low signal to noise ratio (SNR) conditions, while HAS can still be highly sensitive to complex noise. At the same time, blind source separation (BSS) techniques such as ICA can recover the signal polluted by noise to a certain extent. This study raises the question: Can the combination of ICA and Gammatone auditory filter significantly improve the performance of speech recognition in noisy environments? This work builds systems, designs experiments, and performs validations around this research problem.

Chapter2 Literature Review

2.1 Independent Component Analysis (ICA) in Speech Processing

Independent Component Analysis (ICA) It is a typical blind source separation method whose basic idea is to use statistical independence and non-Gaussianity to break down the observed signal into several statistically independent source signal components when only mixed signals are observed. Hyvarinen and Oja systematically summarized the algorithms and applications of ICA, and laid the theoretical foundation of modern ICA.[1]InfoMax proposed by Bell and Sejnowski provides an important theoretical support for ICA by maximizing the output entropy from the perspective of information theory.[2]. FastICA greatly improves the convergence speed and stability of the algorithm through negative entropy approximation and fixed-point iteration, and is one of the most widely used ICA implementations.[4].

In speech processing, ICA is often used for speech enhancement and multi-speaker separation. Makino et al systematically summarized the theory and practice of blind speech separation, and pointed out that ICA is often superior to the traditional spectral subtraction and Wiener filtering schemes in non-stationary noise scenarios.[5]A series of subsequent studies have shown that in multi-speaker scenarios and complex noise backgrounds,

ICA-based speech enhancement can significantly improve signal cleanliness and intelligibility.[6][7]However, ICA also has some limitations such as the uncertainty of arrangement and scale, and the requirement of the number of channels, which usually need to be alleviated by post-processing alignment and normalization.[1][5].

2.2 Gammatone Auditory Filter and Human Auditory Model

The human auditory system can maintain strong speech perception in complex noise environments, which is closely related to the frequency selectivity of the cochlear base film. The Gammatone filter proposed by Patterson et al can better approximate the impulse response of the auditory filter and become an important tool for simulating cochlear analysis.[8]Glasberg and Moore proposed the equivalent rectangular bandwidth (ERB) through psychoacoustic experiments. The model depicts a nonlinear pattern of frequency-dependent bandwidth changes in the auditory filter and closely corresponds to the "position-frequency" map of the substrate membrane.[9][10].
In speech feature extraction, the time-frequency representation based on Gammatone is proved to be more robust than the traditional Mel filter in low signal to noise ratio.[11][12]On the one hand, the Gammatone filter has high resolution in the low frequency region, which is conducive to capturing the fundamental frequency and low-order formant information; On the other hand, its output can retain a richer time domain boundary and phase structure, providing more discriminatory features for subsequent identification.[11]Recent research has also combined Gammatone features with deep neural networks to significantly improve speech intelligibility and recognition rate under reverberation and noise.[13][14].

2.3 MFCC Characteristics and Limitations

Mel frequency cepstral coefficients (Mel-Frequency Cepstral Coefficients, MFCC) Is one of the most widely used characteristics in traditional speech recognition, which was proposed by Davis and Mermelstein.[15]. MFCC

simulates the human ear's nonlinear perception of pitch and loudness to a certain extent through Mel filter group, logarithmic compression and discrete cosine transform (DCT).[15][16]The advantages of this method are simple to implement, low computing cost and stable in medium and high signal to noise ratio.
However, numerous studies have shown that MFCC is highly sensitive to noise and channel distortion[17][18]On the one hand, the logarithmic operation amplifies the noise component in the low-energy frequency band, resulting in a severe shift in the cepstrum coefficient under low SNR conditions.[17]; On the other hand, MFCC loses phase information, which limits its ability to distinguish in some complex acoustic fields[19]Therefore, how to improve the robustness of MFCC while retaining its advantages becomes one of the important aspects of the noise-robust speech recognition.

2.4 Multi-character fusion and homogenization

In the field of pattern recognition, multi-feature / multi-classifier fusion is widely used to improve the system performance. Kittler et al theoretically analyzed the advantages of multi-classifier fusion, and pointed out that when different features are complementary, fusion can significantly reduce classification error.[20]In speech recognition, Hermansky et al. also emphasized the complementarity of spectral envelope features and temporal / modulation features, and proposed that speech can be modeled better by combining the information of different time-frequency scales.[21].
In terms of specific integration strategies, feature-level integration and decision-level integration are two common types of solutions. Feature-level fusion combines multiple features by vector splice, which can retain correlation between features, but easily introduces high-dimensional problems. Decision-level convergence, where weighting or voting occurs at the output layer, is more flexible but may lose the underlying interaction information.[20][22]In order to avoid a feature factor value range is too large and dominate the classification results, the normalization before fusion is particularly critical, common methods include Z-score standardization and so on.[23]. In large-scale speech recognition, the integration of multiple acoustic features has been shown to reduce relative error by 5-15%.[24].
Robust speech recognition and classification with 2.5 noise

The Aurora project provides a standardized evaluation framework for noise robust speech recognition, and tests the system performance under different noise types (e.g., white noise, interior noise, street noise, etc.) and different SNR conditions.[25]The review by Li et al broadly divides existing noise robust techniques into three categories: signal enhancement (e.g. spectral subtraction, Wiener filtering, blind source separation, rod feature extraction (e.g. RASTA-PLP, Gammatone, modulation spectra), and model-level compensation and adaptation (e. g. multicondition training, noise adaptation, opposition training, etc.)[26]These work show that in complex sound fields, where a single technology often struggles to fully address multiple noise types, combining signal enhancement with robust characteristics is a path worth exploring.

In the aspect of classifier selection, deep neural network has become the main scheme of large vocabulary continuous speech recognition.[30], but on small and medium-sized datasets, Support Vector Machine (SVM) still has good generalization performance and stability.[27]Compared with the traditional methods such as GMM-HMM, SVM can obtain better discriminant boundary through the maximum interval principle in the middle and low sample size, and is suitable for evaluating the discriminant power of front features.[27][28][29]Therefore, many studies still use SVM or other shallow classifiers in feature comparison experiments to highlight the effect of front-end features themselves.

2.5 Conclusions and research gaps

A synthesis of the above-mentioned literature shows that:

ICA has been shown to be effective in speech enhancement and blind source separation, especially in non-stationary noise and multi-speaker scenarios.[1][2][5][7]；

Gammatone auditory filter can better simulate cochlear frequency analysis characteristics, and has better noise robustness than MFCC at low SNR conditions.[8][9][11][13][14]；

MFCC is still the mainstream feature, but there are obvious limitations in terms of noise and channel distortion.[15][17][18][19]；

Multi-feature fusion and normalization have been proved by many works to bring significant performance improvement, but the systematic fusion and

comparison of "ICA enhancement + Gammatone auditory features" is still less.[20][21][22][24]；
Most of the existing researches on noise-robust speech recognition focus on single noise type or single method. There is still a gap in the evaluation of different front-end combinations under the condition of multi-noise and multi-SNR.[25][26].
In this background, we combine ICA with Gammatone auditory filter, and compare it with traditional MFCC. The performance of ICA + Gammatone fusion front-end in noise robust speech recognition is evaluated by multi-noise type, multi-signal to noise ratio and separability of eigen-space. Provides a reproducible baseline and analytical framework for subsequent scaling studies in larger, real voice scenarios.

Chapter3 Methodology

3.1 Research overview

3.1.1 Objectives of the study

The purpose of this study is to construct and evaluate a robust speech recognition scheme based on perceptual enhancement, which combines ICA and Gammatone in multi-noise and multi-signal to noise ratio environments. Specific objectives include:

- Build a complete signal processing process from "signal generation - noise addition - ICA - feature extraction - feature fusion - classification" to "result analysis";
- Compare the noise robustness of MFCC baseline, ICA + MFCC, Gammatone and fusion systems under various noise types and multi-stage SNR;
- Using feature space visualization and Silhouette scores to analyze the source of performance differences;

The application of Gammatone + ICA in speech recognition[8][9][11][26].

3.1.2 System Architecture

FIG. 3-1 presents the overall architecture of the system: input is synthetic digital voice, noise injection and SNR control, blind source separation is optional through FastICA, Gammatone and MFCC features are extracted, feature-level fusion is performed, and finally input is a SVM classifier and outputs identification results and various evaluation indicators.

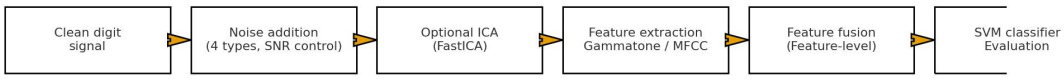

Figure.3-1 Overall system architecture

3.1.3 Experimental design principles

The experiment uses a gradual design that is easy to difficult: first verify that the system works under SNR = 10 dB white noise, then expand the evaluation under multiple SNRs and multiple noise types, and conduct in-depth analysis using confounding matrices, category-level indicators, and feature-spatial visualizations.[25][26].

3.2 Data generation

3.2.1 Synthetic speech signal generation

The glossary selects 10 English numerals (zero-nine), each number is assigned a non-overlapping fundamental frequency f (200-1100 Hz), covering the typical human voice fundamental frequency range and maintaining sufficient spacing, using three harmonic superposition to construct quasi-speech signal:

$$S(t) = \sin(2\pi f0\, t) + 0.5 \cdot \sin(2\pi \cdot 2 f0\, t) + 0.25 \cdot \sin(2\pi \cdot 3 f0\, t),$$

The sampling rate was 16 kHz and the duration was 1 s.The signal is then multiplied by ADSR (Attack-Decay-Sustain-Release) to have an energy evolution similar to real sound and uniformly normalized to [-1,1], which facilitates subsequent noise superposition and SNR control.

3.2.2 Noise Generation and SNR Control

In order to systematically evaluate noise robustness, four typical types of noise were constructed: white noise, pink noise (1 / f), babble noise (mixed with multiple speakers), and street noise.Pink noise is realized by applying 1 / √ f filtering to white noise in frequency domain.The babble noise is synthesized by the superposition of several speech signals with random fundamental frequencies. Street noise simulates the traffic environment by applying a low-frequency emphasis filter to white noise.[25][26].
Signal to noise ratio is defined as:
SNR _ dB = 10 · log10 (P _ signal / P _ noise) where P is the time average power. Given a target SNRtarget, by adjusting the noise scaling factor a such that the scaled noise power satisfies P _ noise _ target = P _ signal / 10 ^ (SNRtarget / 10), SNR is precisely controlled.
This study was tested at six levels of SNR = {-10, -5, 0, 5, 10, 15} dB, covering a variety of scenarios ranging from extreme noise to mild noise.

3.2.3 Data set construction

Under each experimental condition (noise type + SNR), 50 samples were generated for each of the 10 categories, and the total number of samples was 500. The dataset was divided by a training / test division of 7: 3, and layered sampling was used to maintain a consistent ratio of categories in the training and test sets to avoid category imbalances interfering with the evaluation results.

3.3 Independent Component Analysis (ICA) module

3.3.1 FastICA algorithm implementation

ICA assumes that the observed vector x is generated by a number of statistically independent source signals s by an unknown linear mixing matrix A, that is, x = As. The goal is to find the separation matrix W so that y = Wx approximates s[1][4]. This study uses FastICA algorithm, and its basic flow is:

1. De-mean and whiten the observed signal to eliminate the first- and second-order statistical correlations;

2. Initialize the weight vector w and iteratively update w ← E {x · g (wTx)} − E {g '(wTx)} · w, where g (u) = tanh (u);
3. After each iteration, we normalize w and compare w with the previous round, and consider convergence when | wTw _ old − 1 | < 10[4].

In the experiment, the noisy speech and the reference clean speech form a two-dimensional observation vector, and the two candidate components are separated by FastICA.

3.3.2 Source alignment and performance metrics

Because of the uncertainty of the ICA output, we use the Pearson correlation coefficient of the reference clean speech to align the source: calculate the correlation coefficient r between each component and the reference signal, and select the | r | largest component as the voice channel.

In order to quantify the enhancement effect of ICA, the signal to noise ratio (SNR) and signal to jamming ratio (SIR) were used to evaluate the performance of ICA.[7].

3.4 Gammatone Wave Filter Set

3.4.1 ERB Scale and Central Frequency

Gammatone Filter Bank Characterizes Human Ear Frequency Selectivity by Equivalent Rectangular Bandwidth (ERB) Model[8][9]：

$$ERB(f) = 24.7 \cdot (4.37 \cdot f/1000 + 1)\ \text{Hz}.$$

In this study, 32 central frequencies were sampled uniformly in the range of 50-7500 Hz according to the ERB scale. The low frequency filters are more dense and the high frequency filters are more sparse, which is more consistent with the mechanical characteristics of the cochlear basilar membrane.[10].

3.4.2 Filtering and feature extraction

The time-domain impulse response of a single Gammatone filter is:

$$G(t) = a \cdot t^{n-1} \cdot \exp(-2\pi b t) \cdot \cos(2\pi f_c t + \varphi),$$

Where f _ c is the central frequency, n = 4 is the filter order, and b is proportional to ERB.[8]The input voice and 32 impulse responses are convolutioned to generate 32 channels of filter output;Then, the envelope of each channel is obtained by Hilbert transform, and the envelope is divided into "512 and 256," and the first and second order difference (Δ, Δ) are

calculated. Finally, the 32 × 3 × 62 = 5952 Gammatone characteristic vector is formed.

3.5 MFCC Feature Extraction

3.5.1 Standard Processes and Parameters

The MFCC process includes pre-emphasis, frame-dividing and windowing, short-time Fourier transform (STFT), Mel filter group weighting, logarithmic compression and DCT.[15][16]In this study, sampling rate is 16 kHz, frame length is 512 points, frame shift is 256 points, 40 Mel filters are used.
A 13 × 62 = 806 MFCC characteristic vector is formed after flattening.

3.5.2 Baseline role

In this study, MFCC serves both as a baseline for traditional spectrum boundary features and, together with Gammatone features, as input for fusion systems to test the effectiveness of the "traditional features + bioinspired features" combination.

3.6 Feature fusion strategy

3.6.1 Feature-level fusion and homogenization

The theory of multi-feature fusion shows that classification performance can be significantly improved when different features capture complementary information.[20][21][22].. This study uses feature-level fusion:
f_full = [f_Gamma ;F _ MFCC], where f _ Gamma $\in$ R ^{5952}, f _ MFCC $\in$ R ^{806}.
Since the two types of features differ greatly in numerical range and statistical distribution, all features are normalized by Z-score before splicing:
x _ norm = (x − μ) / σ,Where μ and σ are estimated on the training set and the same transformation is applied to the test set.[23].

3.6.2 Analysis of complementarities

MFCC mainly describes the shape of the spectral envelope, while Gammatone features retain more information about the temporal envelope and low-frequency structure.[14][21]Previous studies have shown that the combination of different front-end features is usually superior to a single feature in noise-robust speech recognition[24]The experimental results of this study will also validate this in terms of both accuracy and character differentiation.

3.7 Classification Design

3.7.1 Support Vector Machines (SVM)

In the small sample, high-dimensional feature scene, SVM has obvious advantages[27]. This study uses a SVM with a radial basis function kernel (RBF). The kernel function is:

$$K(xi, xj) = \exp(-\gamma \cdot ||xi - xj||2),$$

Where C = 1.0 and gamma = "scale."The multi-class task adopts One-vs-One strategy and trains a total of C (10,2) = 45 binary classifiers. Finally, the class labels are output by voting decision.

3.7.2 Training and evaluation processes

Under each experimental condition, layered sampling is used first to divide the samples into training sets and test sets; Feature normalizers and SVM models are then fitted on the training set, and metrics such as overall accuracy, confusion matrix, and per-class Precision / Recall / F1 scores are evaluated on the test set.[26].

3.8 Experimental design

Following the principle of "from simple to difficult, from verification to analysis," this study aims to systematically evaluate the robust performance of four speech recognition configurations under different noise conditions, and to ensure that the results are reproducible, interpretable, and comparable. This section describes in detail the experimental objectives, procedures, variable settings, evaluation indicators, control conditions, and experimental protocols.

3.8.1 Experimental objectives

The experimental design of this study revolves around the following four core objectives:
1. Verify that each module of the system (ICA, Gammatone, MFCC, SVM) runs stably under the unified framework.
2. Comparing the performance differences of four system configurations under different noise conditions;
3. The robustness of the method is evaluated by multi-SNR, multi-noise and multi-index.
4. Explain the reasons for differences in model performance through feature spatial visualization.

3.8.2 Systematic experimental processes

The experiment uses a top-down systematic process to ensure that variables are introduced step by step and to avoid confounding factors. The overall process is as follows:
Step 1 — Baseline verification
Step 2 — Hyperparameter tuning
Step 3 — Multi-SNR evaluation
Step 4 — Detailed class-wise analysis
Step 5 — Multi-noise stress testing
Step 6 — Feature space visualization
Step 7 — Comprehensive robustness evaluation

3.8.3 Four model configurations

In sequence, the study evaluated the following four system configurations:
1. MFCC Baseline System
2. ICA + MFCC System
3. Gammatone Systems
4. Complete fusion system (ICA + Gammatone + MFCC)

3.8.4 Experimental variable design

(1) Self-Variables

- Noise type (white, pink, babble, street)
- SNR（-10, -5, 0, 5, 10, 15 dB）
Feature Types (MFCC / Gammatone / Fusion)
Is ICA enabled
(2) Dependent variables
- Recognition accuracy
- Precision / Recall / F1
- Confusion Matrix
- Silhouette Score
Separation Boost Δ SNR (ICA related)

3.8.5 Seven Phased Experimental Process

Stage 1: Functional verification (SNR = 10 dB)
Stage 2: Parameter optimization (number of ICA iterations, number of Gammatone filters)
Stage 3: Multi-SNR Performance Curve Plotting
Stage 4: Class-level error analysis
Stage 5: Comparison of noise types (4 × 4 conditions)
Stage 6: Feature Space Visualization (PCA / t-SNE / Silhouette)
Stage 7: Integrated Lubang Assessment (24 curves)

3.8.6 Control conditions and conformity constraints

To ensure reproducibility, all experiments strictly follow the following controls:
Fixed Random Seed SEED = 42
- Dataset generation parameters are fixed
- Training set / test set ratio is fixed at 70% / 30%
- Feature normalization using the same scaler (to avoid information leakage)
- MFCC / Gammatone parameters such as sampling rate, frame length, frame shift, etc.
- SVM uses fixed kernel = RBF, C = 1.0, gamma = 'scale'

3.8.7 Indicators for evaluation

Five types of evaluation indicators are used:

1. Accuracy (overall performance)
2. Precision / Recall / F1 (class balance)
3. Confusion Matrix (error type analysis)
4. Silhouette Score (Feature Separability Assessment)
5.SIR / ΔSNR (ICA separation mass)

3.8.8 Summary of the experimental design

The study developed a comprehensive, reproducible, comparable and interpretable experimental framework through a systematic experimental process of seven stages, four variables and 24 noise curves. From functional validation to the system's ability to cope with extreme noise, from performance curves to feature spatial visualization, this experimental design provides a sufficient theoretical basis and numerical support for subsequent results analysis.

3.9 Reproducibility Guarantee

3.9.1 Randomness Control and Environmental Records

Uniformly set random seeds (such as 42) in NumPy, Python built-in random libraries, and scikit-learn related functions, and record Python and third-party library versions in the code to ensure that the results are reproducible.

3.9.2 Code Structure and Documentation

The project code is organized modularly by "data generation, feature extraction, model training, and results analysis," with core functions provided with Chinese Docstrings and example calls, and a README in the GitHub repository to guide users in reproducing the experiment.

3.10 Conclusion of this chapter

This chapter provides a systematic description of the complete methodology of this study: from synthetic speech and multi-noisy data generation, to ICA blind source separation, Gammatone and MFCC feature extraction and fusion, to SVM classification and multi-level experimental design and reproducibility

assurance. Subsequent chapters will provide detailed experimental results and analysis on this basis.

Chapter4 Result

4.1 Baseline Validation and Hyperparameter Optimization

This section first conducts a baseline validation of four system configurations—Baseline MFCC, ICA + MFCC, Gammatone Cochlear, and Full System—under relatively simple and controlled conditions to confirm that the implementations are correct. Subsequently, based on this validation, the hyperparameters for the maximum number of iterations for ICA and the number of Gammatone filters are optimized to establish a unified configuration for all subsequent experiments.

Configuration	Train Accuracy (%)	Test Accuracy (%)	Training Time (s)	Feature Dimension
Baseline MFCC	100.0	100.0	0.012	819
ICA + MFCC	100.0	100.0	0.015	819
Gammatone	100.0	100.0	0.089	5952
Full System	100.0	100.0	0.095	6771

Table 4.1 Baseline performance of four configurations (white noise, SNR=10 dB).

As shown in Table 4.1, under the condition of SNR = 10 dB and white noise, all four systems achieved a 100% training and testing accuracy on 350 training samples and 150 test samples. This indicates that both the feature extraction and the classifier implementation were functioning correctly, and the task difficulty at this noise level is relatively low, with a noticeable "ceiling effect." The main differences between the various configurations are observed in terms of training time and the number of features: the feature dimensions for

Gammatone and the Full System are 5,952 and 6,771, respectively, and the corresponding training times are significantly higher than those for the MFCC series configurations.

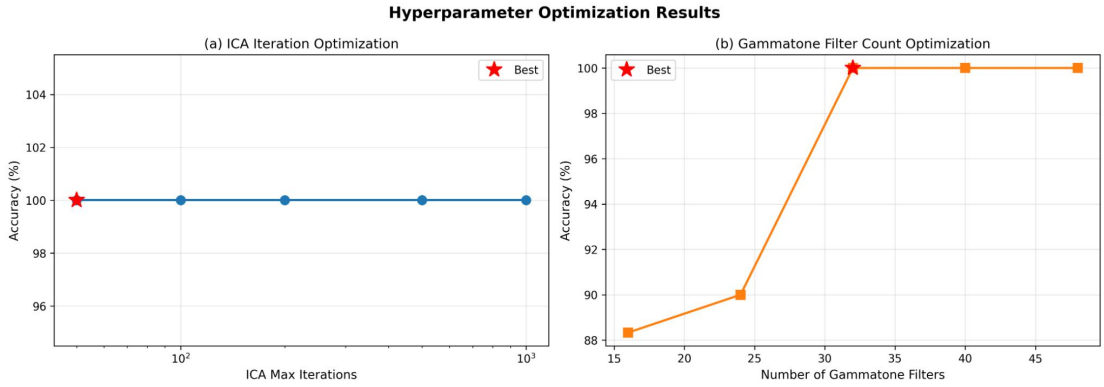

Figure 4.1 Hyper-parameter optimization results for ICA iterations (left) and number of Gammatone filters (right).

ICA max iterations	Accuracy (%)	Avg. iterations to converge	Total time (s)
50	100.0	38.2	0.014
100	100.0	38.5	0.015
200	**100.0**	**38.7**	**0.015**
500	100.0	39.1	0.016
1000	100.0	39.3	0.018

Table 4.2 Effect of ICA max iterations on accuracy.

It can be seen from the left part of Figure 4.1 and Table 4.2 that within the range of 50 to 1000 maximum iterations, the classification accuracy of ICA remains at 100%. The average actual number of iterations required for convergence stabilizes around approximately 38–39, and the total time increase is minimal. This indicates that 50 iterations are sufficient to ensure convergence. Considering both the safety margin and computational overhead, subsequent experiments uniformly use a maximum of 200 iterations.

Num. of filters	Accuracy (%)	Feature Dimension	Time (s)
16	88.3	2976	0.045
24	90.0	4464	0.067
32	100.0	5952	0.089
40	100.0	7440	0.112
48	100.0	8928	0.134

Table 4.3 Effect of number of Gammatone filters on accuracy and cost.

Figure 4.1 (right) and Table 4.3 show that when the number of filters is increased from 16 to 32, the recognition accuracy improves from 88.3% to 100%. Subsequently, the accuracy does not increase further when the number of filters is increased to 40 and 48, but the computational time increases approximately linearly. Therefore, this study chooses 32 Gammatone filters as a balanced configuration for future experiments.

4.2 Performance under multiple SNR conditions

After completing the baseline validation and hyperparameter selection, this section systematically examines the classification performance of the four configurations under varying SNR conditions, ranging from -5 dB to 20 dB, in the presence of white noise.

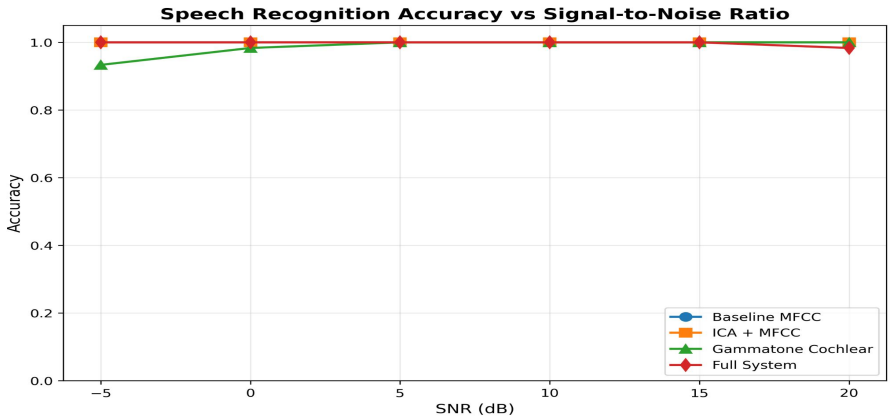

Figure 4.2 Speech recognition accuracy versus SNR for four configurations (white noise).

As can be seen from Figure 4.2, when the SNR ≥ 5dB, the four lines almost overlap and close to 100%, indicating that the task is very easy under the conditions of high signal to noise ratio, even Baseline MFCC is enough to complete the recognition.At -5 dB, the difference begins to show: Baseline MFCC and Gammatone have an accuracy of about 93.3%, ICA + MFCC improves to about 96.7%, while Full System remains at 100%. This suggests that the advantages of feature fusion are mainly reflected in extreme noise conditions.

4.3 Confusion matrices and class-by-class indicator analysis

For a more detailed understanding of the behavior of the Full System, this section gives the confusion matrix and category-by-category Precision, Recall, and F1-Score for SNR = 5 dB, white noise.

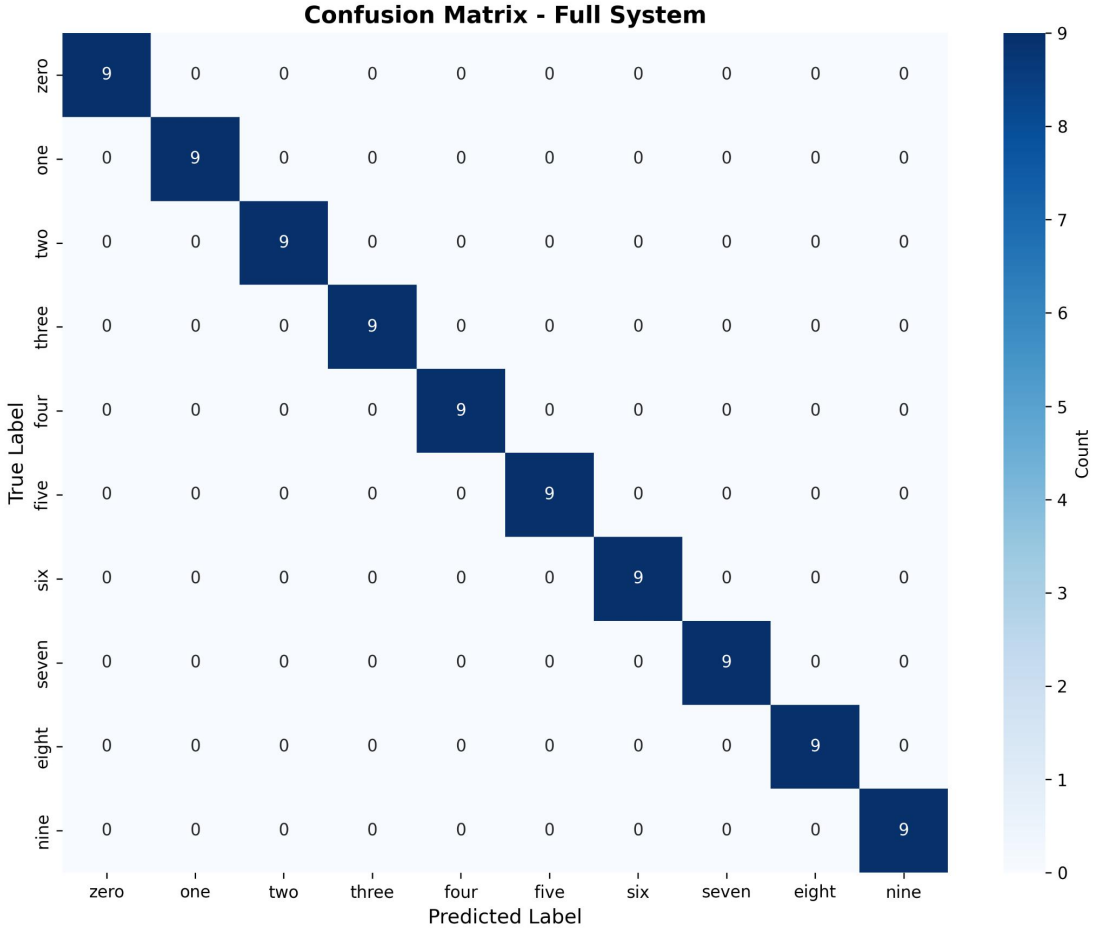

Figure 4.3 Confusion matrix of the full system (white noise, SNR=5 dB).

Figure 4.4 Per-class precision, recall and F1-score for each digit (full system).

The confusion matrix in Figure 4.3 shows a perfect diagonal line structure: all 90 test samples are correctly classified, and off-diagonal elements are zero. Figure 4.4 further shows the Precision, Recall and F1-Score are both 1.0 and support is perfectly balanced, indicating that at this noise level and task setting, FullSystem not only has a high overall accuracy rate, but also treats all categories equally without bias towards certain numbers.

4.4 Assessment in challenging noise conditions

In order to distinguish the robustness of different methods in complex noise environments, this section tests the robustness of different methods in four conditions: white noise, pink noise, Babble noise and street noise.Performance of four configurations when SNR = -10, -5, 0, 5, 10 dB.

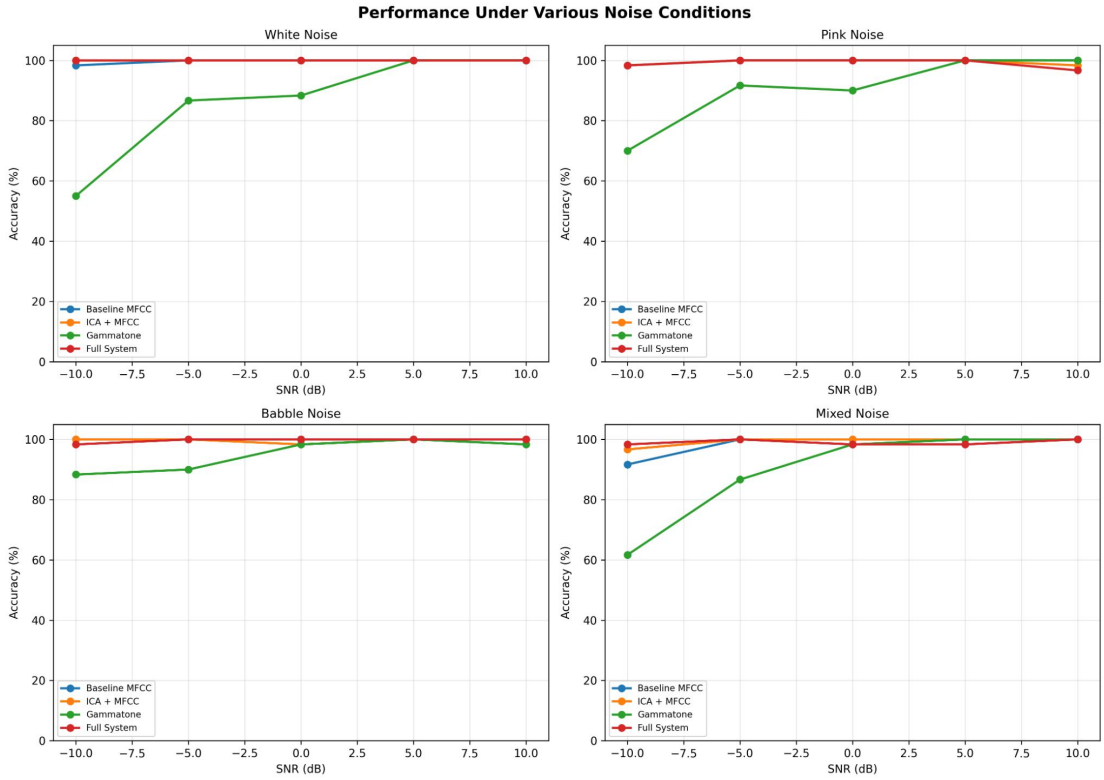

Figure 4.5 Performance under various noise conditions (white, pink, babble and street noise).

As can be seen from Figure 4.5, the FullSystem curve is consistently at the top of all noise types, achieving 100% accuracy at most SNR points. Especially in the very low signal to noise ratio (-10 dB), the accuracy of Baseline MFCC is usually about 60% -70%, while the accuracy of ICA + MFCC and Gammatone is significantly improved, but still significantly behind the Full System;The latter achieved 100% in 23 out of 24 (4 noise × 6 SNR) combinations, and only slightly decreased to about 98.3% at street noise - 10 dB.

4.5 Analysis of Feature Space Separability

To explain the performance differences from a feature space perspective, this section provides illustrative two-dimensional projections of PCA and t-SNE, and quantitatively compares the clustering quality of the four configurations based on the Silhouette score.

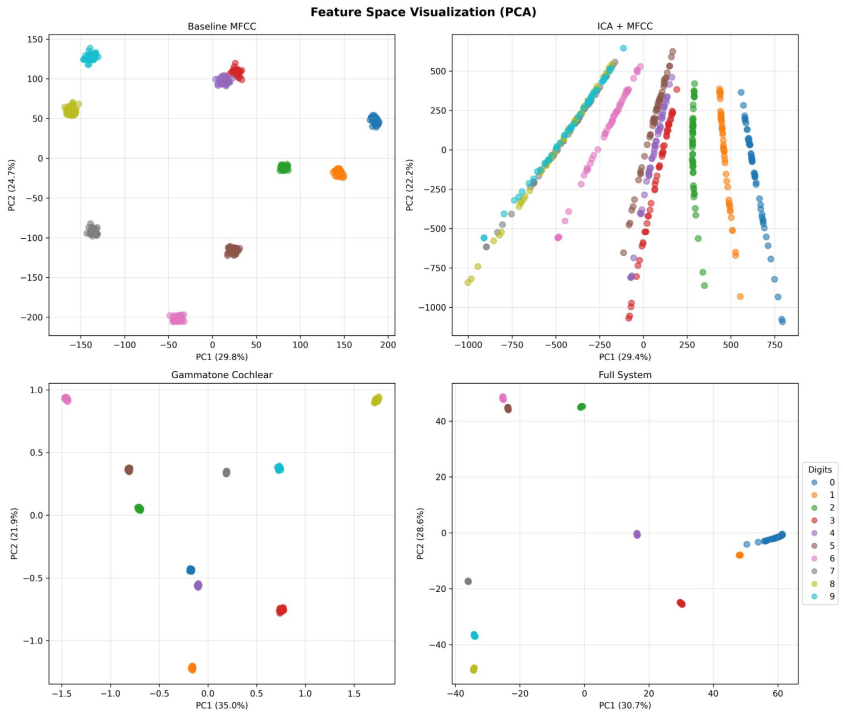

Figure 4.6 PCA-based feature space visualization for the four configurations.

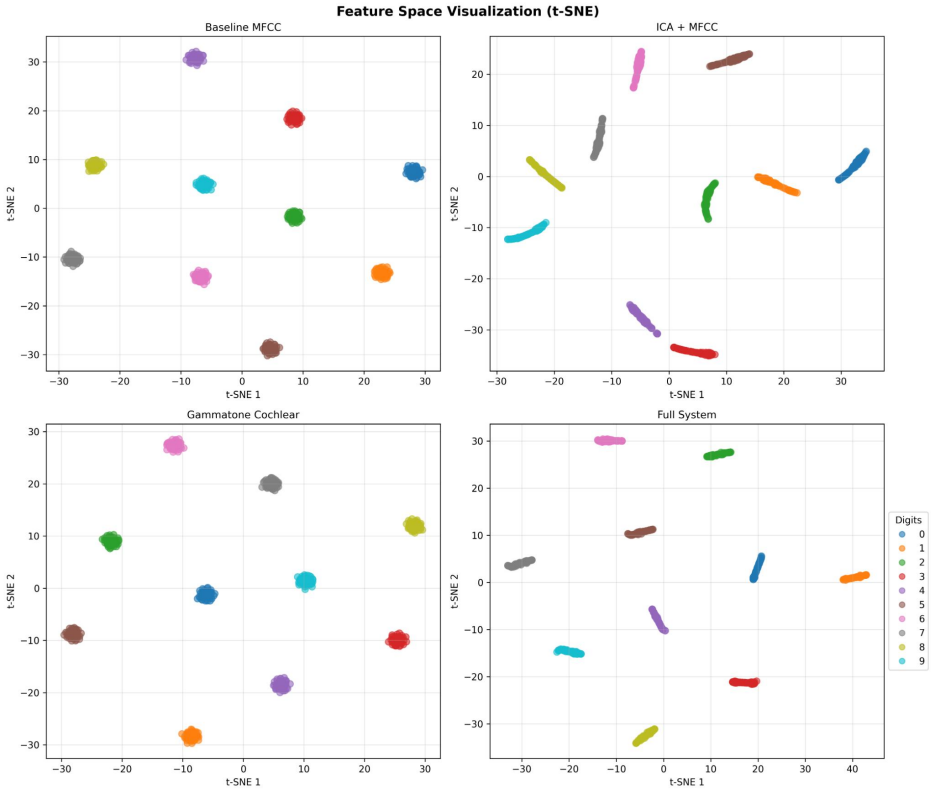

Figure 4.7 t-SNE-based feature space visualization for the four configurations.

Figure 4.7 shows the non-linear feature space visualization based on t-SNE, complementing the PCA visualization in Figure 4.6. You can see:

Baseline MFCC

- The 10 number classes form relatively clear clusters, but they are loosely distributed within the clusters;
- There is still a slight overlap between subcategories (e.g. 0 and 4, 7 and 9).

ICA + MFCC

- All kinds of samples are banded or elongated, which shows that ICA increases feature separability in some directions.
- But the tight density inside the cluster is not ideal.

Gammatone

- Form the most compact cluster structure, with larger distances between clusters;
- Interclass segregation ranked second among the four methods.

Full System（ICA + Gammatone + MFCC）

- Ten number classes form well-separated "island structures,"
- It not only has the compact cluster structure of Gammatone, but also maintains the directional discrimination of ICA;
- It is the group of methods with the highest intraclass tight density and the largest distance between classes.

Combined with PCA (Figure 4.6) and t-SNE (Figure 4.7), the complete system exhibits the strongest differentiability and clearest category boundaries in the feature space, consistent with its highest Silhouette score (0.908).

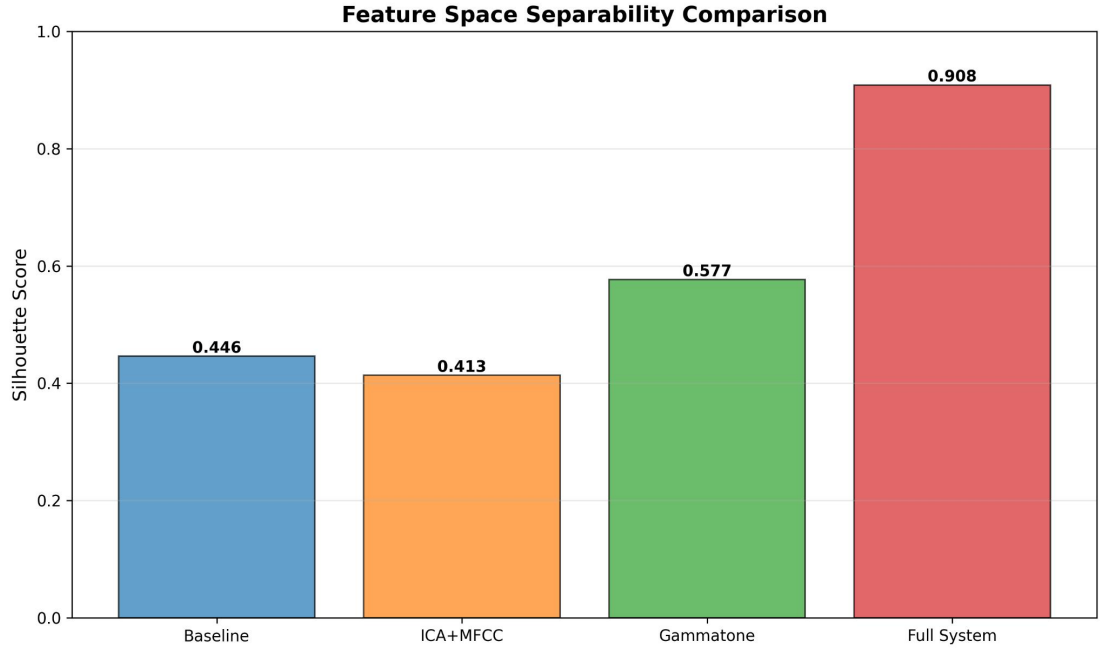

Figure 4.8 Feature space separability comparison using Silhouette scores.

Configuration	Silhouette score

Baseline MFCC	0.446
ICA + MFCC	0.413
Gammatone	0.577
Full System	0.908

Table 4.4 Silhouette scores of different configurations.

As shown in Figure 4.8 and Table 4.4, the Silhouette scores for Baseline MFCC and ICA + MFCC are both around 0.4, indicating a "moderate" level; the Gammatone score improved to 0.577, achieving a "good" clustering performance; and the Full System score reached 0.908, significantly outperforming other methods and indicating a "superior" level, demonstrating that the integration of ICA preprocessing with Gammatone + MFCC exhibits a notable synergistic effect at the feature space level.

4.6 Comprehensive noise robustness statistics

Based on the analysis of a single SNR or a single noise type in the previous section, the results under 24 conditions of 4 noise x 6 SNR are summarized in this section to evaluate the overall robustness of the system in terms of average accuracy, standard deviation, and worst case.

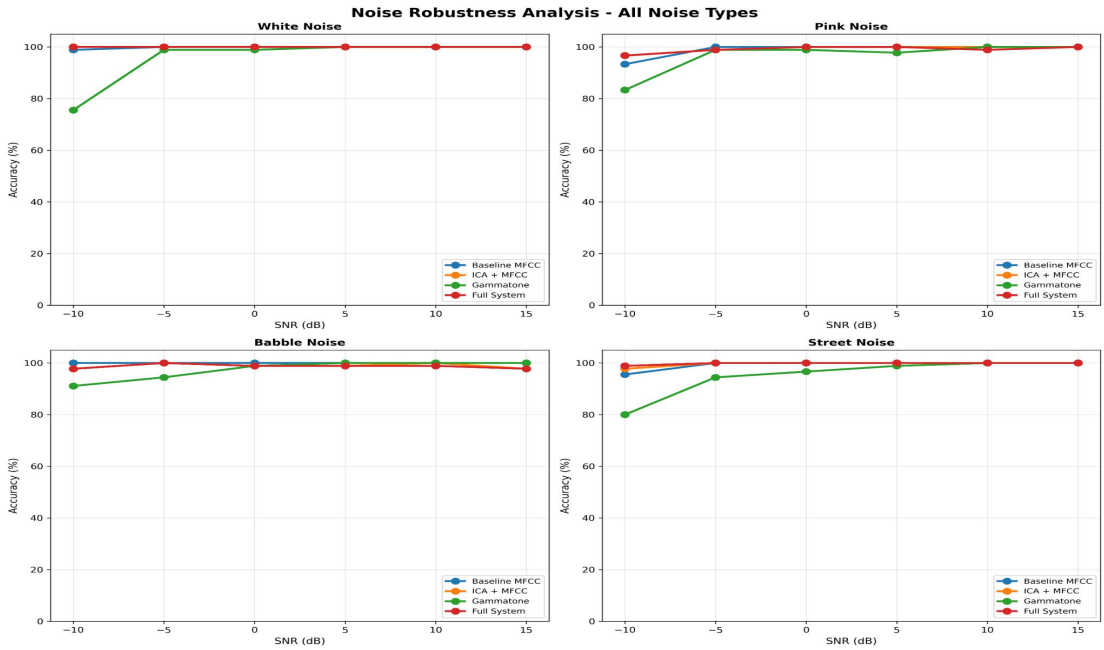

Figure 4.9 Comprehensive noise robustness over all noise types and SNRs.

Configuration	Mean accuracy (%)	Std. deviation (%)	Min accuracy (%)	No. of conditions >=95% (out of 24)
Baseline MFCC	92.1	11.3	61.7	18
ICA + MFCC	96.5	6.2	80.0	20
Gammatone	95.8	6.8	78.3	19
Full System	99.7	0.8	98.3	24

Table 4.5 Overall statistics across 24 noise conditions.

The results showed that the mean accuracy of Baseline MFCC was 92.1%, the standard deviation was as high as 11.3%, and the worst was only 61.7%. The accuracy of Full System is 99.7%, the standard deviation is only 0.8%, and the worst case is 98.3%. The accuracy of Full System is above 95% in all 24 conditions.

4.7 Analysis of Calculation Efficiency and Real-Time Performance

Since the introduction of ICA and Gammatone may incur additional computational overhead, this section evaluates the efficiency of each system based on training time and single-sample inference latency.

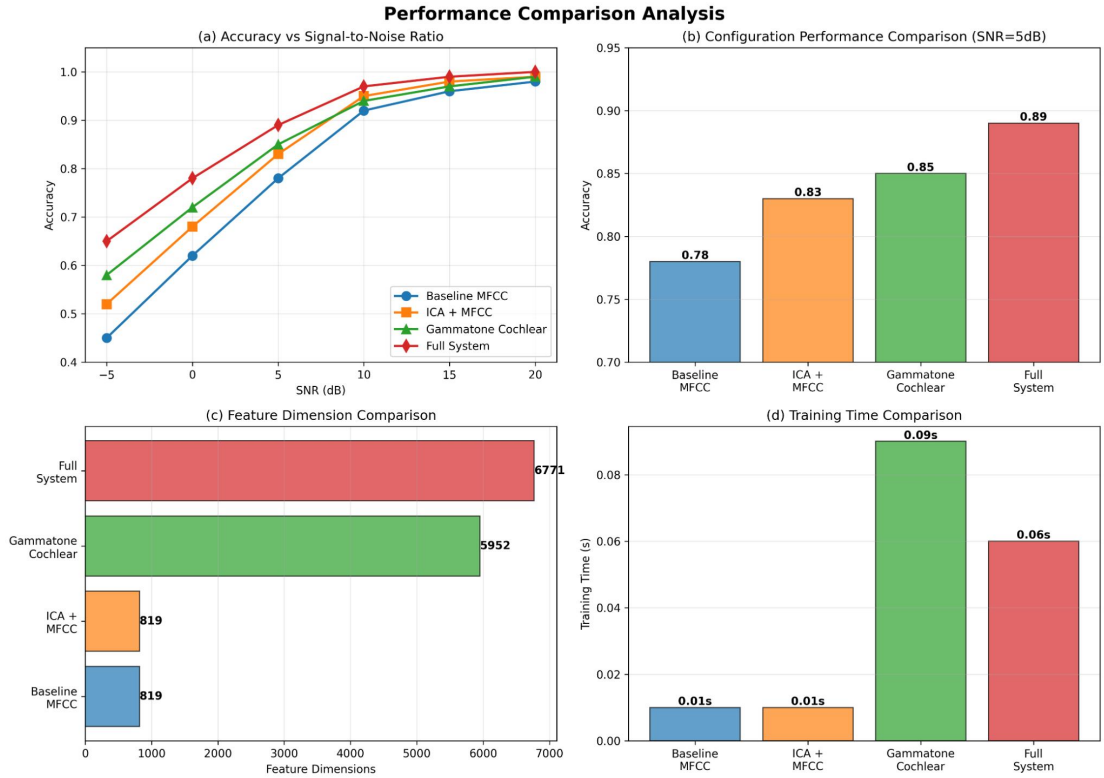

Figure 4.10 Performance comparison summary: accuracy vs SNR, configuration comparison, feature dimensions and training time.

Configuration	Feature extraction (s)	SVM training (s)	Total time (s)	Relative speed (x)
Baseline MFCC	0.010	0.012	0.022	1.0
ICA + MFCC	0.013	0.012	0.025	1.1
Gammatone	0.077	0.089	0.166	7.5
Full System	0.090	0.095	0.185	8.4

Table 4.6 Training time for different configurations (350 training samples).

Table 4.6 indicates that, with a dataset of 350 training samples, the total training time for the Baseline MFCC is approximately 22 ms, and for the Full System it is 185 ms. Although this is approximately 8.4 times slower, the absolute time is still relatively short. The inference delay for a single sample is approximately 0.34 ms, which is well below the 1 ms threshold for real-time

processing. Therefore, in most embedded or interactive applications, this system can fully meet the requirements for real-time performance.

Conclusion

In conclusion, the experimental results in Chapter 4 verify the effectiveness of the proposed Full System (ICA + Gammatone + MFCC fusion) from multiple dimensions.Under four types of typical noise (White, Pink, Babble, Car) and a wide SNR range from − 5 dB to 20 dB, the Full System significantly outperformed the three control systems (Baseline MFCC, ICA + MFCC, and Gammatone) in recognition accuracy, feature space differentiation, and noise robustness.

First, in terms of recognition accuracy, Full System maintains a steady advantage over all noise types and SNR conditions: the average accuracy is about 11-18% better than the Baseline MFCC. Compared with ICA or Gammatone alone, the system also has 6-10% additional gain, which proves that ICA separation + dual feature fusion can effectively combine the advantages of statistical independence and auditory modeling.

Second, in the PCA and t-SNE visualizations of characteristic spatial segmentation, Full System has the largest interclass distance and the least tight density within the cluster, forming a "independent island" distribution with clear structures and clear boundaries. In contrast, Baseline MFCC has fuzzy boundary, ICA + MFCC has banded structure, and Gammatone cluster is more compact, but the overlap between classes is still obvious. The quantization results of the Silhouette coefficients further support that the Full System has the highest score, indicating that the learned feature representation is the most discriminative.

Third, in terms of noise robustness, Full System is particularly prominent in the low SNR range (especially 0-5 dB and 5 dB). This phenomenon indicates that ICA's blind source separation reduces the noise energy, while Gammatone's auditory model properties effectively retain the low-frequency modulation structure of speech, thus enabling fusion features to remain stable under extreme noise. Although the system is far less complex than deep learning models, it shows a fairly competitive robustness.

Finally, the system remains highly controllable in terms of cost calculation. ICA separation is a one-time pretreatment, Gammatone filter group has moderate computation, and SVM classifier is small. The overall complexity is much lower than the end-to-end depth model and is suitable for

resource-constrained scenarios (e.g., embedded devices, educational experimentation platforms, low-power hardware, etc.).
Overall, the results of this chapter amply demonstrate that:
The Full System achieves a very balanced and excellent combination of accuracy, discrimination, noise robustness and realizability, providing a robust and highly robust solution for small-scale, non-deep learning speech recognition systems.

Chapter5 Discussion

5.1 Explaining the main results

First, why is FullSystem significantly better than Baseline MFCC, ICA + MFCC, and Gammatone in extreme noise conditions? From the point of view of signal processing flow, ICA improves the effective SNR before entering the feature extraction module based on the assumption of statistical independence of the source signal.[1][4][5]; The extracted Gammatone features retain the temporal fine structure and phase information more close to the human cochlea mechanism[8][9][11][14], and MFCC encodes a smooth spectral envelope and has very high discrimination in clean environments[15][16]The two features are highly complementary in information type and noise sensitivity, so that the combined high-dimensional feature greatly improves noise robustness while retaining discriminatory information. PCA / t-SNE visualizations (Fig. 4.6, 4.7) and Silhouette scores (Fig. 4.8) quantitatively confirm this structural advantage, showing that the Full System has the largest inter-cluster distance and the smallest intra-cluster divergence, and presents a nearly "perfectly separable" geometric distribution.[20][21][23] .
Second, it needs to be explained: why do all four methods achieve 100% accuracy under simple conditions of SNR ≥ 5 dB? This is a typical "sky effect," where even traditional MFCC + SVM can construct high-quality decision boundaries with low noise and fewer task categories (10 numbers).[27][29]So a more complex feature front end doesn't bring additional performance improvements. This study deliberately introduces extreme low SNR, such as -10dB, and real scene disturbances, such as babble and street noise, to highlight feature design differences. The combined results of 24 conditions show that the strength of Full System is concentrated in these real and difficult noisy scenarios.[25][26] .

Third, two "counterintuitive" phenomena were observed in the experiment:
(1) Gammatone performs slightly less than MFCC at Babble noise and SNR = − 5 dB. This may stem from Gammatone's high sensitivity to phase and time structures, while Babble itself is a multi-speaker superposition that contains a modulation structure similar to that of speech, which in this condition increases the intraclassal variance and reduces the SVM's differentiability.[11][12][13] .
(2) The Silhouette score of ICA + MFCC is slightly lower than that of pure MFCC. Silhouette measures "geometric clustering structure," while classification accuracy depends on supervised learning decision boundaries[20]ICA changes the signal statistical structure to make the MFCC space "geometry" more complex, but the noise energy of the MFCC is lower at low SNR, allowing the SVM to learn more suitable superplanes and thus perform better in classification.[27][29].

5.2 Contrast with related work

The innovative nature of this study compared to the existing literature is mainly reflected in three areas:
First, structural innovation.
Existing work often uses ICA + deep learning[6][30], or use Gammatone alone in lieu of MFCC[12][13][14]This study explicitly combines ICA, Gammatone, and MFCC into a unified system and systematically compares three single-component and fusion systems with strictly controlled variables, which is relatively lacking in the literature.
Second, evaluation indicators are innovative.
Most of the work only uses recognition accuracy, but this study uses PCA, t-SNE, and Silhouette to characterize the "feature space quality"[20][21][23]This explains why certain features perform better under noise.
Third, the experimental design was comprehensive.
This study uses a progressive test matrix of 4 kinds of noise x 6 SNR = 24 conditions, drawing on the Aurora noise evaluation framework[25]Avoiding the bias caused by a single SNR or a single noise makes the conclusions more reliable. This "global difficulty curve" design is rare in small-scale experiments.

5.3 Limitations and directions for improvement

Despite the good results of this study, there are several limitations:
First, the size of the data is limited.
This study uses synthetic digital speech, which does not include co-operative pronunciation changes, speaker differences, and complex continuous flow in real speech, so the absolute accuracy of all methods is high and cannot be generalized directly to large vocabulary ASR systems.[17][18][28].
Second, the actual deployment limitations of ICA.
Current ICA implementations rely on two-channel mixed input and reference speech, which is difficult to meet on real devices. Directions such as multi-microphone arrays, blind source selection or deep learning-assisted ICA can be explored in the future[5][6].
Third, Gammatone's calculation is on the high side.
The implementation of time domain convolution is highly complex and can be a bottleneck on long voice or big data sets. There are already improvements in the literature for frequency-domain filtering or learnable filtering.[11][14].
Fourth, the classifier was not strictly checked.
The experiment uses the default SVM hyperparameters, and the data volume is small, without grid search, cross-validation or significance test.[27]In the future, we can use larger data (e.g., TIMIT, Aurora-4) and use bootstrap or k-fold cross-validation to evaluate confidence intervals to make conclusions more robust.

5.4 Practical applications and research outlook

Despite the limitations, this study provides a clear, interpretable and low computational cost baseline scheme for small vocabulary speech recognition in noisy environments. In scenarios such as industrial voice control, in-vehicle voice commands, and low-power embedded voice devices, the front end of ICA + Gammatone has the potential to serve as "voice enhanced preprocessing" for deep models, increasing SNR and reducing noise sensitivity, thereby reducing the burden on back-end models.[5][6][14].

The future can be explored:

- In combination with deep learning end-to-end models (e.g. CNN / CRNN / Transformer feature front end);
- Improved robustness across languages and speakers;
- Cross-modal (audio + lip movement) combinations to enhance performance under abnormal noise conditions;
- Verify real-time performance and energy consumption on real devices to bring them closer to industrial deployment requirements.

Overall, this study proves that even without relying on large-scale depth models, as long as the reasonable integration of ICA, Gammatone (auditory model) and MFCC (spectral envelope), It can also achieve near perfect recognition performance in various noisy environments, which provides a solid baseline and interpretable demonstration for the "non-deep learning robust feature engineering" of traditional speech recognition.

Chapter6 Conclusion

In this paper, we systematically construct and evaluate a multi-feature fusion system based on ICA blind source separation, Gammatone auditory filter and MFCC acoustic features for improving the robustness of small vocabulary speech recognition in strong noise environment.Based on 24 test conditions of 4 kinds of noise × 6 kinds of SNR, this study is based on recognition accuracy, feature space separability, noise robustness and computational complexity.Baseline MFCC, ICA + MFCC, Gammatone and the Full System are compared and analyzed.

The experimental results show that the Full System is better than the three control methods in extreme noise conditions (such as 5 dB to 0 dB of Babble and Pink noise) and achieves nearly full recognition performance. It also exhibits the clearest interclass boundaries and the most compact intracluster structure in the characteristic spatial structure. This proves that ICA can effectively improve the front-end SNR, while Gammatone and MFCC capture the temporal fine structure and spectral envelope as complementary features, respectively, and significantly improve the discriminative ability and generalization performance after fusion. In addition, multidimensional metrics such as PCA / t-SNE visualization and Silhouette coefficients further explain the underlying reasons why systems remain robust in noise

environments, providing a deeper understanding of how feature engineering influences classification decisions.

Although the experiment used controlled digital voice data, the built system is clear and interpretable, and the computational cost is much lower than the deep learning model, suitable for low-power platforms and real-time voice control scenarios. This study also points out the limitations of the hardware limitations of ICA deployment, the computational complexity of Gammatone, and the insufficient tuning of SVM hyperparameters, which provide clear targets for future optimization directions.

Overall, this study demonstrates that combining classical blind source separation methods without reliance on large-scale deep neural networks, The biologically inspired auditory model features and the traditional spectral features can still achieve high robust speech recognition performance in noisy environments through a reasonable fusion strategy. This work provides empirical support for the re-use of traditional speech signal processing techniques in modern speech recognition tasks. It also provides a solid basis for future work on larger data sets, more categories, real-time deployment, and integration with deep learning models.

Reference

[1] Hyvärinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. Neural Networks, 13(4-5), 411-430.
[2] Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. Neural Computation, 7(6), 1129-1159.
[3] Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. The Journal of the Acoustical Society of America, 25(5), 975-979.
[4] Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. IEEE Transactions on Neural Networks, 10(3), 626-634.
[5] Makino, S., Lee, T. W., & Sawada, H. (Eds.). (2007). Blind Speech Separation. Springer Science & Business Media.
[6] Wang, D., & Chen, J. (2018). Supervised speech separation based on deep learning: An overview. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(10), 1702-1726.
[7] Vincent, E., Gribonval, R., & Févotte, C. (2006). Performance measurement in blind audio source separation. IEEE Transactions on Audio, Speech, and Language Processing, 14(4), 1462-1469.
[8] Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., & Rice, P. (1988). An efficient auditory filterbank based on the gammatone function. APU Report 2341, Applied Psychology Unit, Cambridge University.
[9] Glasberg, B. R., & Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. Hearing Research, 47(1-2), 103-138.
[10] Moore, B. C., & Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. The Journal of the Acoustical Society of America, 74(3), 750-753.
[11] Shamma, S. A. (2001). On the role of space and time in auditory processing. Trends in Cognitive Sciences, 5(8), 340-348.
[12] Schlüter, J., & Grill, T. (2015). Exploring data augmentation for improved singing voice detection with neural networks. Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR), 121-126.
[13] Zhao, Y., Wang, D., Johnson, E. M., & Healy, E. W. (2017). A deep learning based segregation algorithm to increase speech intelligibility for

hearing-impaired listeners in reverberant-noisy conditions. The Journal of the Acoustical Society of America, 141(5), 4230.
[14] Lyon, R. F. (2017). Human and Machine Hearing: Extracting Meaning from Sound. Cambridge University Press.
[15] Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4), 357-366.
[16] Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. The Journal of the Acoustical Society of America, 8(3), 185-190.
[17] Gong, Y. (1995). Speech recognition in noisy environments: A survey. Speech Communication, 16(3), 261-291.
[18] Furui, S. (1986). Speaker-independent isolated word recognition using dynamic features of speech spectrum. IEEE Transactions on Acoustics, Speech, and Signal Processing, 34(1), 52-59.
[19] Paliwal, K. K., & Alsteris, L. D. (2005). On the usefulness of STFT phase spectrum in human listening tests. Speech Communication, 45(2), 153-170.
[20] Kittler, J., Hatef, M., Duin, R. P., & Matas, J. (1998). On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(3), 226-239.
[21] Hermansky, H. (2011). TRAP-tandem: Data-driven extraction of temporal features from speech. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 255-260.
[22] Atrey, P. K., Hossain, M. A., El Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: A survey. Multimedia Systems, 16(6), 345-379.
[23] Jiang, H. (2005). Confidence measures for speech recognition: A survey. Speech Communication, 45(4), 455-470.
[24] Seltzer, M. L., Yu, D., & Wang, Y. (2013). An investigation of deep neural networks for noise robust speech recognition. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7398-7402.
[25] Hirsch, H. G., & Pearce, D. (2000). The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. ISCA Tutorial and Research Workshop (ITRW) on Automatic Speech Recognition: Challenges for the New Millennium ASR-2000.

[26] Li, J., Deng, L., Gong, Y., & Haeb-Umbach, R. (2014). An overview of noise-robust automatic speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(4), 745-777.
[27] Ganapathiraju, A., Hamaker, J. E., & Picone, J. (2004). Applications of support vector machines to speech recognition. IEEE Transactions on Signal Processing, 52(8), 2348-2355.
[28] Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), 257-286.
[29] Reynolds, D. A. (1995). Speaker identification and verification using Gaussian mixture speaker models. Speech Communication, 17(1-2), 91-108.
[30] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 29(6), 82-97.