

Voice Identification Based on the Use of Chirplet Transforms

Aliaskar Magzhan^{1*}, Mazakov Talgat², Mohamed Othman³, Jomartova Sholpan⁴, Mazakova Aigerim⁵

International Engineering and Technology University, 050060, Almaty, Kazakhstan^{1,2}

Al-Farabi Kazakh National University, 050040, Almaty, Kazakhstan^{1,2,4,5}

Malaysia Department of Communication Technology and Networks, Universiti Putra Malaysia, Serdang, Malaysia³

Laboratory of Computational Science and Mathematical Physics, Institute for Mathematical Research, Universiti Putra Malaysia, Serdang, Malaysia³

Abstract—Voice biometrics is based on extracting stable individual features from a speech signal. However, practical identification accuracy is significantly reduced by short fragments, variations in speech rate, noise, and mismatched recording channels. This study explores the potential of chirplet transforms as a specialized tool for analyzing the non-stationary time-frequency structure of speech. Unlike classical spectral analysis, the chirplet basis contains an additional parameter for the rate of change of instantaneous frequency and is therefore capable of describing local frequency trajectories associated with pitch dynamics, formant transitions, and articulatory features. A formalized feature extraction scheme is proposed, including preprocessing, cropping, calculation of a linear chirplet bank, normalized energy estimation, aggregation of frequency and chirp-rate statistics, and subsequent matching of feature vectors. Using a pilot dataset of 50 speakers, previously compiled from recordings of pronouncing the numbers from one to ten in Russian, the chirplet approach achieved identification accuracy of approximately 73-75%, including the ability to identify speakers not in the database. This article describes a fixed-dimensionality chirplet descriptor, normalization alternatives, an evaluation pipeline with leakage control, discriminability hypotheses, baselines, open-set calibration, and a specific score-fusion experiment for future validation. The proposed hybrid system remains a hypothesis until verified by independent methods. Therefore, the primary contribution lies in the analytical refinement and verifiable methodology, rather than in demonstrating superiority over existing speech recognition systems.

Keywords—Voice identification; voice biometrics; chirplet transform; time-frequency analysis; non-stationary signal; instantaneous frequency; speaker embeddings; noise resistance

I. INTRODUCTION

Automatic speaker recognition identifies or verifies speakers using measurable speech characteristics. Unlike automatic speech recognition, which evaluates linguistic content, speaker identification looks for acoustic features that remain relatively stable for an individual, despite variations in the spoken phrase or recording conditions. The vocal tract, laryngeal properties, articulation, tempo, and prosody all contribute to speaker variability [1]

Speaker-recognition research developed through several distinct contributions. Kinnunen and Li [1] surveyed text-independent speaker recognition from acoustic features to

statistical speaker models. Dehak et al. [2] proposed the total-variability i-vector representation. Variani et al. [3] investigated deep neural speaker representations for text-dependent verification. Snyder et al. [4] introduced x-vectors with statistical pooling. Desplanques et al. [5] developed ECAPA-TDNN with channel attention and multilevel feature aggregation. Wang et al. [6] proposed CAM++, emphasizing efficient speaker verification. Chen et al. [7] introduced WavLM as a self-supervised speech representation. Bai and Zhang [8] reviewed deep-learning approaches to speaker recognition and their major design choices. Sharma et al. [9] surveyed milestones and evolving techniques in speaker recognition. Nagrani et al. [10] introduced the VoxCeleb speaker-identification dataset, while Chung et al. [11] presented VoxCeleb2 as a larger speaker-recognition resource. Panayotov et al. [12] released LibriSpeech, a read-speech corpus suitable for controlled speech experiments, although it was designed primarily for speech recognition. These works provide the methodological and dataset context; none constitutes evidence that chirplet features outperform established baselines.

Speech is non-stationary: its instantaneous frequency, amplitude, and spectral envelope can change within individual phonemes. A global Fourier transform does not explicitly localize these changes in time. The short-time Fourier transform (STFT) introduces window-based localization, but retains a window-dependent resolution. Wavelets vary time-frequency resolution across scales, but do not, in their conventional form, explicitly parameterize a linear frequency slope. Mann and Haykin [13] established the physical and time-frequency foundations of chirplet analysis, describing additional transformations beyond conventional Gabor atoms. Mihovilovic and Bracewell [14] investigated adaptive chirplet representations on the time-frequency plane, motivating signal-dependent selection of analysis atoms. Peng et al. [15] extended the approach to polynomial chirplets for tracking nonlinear instantaneous-frequency trajectories. Yu and Zhou [16] formulated the general linear chirplet transform for time-frequency analysis. Together, these four works motivate explicit treatment of the chirp-rate parameter; they do not independently establish its biometric discriminative value.

Speech-oriented studies provide related, but not equivalent, evidence. Do et al. [17] evaluated linear-chirplet speech features for classification tasks. Do-Duc et al. [18] developed polynomial-chirplet-based speech feature extraction. Mishra et

*Corresponding author.

al. [19] explored chirplet time-frequency analysis for speech emotion recognition. Ge et al. [20] investigated graph-attention pooling for speaker recognition using self-supervised representations, offering a modern learned-feature comparison point rather than a chirplet baseline. Liu et al. [21] proposed a residual-optimized general linear chirplet transform for time-frequency feature extraction, illustrating developments in adaptive chirplet analysis. None of these results alone demonstrates improved open-set biometric speaker identification from chirp-rate statistics.

The earlier local pilot by Aliaskar et al. [22] considered 50 recordings of speakers reading Russian numbers from one to ten and reported approximately 73–75% identification accuracy using dominant chirplet-derived frequency and amplitude components. Aliaskar et al. [23] investigated voice identification using fundamental-harmonic detection, providing an earlier compact-feature approach. Shormanov et al. [24] investigated neural-network voice identification with attention to the Kazakh accent, offering related local speaker-recognition context. The pilot report [22] did not document repeated independent enrollment/test partitions, full decision logs, confidence intervals, or calibrated open-set trials. A single additional non-enrolled recording is insufficient to establish open-set rejection performance. The reported accuracy is therefore historical feasibility evidence, not a result of new experiments in this study.

The present contribution is a mathematical account of local chirp-rate matching together with an operational specification for a future, independently validated speaker-identification study. Its falsifiable question is whether chirp-rate statistics contribute speaker-specific information after controlling for spectral content, signal normalization, classifier capacity, and recording conditions. The proposed score fusion with deep embeddings is an experimental candidate, not an empirically established improvement.

Classic speaker recognition schemes were built around short-term spectral characteristics. The most common features were MFCCs, which approximate the spectrum to the nonlinear frequency scale of auditory perception and then decorrelate the logarithmic energies of the filters. Such features are convenient, compact, and remain a strong baseline for small datasets. A limitation of MFCCs is that they describe the local spectrum of a frame, but in their original form, they do not explicitly model the trajectory of instantaneous frequency change within a frame.

A valid evaluation must contrast the chirplet representation against MFCC statistics, a matched-capacity nonstationary baseline, and pretrained speaker embeddings on identical test trials. Performance claims must be based on repeated independent observations, uncertainty estimates, and ablation tests. The earlier pilot observations do not meet these requirements.

Speech contains quasi-periodic sections, noise-like consonants, pulse transitions, formant movements, and pitch modulations. Under these conditions, the assumption of stationarity is valid only over a limited interval. Even with a 20–30 ms frame rate, rapid transitions between phonemes and intonation changes can manifest as oblique or curvilinear structures on the time-frequency plane. If the analyzing atom is

not aligned with such a structure, the signal energy is distributed across several coefficients, which reduces the concentration of the representation.

However, these tasks are not equivalent to biometric identification. Gender or emotion recognition allows for the use of features common to large groups of speakers, whereas speaker identification requires maximizing inter-speaker distinctiveness while simultaneously maintaining robustness to text, emotional state, and channel. Therefore, transferring the chirplet representation to voice biometrics requires separate research: it is necessary to determine which transformation parameters yield the most stable speaker-specific signatures and whether the resulting coefficients are excessively sensitive to phonetic content.

An additional gap arises from the open set. In real-world biometric systems, the algorithm must not only select the most similar individual from the database, but also reject recordings of unknown speakers. This transforms the task from a standard multi-class classification into a threshold-based decision problem and requires analysis of ROC/DET, EER and minDCF. This protocol is proposed below as a mandatory step in full-fledged validation.

II. MATERIALS AND METHODS

A. Mathematical Model of Linear Chirp and Localized Chirplet

The simplest complex linear chirp with an initial frequency f_0 and a constant rate of change of frequency k can be written as:

$$s(t) = A \cdot \exp\{j2\pi[f_0 t + \frac{1}{2}kt^2 + \varphi_0]\} \quad (1)$$

where, A is the amplitude, φ_0 is the initial phase, and the instantaneous frequency is determined by the derivative of the phase: $f_i = f_0 + kt$. The quantity k , measured in Hz/s, directly determines the slope of the frequency trajectory. At $k = 0$ the atom degenerates into a local harmonic component; at $k \neq 0$ it becomes consistent with the increasing or decreasing instantaneous frequency.

For local analysis, the chirp is multiplied by a window function. A Gaussian window is a convenient choice, as it provides smooth localization and is well-related to classical Gabor atoms. A normalized linear Gaussian chirplet can be defined as:

$$\psi(t; \tau, \sigma, f, k) = (\pi\sigma^2)^{-\frac{1}{4}} \cdot \exp\left[-\frac{(t-\tau)^2}{2\sigma^2}\right] \cdot \exp\{j2\pi[f(t - \tau) + \frac{1}{2}k(t - \tau)^2]\} \quad (2)$$

The parameters τ , σ , f and k correspond to the time center, scale, central frequency, and chirp rate, respectively. A family of such functions forms a dictionary that can be discretized across multiple scales, frequencies and rates of frequency change.

For a signal, $x(t)$ the chirplet transform coefficient is determined by the scalar product with the complex conjugate of the atom:

$$C(\tau, \sigma, f, k) = \int x(t) \cdot \psi^*(t; \tau, \sigma, f, k) dt \quad (3)$$

In discrete form, the integral is replaced by a sum for samples $x[n]$. For biometric applications, what is most interesting is not the phase of an individual coefficient as such, but the stable distribution of the normalized energy $E = |C|^2$ over frequencies and chirp-rate. Unlike a one-dimensional spectrum, such a distribution distinguishes, for example, energy at a frequency of 800 Hz, which remains stationary, and energy passing through 800 Hz with a positive or negative slope.

$$E(\tau, f, k) = \frac{|C(\tau, f, k)|^2}{\varepsilon + \sum_{f, k} |C(\tau, f, k)|^2} \quad (4)$$

Normalization by local energy reduces the influence of absolute loudness and distance to the microphone. This parameter ε prevents division by zero. In practice, logarithmic compression $\log(E + \varepsilon)$ and normalization by utterance or by the accumulated statistics of the training set are also recommended.

III. RESULTS

A. Theoretical Part

For a rigorous analysis, let us consider the space $L^2(R)$ and the normalized Gaussian chirplet (2), for which $\|\psi(\cdot; \tau, \sigma, f, k)\|_2 = 1$. Let us denote the set of parameters by $\theta = (\tau, \sigma, f, k)$ and the signal transformation coefficient for a signal x by $C_x(\theta)$. The following properties formalize the stability, the transformation of frequency-time geometry, and the ability of the chirplet basis to concentrate the energy of locally linearly frequency-modulated speech components.

Theorem 1 (Lipschitz stability of chirplet coefficients). Let $x, y \in L^2(R)$, while $C_x(\theta)$ and $C_y(\theta)$ be calculated using Eq. (3) with the same normalized atom. Then, for any θ the estimate holds.

$$|C_x(\theta) - C_y(\theta)| \leq \|x - y\|_2 \quad (5)$$

In particular, with an additive observation model $y = x + n$ the change in the energy of an individual coefficient is limited by the inequality

$$|C_{x+n}(\theta)|^2 - |C_x(\theta)|^2 \leq 2\|x\|_2\|n\|_2 + \|n\|_2^2 \quad (6)$$

Proof. From the linearity of the scalar product $C_x(\theta) - C_y(\theta) = \langle x - y, \psi \rangle$. By the Cauchy-Bunyakovsky inequality $|\langle x - y, \psi \rangle| |\langle x - y, \psi \rangle| \leq \|x - y\|_2 \|\psi\|_2$, but normalizing the atom yields (5).

For $y = x + n$, let $\delta = \langle n, \psi \rangle$. Then $|C_x + \delta|^2 - |C_x|^2 \leq 2|C_x| * |\delta| + |\delta|^2$. Using $|C_x| \leq \|x\|_2$ and $|\delta| \leq \|n\|_2$, we obtain (6). The theorem is proven.

A consequence for voice identification: For a fixed bank of normalized chirplets, the “signal $\rightarrow$ individual coefficient” mapping does not amplify a small L^2 -perturbation. This does not prove the robustness of the entire classifier, but it does provide a strict upper bound on the sensitivity of the primary chirplet representation to additive noise and justifies subsequent studies of its robustness at different SNRs.

Theorem 2 (covariance with respect to quadratic phase modulation).

Let $y(t) = x(t) \cdot \exp\{j\pi k t^2\}$, where k is the constant rate of incremental change in instantaneous frequency. For the chirplet transform (3), the following holds:

$$C_y(\tau, \sigma, f, k) = \exp(j\pi k \tau^2) \cdot C_x(\tau, \sigma, f - \kappa \tau, k - \kappa) \quad (7)$$

Proof. In integral (3), let us introduce the local variable $u = t - \tau$. Then $t^2 = u^2 + 2u\tau + \tau^2$ and the quadratic phase factor are represented as $\exp(j\pi k \tau^2) * \exp(j2\pi k \tau u) * \exp(j\pi k u^2)$. After multiplying by the complex conjugate phase of the atom (2), the coefficients of u and u^2 become, respectively $f - \kappa \tau$ and $k - \kappa$. The factor $\exp(j\pi k \tau^2)$ independent of u is factored out of the integral, which directly yields (7). The theorem is proven.

Consequently, the chirplet representation module retains its shape under quadratic phase modulation and undergoes a predictable shift along the chirp-rate axis, with a simultaneous shift in the local center frequency. For a voice signal, this is important when analyzing sections with similar spectral envelopes, but different pitch or formant velocities: such dynamics do not disappear, but are encoded in the k coordinate.

Theorem 3 (energy concentration of a matched chirplet). Let a local voice fragment in the window under consideration be approximated by a single linear frequency-modulated component $x(t) = A\psi(t, \tau, \sigma, f_0, k_0)$, where ψ is given by Eq. (2). For fixed τ and σ we denote $\Delta f = f - f_0$ and $\Delta k = k - k_0$. Then the modulus of the coefficient for arbitrary f and k is equal to

$$\frac{|C_x(\tau, \sigma, f, k)|}{|A|} = [1 + \pi^2 \sigma^4 \Delta k^2]^{-\frac{1}{4}} \cdot \exp\left\{-\frac{\pi^2 \sigma^2 \Delta f^2}{1 + \pi^2 \sigma^4 \Delta k^2}\right\} \quad (8)$$

Proof. Since both atoms have the same τ and σ , after the substitution $u = t - \tau$ the scalar product reduces to a Gaussian integral with a complex quadratic coefficient:

$$\frac{C_x}{A} = (\sqrt{\pi\sigma})^{-1} \int \exp(-u^2/\sigma^2 - j2\pi\Delta f u - j\pi\Delta k u^2) du,$$

where, $\Delta f = f - f_0$ and $\Delta k = k - k_0$.

Setting $a = \sigma^{-2} + j\pi\Delta k$ and $b = -j\pi\Delta f$, we obtain $\int \exp(-au^2 + bu) du = \sqrt{(\pi/a)} \exp\left(\frac{b^2}{4a}\right)$, $\text{Re}(a) > 0$.

After substitution and taking the modulus, we obtain (8). The first factor on the right-hand side does not exceed unity and is equal to unity only when $\Delta k = 0$; the exponential factor also does not exceed unity and when $\Delta k = 0$ reaches unity only when $\Delta f = 0$. Consequently, the global maximum $|C_x| = |A|$ is achieved when $f = f_0$ and $k = k_0$. The theorem is proved.

For comparison with windowed Fourier/Gabor analysis, we set $k = 0$. If the true local rate of change of frequency $k_0 \neq 0$ and the center frequency are chosen correctly ($f = f_0$), the relative matching amplitude is

$$R_{\text{Gabor/STFT}} = \frac{|C_x(\tau, \sigma, f_0, 0)|}{|C_x(\tau, \sigma, f_0, k_0)|} = (1 + \pi^2 \sigma^4 k_0^2)^{-\frac{1}{4}} < 1, \quad k_0 \neq 0 \quad (9)$$

Eq. (9) quantitatively demonstrates the source of the advantage of the additional chirp-rate coordinate: for a non-zero slope of the instantaneous frequency, a fixed Gabor/STFT atom

loses energy concentration, with the loss increasing with $|k_0|$ the effective window duration σ . A matched chirplet $k = k_0$ eliminates this mismatch. For speech, this statement applies locally to quasi-harmonic components, formant transitions, and pitch-glide, which allow a linear approximation of the instantaneous frequency over a short interval.

A set of three theorems provides the mathematical basis for the further experiment. Theorem 1 answers the question of the sensitivity of the representation to noise; Theorem 2 describes the geometric effect of changing the frequency slope; Theorem 3 shows that the inclusion of the parameter k is not just a heuristic expansion of the vocabulary but leads to an analytically measurable energy concentration for locally chirp-like components.

B. Experimental Part

Direct use of the full four-dimensional coefficient array is computationally expensive and sensitive to the recording length.

Therefore, it is proposed to combine it into a fixed vector. Let the tensor be calculated for each frame $E_m(f, k)$, where m is the frame number. For each frequency range and each chirp rate, the mean, standard deviation, and high energy quantile are calculated. Additionally, features of temporal-frequency concentration, entropy, and symmetry with respect to are introduced $k = 0$.

$$z = [\mu(E), \sigma(E), Q_{0.9}(E), H(E), \rho_+, \rho_-, cTF] \quad (10)$$

Here $H(E)$ is the entropy of the normalized distribution, ρ_+ and ρ_- are the energy fractions at positive and negative chirp rates, and is the concentration index. A higher concentration means that a significant portion of cTF the energy is described by fewer atoms, which indirectly indicates a good fit of the dictionary to local signal trajectories. For speaker identification, it is possible to aggregate these statistics across multiple frequency ranges, including the low-frequency pitch region and the mid-frequency formant regions.

In [22], a significantly more compact description was used: after sorting the significant components, three harmonics were selected, which yielded six characteristics-three frequency and three amplitude coefficients. This scheme demonstrates the principle well, but to improve accuracy, it is advisable to move from six coefficients to a statistically stable multidimensional description. In this case, the original six-dimensional model is preserved as an ablation baseline: it allows you to quantitatively check whether the chirp-rate extension really gives an additional effect.

For closed-loop identification, the simplest option is to compare the test record vector z with reference centroids $\bar{z}_i$. After feature standardization, cosine similarity or the Mahalanobis distance can be used. For more complex decision boundaries, SVM, LDA/PLDA, or a compact multilayer network are applicable. For small samples, SVM provides a reasonable baseline, as it does not require a large number of parameters.

$$\text{sim}(z, \bar{z}_i) = \frac{z^T \bar{z}_i}{\|z\|_2 \cdot \|\bar{z}_i\|_2} \quad (11)$$

In an open set, an identity is selected only if a threshold is exceeded θ : if $\max_i \text{sim}(z, \bar{z}_i) < \theta$, the entry is marked as unknown. The threshold should be selected exclusively for the development subset and then be fixed until the test is evaluated on the test subset. For the verification task, the threshold is varied, and ROC/DET curves are plotted; the EER corresponds to the point where false acceptance rate and false rejection rates match.

The previously described set of speech recordings [22] was used as the initial proof-of-concept. It included 50 WAV files prepared by students from the International University of Engineering and Technology (Almaty). Each participant recited a sequence of numbers from one to ten in Russian. In addition, a recording of a person not included in the database was separately analyzed, allowing us to test the fundamental possibility of rejecting an unknown speaker.

The 73–75% figure is taken from [22], not obtained through a new experiment in this article. The exact enrolled/test division, trial-level outputs, number of trials, and uncertainty were not reported, and cannot be reconstructed from the aggregate figure. Accordingly, Tables I–III contain a proposed design and a critical summary rather than new observations or a comparative benchmark.

Table I distinguishes historical pilot observations from planned independent validation; its extended-design column reports no completed experiment.

It is important to emphasize that these data should be considered a pilot evaluation of the method's feasibility. The published description lacks a full protocol for multiple train/test splits, the number of independent trials, and confidence intervals. Therefore, the 73–75% range cannot be used as evidence of statistical superiority over alternative models. This study retains this result as a starting point and formulates an expanded protocol that addresses key limitations.

TABLE I. TRANSITION FROM THE PILOT TO THE EXTENDED EXPERIMENTAL DESIGN

Characteristic	Pilot design [22]	Extended design
Number of speakers	50	50 local speakers + public corporation
Content	Numbers 1–10, Russian	Text-dependent and text-independent modes
Format	WAV	16 kHz mono WAV/FLAC after standardization
Features	3 frequencies + 3 amplitudes	chirp-rate statistics + MFCC + embeddings
Setting	Closed-set + one unknown	Closed-set, verification, open-set
Distortions	Not systematically evaluated	AWGN, reverberation, short utterances
Metrics	Accuracy	Accuracy, macro-F1, EER, minDCF, CMC

For a reproducible experiment, all recordings must be converted to a single format: one channel, 16 kHz, the same bit depth or floating-point representation. The signal level is normalized by RMS or by the loudness standard. Normalization should be applied without aggressive compression that could distort biometric characteristics. It is recommended to use voice

activity detection before feature extraction. Activity detection should remove long pauses and leave only speech intervals.

Noise reduction should be separated from feature extraction itself. If the denoising algorithm is optimized on test data, leakage occurs, and the result becomes overestimated. Therefore, the FFT parameters, thresholding, and spectral gating or other filtering are defined on the development subset. In the main series of experiments, it's useful to compare two modes: without pre-denoising and with fixed denoising. This will help us understand whether the chirplet representation offers inherent noise immunity or benefits solely from pre-processing.

Speech is divided into overlapping frames 25-40 ms long. For chirplet analysis, it is useful to additionally examine longer windows of 60-100 ms, as frequency slope estimation requires observing the trajectory over a finite interval. This is a key hyperparameter that should be investigated within the framework of mathematical analysis.

The basic configuration proposes using several σ scales, frequencies from 80 to 4000 Hz, and a symmetrical set of chirp rates. The lower limit is determined by the typical range of the fundamental tone, and the upper limit covers a significant part of the formant structure of speech at 16 kHz sampling. The chirp-rate range is selected empirically for development. subset. For short frames, a practical initial set can be $k \in \{-4000, -2000, -1000, 0, 1000, 2000, 4000\}$ Hz/s with subsequent refinement. It is important to note that these are not “optimal” values, but a starting grid for a systematic search.

To avoid excessive computational complexity, the bank can be implemented via FFT convolution, pre-generated atoms, or GPU-batched inner products. An additional option is to first estimate the coarse local instantaneous frequency and then calculate chirplets only in a limited neighborhood of the most probable ones f and k . This coarse-to-fine approach is consistent with the idea of adaptive chirplet analysis and significantly reduces the number of operations.

TABLE II. RECOMMENDED SET OF COMPARED SYSTEMS

Model	Input	Training	Role in comparison
FFT harmonics	dominant frequencies/amplitudes	minimal	control of the original idea
MFCC + SVM	MFCC statistics	small data	classical baseline
Chirplet + SVM	chirp-rate statistics	small data	main method under study
ECAPA-TDNN	log-mel / embedding	pretrained model	strong modern baseline
Chirplet + ECAPA	fusion features/scores	small classifier	test of the added value of chirplets

Table II defines the required controlled comparisons: the six-feature historical method; fixed-capacity FFT-derived features; MFCC statistics with the same SVM classifier; frequency-only chirplet statistics; full chirplet statistics; pretrained ECAPA-TDNN and the proposed hybrid system. Fig. 1 summarizes the proposed pipeline. Fig. 2 illustrates why a sloping instantaneous-frequency component motivates chirp-rate analysis, but it is not experimental evidence of identification improvement.

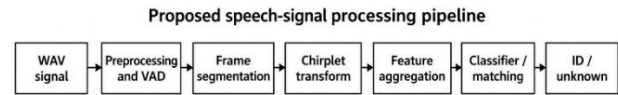


Fig. 1. General speaker-identification pipeline using chirplet features.

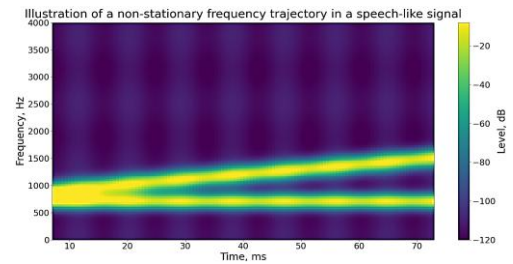


Fig. 2. Example of a non-stationary frequency trajectory.

To correctly test the hypothesis, it is necessary to include several baseline levels. The first level is simple features: FFT-dominant harmonics and MFCC. The second is a statistical classifier on the same number of training records: SVM or LDA for MFCC statistics and chirplet statistics. The third is a modern pretrained speaker embedding model, such as ECAPA-TDNN, with cosine scoring or PLDA. The fourth is a hybrid system where a standardized chirplet vector is combined with a speaker embedding before the final classifier or at the score level fusion. Table II lists the parameters for the models being compared.

The main requirement for a speaker Recognition – splitting by utterance and, if necessary, by session so that test recordings do not have technical duplicates or adjacent fragments in the training set. If a long recording is split into several segments, all segments of a single source file should fall into only one subset. Otherwise, the algorithm may recognize the channel, background, or acoustics of the recording instead of the individual.

For a local dataset, nested is recommended. Cross-validation or repeated stratified group split. The chirplet bank hyperparameters, SVM, and unknown threshold are selected in the inner loop, and the final metric is calculated in the outer test fold. At least five independent repetitions with fixed random values are recommended. seeds and publication of the mean, standard deviation, and 95% confidence interval.

For public independent verification, it is appropriate to use LibriSpeech as a controlled corpus of read speech and VoxCeleb as a more sophisticated “in-the-wild” benchmark. LibriSpeech contains approximately 1000 hours of English speech recorded at 16 kHz [12], and VoxCeleb/VoxCeleb 2 were specifically formed for speaker recognition in variable conditions [10,11]. It is important not to train the model on VoxCeleb test identities and to adhere to official protocols.

For closed-set identification, the primary metric is accuracy, but if the number of recordings across speakers is unbalanced, macro-F1 should also be included. The CMC-curve and Top-k accuracy are useful for ranking analysis. Open-set identification also requires calculating the false identification rate and false rejection rate at various thresholds. EER and minDCF are commonly used for verification, as they characterize the tradeoff

between false acceptance and false rejection and allow for comparison of systems under different operating conditions.

$$\text{Accuracy} = \frac{N_{\text{correct}}}{N_{\text{total}}} \quad (12)$$

$$F1_{\text{macro}} = \frac{1}{C} \sum_i \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \quad (13)$$

In addition to accuracy, it is necessary to record computational indicators: feature extraction time, vector size, number of classifier parameters, and real-time factor. This is especially important for justifying the use of chirplets in edge scenarios, where a large pretrained network may be impractical.

A non-stationary presentation is only meaningful if it exhibits an advantage in complex conditions. Therefore, testing should include: 1) additive white noise and, if possible, real household noise; 2) SNR of 20, 10, 5, and 0 dB; 3) artificial or measured reverberation with several RT 60; 4) short segments of 1, 2, and 3 s; 5) microphone and channel differences; 6) text-dependent and text-independent staging; 7) unknown speakers. For each distortion, the parameters are generated after the train/test split, so that the same noise realization does not end up in both parts.

Fig. 3 lists the predefined stress conditions. Apply additive noise at 20, 10, 5, and 0 dB SNR, room reverberation over published RT60 values, truncated 1-, 2-, and 3-second speech, channel mismatch, and unknown speakers. Noise realizations and room responses must be assigned after partitioning to avoid train/test leakage. Each condition requires sufficient independent utterances and session-level confidence intervals; the existing pilot does not support these estimates.

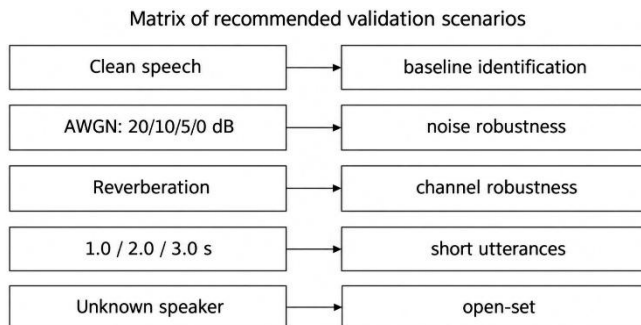


Fig. 3. Scenario matrix.

In [22-24], it was established that three dominant harmonics are sufficient for the scheme under consideration; after processing each recording, six characteristics were formed – three frequency and three amplitude coefficients. For 50 speaker recordings, the identification accuracy was in the range of 73-75%. An additional recording of a speaker not included in the database differed from the existing samples, confirming the fundamental feasibility of an open-source solution.

Table III summarizes the interpretation and evidential limits of the pilot result. From a practical standpoint, a value of 73-75% is not sufficient for a complete biometric system. However, its scientific significance lies elsewhere: a very low-dimensional vector of six features already contains noticeable speaker discriminative information. Therefore, the space of chirplet coefficients deserves a more comprehensive study, where

information about the chirp rate is not reduced to a few maxima, but is aggregated statistically across the entire record.

TABLE III. INTERPRETATION OF THE PILOT RESULTS

Indicator	Pilot result	Interpretation
Number of speakers	50	sufficient for proof of concept, insufficient for a benchmark
Number of features	6	very compact representation
Accuracy	73–75%	presence of discriminative information, but not production-level
Unknown speaker	separated from the database	supports the open-set formulation
Strong baselines	not reported	does not permit claims of superiority
Confidence intervals	not reported	requires repeated statistical evaluation

The published paper's experimental plots page shows the original and cleaned signals, the chirplet representation, and the amplitude spectrum. The plots demonstrate that after filtering, the structure of speech pulses is preserved, and the chirplet representation produces distinct energy regions suitable for further feature extraction. These visual results support the interpretation of the method as a locally consistent analysis, but do not, by themselves, replace quantitative evaluation using an independent test set.

Selecting the three strongest harmonics makes the algorithm computationally simple, but it introduces several risks. First, the order of the dominant components can change due to phonetic content: different numbers contain different vowels and consonants, meaning spectral maxima don't always correspond to the same acoustic structures. Second, noise can alter the local harmonic ranking. Third, such a vector makes virtually no use of the main advantage of chirplets – the explicit description of the rate of frequency change.

IV. DISCUSSION

The scientific significance of the obtained result is primarily related not to the absolute accuracy value, but to the specifics of the feature representation used. The chirplet transform is designed to describe locally non-stationary signal components and, unlike stationary harmonic analysis, contains a chirp-rate parameter that directly characterizes the direction and rate of change of the instantaneous frequency. Consequently, the potential advantage of this approach lies in its ability to account for the dynamic properties of the speech signal, which are only indirectly represented in traditional spectral characteristics.

The speech signal contains pronounced temporal-frequency structures, including pitch changes, formant transitions, short-term modulations, and transition regions between phonemes. For such components, matching the analysis atom to the local frequency trajectory can provide a more concentrated representation of energy compared to a stationary sinusoidal atom. The theoretical results presented in the theoretical analysis formalize this effect and show that in the presence of local linear frequency modulation, taking the chirp-rate into account can reduce the mismatch between the signal and the analysis function.

However, the experimental data from this study do not yet allow us to directly assert that the chirp-rate coordinate was responsible for the observed identification accuracy. In the original pilot representation, the transformation information was reduced to a few dominant components, resulting in only partial utilization of the main advantage of the full chirplet parameterization. Furthermore, the position of the strongest components may depend on the phonetic content and acoustic noise level.

The central empirical test is an incremental-feature comparison: estimate per-feature and aggregated between/within-speaker separability for frequency-only and chirp-rate descriptors, then compare identification and verification on the same held-out trials.

The main argument in favor of chirplets is not statistical but structural. Speech contains dynamic spectral elements, and a chirplet atom may be better aligned with them than a stationary sine wave. If this alignment is truly related to individual speaker characteristics, then the energy distribution by chirp rate should increase interspeaker distance or decrease intraspeaker variability. This is precisely what needs to be verified using the F-ratio, Fisher criterion, mutual information, and downstream accuracy.

The second potential strength is interpretability. In deep embedding, a single vector coordinate typically doesn't have a simple physical meaning. In chirplet space, the central frequency, scale, and direction of frequency variation can be interpreted. For biometric systems used in access control or forensics, such explainability can be an additional advantage, although it doesn't replace the requirements for accuracy and spoofing protection.

Computational efficiency is a limitation: a directly implemented chirplet dictionary has more degrees of freedom than an FFT. With T frames, F frequencies, K rates, S scales, and W window samples, naive arithmetic cost scales as $O(TFKSW)$. Batched FFT convolution, sparse frequency neighborhoods, and on-the-fly pooled statistics may reduce practical time and memory.

For edge devices, it is necessary to compare not only accuracy, but also real-time factor and power consumption. If a compact chirplet bank and SVM achieve acceptable quality without a GPU, such a tradeoff may be practical even at the expense of a large ECAPA-TDNN in absolute accuracy. Conversely, if the computational cost of chirplets approaches the cost of a deep model, then the rationale should be based on increasing stability, interpretability, or fusion gain.

Speaker identification is vulnerable to replay attacks, synthetic speech, and voice conversion. Chirplet-derived identity scores do not constitute presentation-attack detection.

At the same time, it can be assumed that the local dynamics of the chirp rate contain synthesis artifacts, and this area deserves separate study. However, until a dedicated ASVspoof experiment is conducted, it cannot be concluded that chirplets improve security. In this study, this possibility is considered only as a promising hypothesis.

The research conducted in this work is limited by the following circumstances.

The first limitation is related to the size and structure of the original local dataset. Fifty recordings allow us to demonstrate the principle's effectiveness, but they do not cover intra-speaker variability across days, microphones, emotional states, and phonetic content. For a comprehensive biometric assessment, each speaker requires several independent sessions.

A second limitation is the unavailability of original trial-level decisions and sufficiently detailed repeated speaker recordings in the supplied pilot results. Confidence intervals, independent test accuracy, ROC/DET curves, EER, and calibrated unknown rejection cannot be validly reconstructed from the reported 73–75% aggregate.

A third limitation is the lack of a direct controlled comparison with MFCC, wavelet/AM–FM alternatives, x-vectors, or ECAPA-TDNN.

The fourth limitation is possible text dependence. The pilot recordings contain the same sequence of numbers, meaning the production is close to text-dependent identification. To generalize to free speech, a text-independent test is required, where the test phrases do not match the enrollment.

V. CONCLUSION

This study examines the applicability of the chirplet transform to the problem of automatic speaker identification, focusing on the analysis of the locally non-stationary structure of the speech signal. Unlike traditional spectral representations, the approach under consideration incorporates the chirp-rate parameter, which allows for the direct accounting of the rate of change of instantaneous frequency and thereby characterizing the dynamic temporal-frequency components of speech.

The theoretical part of the study allowed us to formalize the key properties of the normalized Gaussian chirplet representation. Specifically, the Lipschitz stability of individual coefficients to additive perturbations was established, covariance with respect to quadratic phase modulation was demonstrated, and the property of maximum energy concentration was proven when matching the parameters of the analyzing atom with the local linear frequency-modulated component of the signal. The resulting relations show that, for a nonzero rate of change of instantaneous frequency, the constraint $k=0$, characteristic of the local Fourier/Gabor representation, leads to a decrease in the degree of matching, whereas a chirplet with the corresponding chirp-rate value maintains maximum response.

The only quantitative speaker-identification result discussed is the previously reported pilot accuracy of approximately 73–75% for 50 recordings using three frequencies and three amplitudes [22].

The coefficient results demonstrate mathematical localization and mismatch behavior, not proven improvements over speaker-recognition baselines. The contribution of the present revision is a reproducible descriptor specification, explicit discrimination hypotheses, and calibrated protocols for comparison, robustness, computation, and score fusion. The

proposed development of the method involves moving from the selection of individual dominant components to a multidimensional feature representation that aggregates the energy distribution by frequency and chirp-rate. This approach allows for more comprehensive use of information about the local frequency dynamics of the speech signal and provides the basis for subsequent analysis of whether chirp-rate provides additional inter-speaker distinctiveness compared to classical spectral and cepstral characteristics.

Thus, this study demonstrates that the primary scientific value of the chirplet approach lies not in its direct replacement of modern neural network systems, but in its ability to generate an additional, interpretable time-frequency representation of the speech signal. The further value of this approach should be determined through rigorous comparative validation and quantitative assessment of the chirp-rate's contribution to the accuracy, robustness and interpretability of voice biometric systems.

REFERENCES

- [1] T. Kinnunen, H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52(1), 2010, pp.12–40, doi:10.1016/j.specom.2009.08.009.
- [2] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol.19(4), 2011, pp.788–798, doi:10.1109/TASL.2010.2064307.
- [3] E. Variani, X. Lei, E. McDermott, I. Lopez Moreno, J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," *Proc. ICASSP*, 2014, doi:10.1109/ICASSP.2014.6854363.
- [4] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," *Proc. ICASSP*, 2018, pp.5329–5333, doi:10.1109/ICASSP.2018.8461375.
- [5] B. Desplanques, J. Thienpondt, K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation," *TDNN Based Speaker Verification*, Interspeech, 2020, pp.3830–3834, doi:10.21437/Interspeech.2020-2650.
- [6] H. Wang, S. Zheng, Y. Chen, L. Cheng, Q. Chen, "CAM++: A Fast and Efficient Network for Speaker Verification Using Context-Aware Masking," *Interspeech*, 2023, pp.5301–5305, doi:10.21437/Interspeech.2023-1513.
- [7] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Yanmin. Qian, Yao. Qian, J. Wu, M. Zeng, X. Yu, F. Wei, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16(6), pp. 1505-1518, Oct. 2022, doi: 10.1109/JSTSP.2022.3188113.
- [8] Z. Bai, X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, 2021, vol.140, pp.65–99, doi:10.1016/j.neunet.2021.03.004.
- [9] R. Sharma, D. Govind, J. Mishra, A.K. Dubey, K.T. Deepak, "Prasanna SRM Milestones in speaker recognition," *Artificial Intelligence Review*, 2024, pp.57-58, doi:10.1007/s10462-023-10688-w.
- [10] A. Nagrani, J.S. Chung, A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," *Interspeech*, 2017.
- [11] J.S. Chung, A. Nagrani, A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," *Interspeech*, 2018, pp.1086–1090, doi:10.21437/Interspeech.2018-1929.
- [12] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, "LibriSpeech: An ASR Corpus Based on Public Domain Audio Books," *Proc. ICASSP*, 2015, pp.5206–5210, doi:10.1109/ICASSP.2015.7178964.
- [13] S. Mann, S. Haykin, "The chirplet transform: physical considerations," *IEEE Transactions on Signal Processing*, 1995, vol.43(11), pp.2745–2761, doi:10.1109/78.482123.
- [14] D. Mihovilovic, R.N. Bracewell, "Adaptive chirple representation of signals on time-frequency plane," *Electronics Letters*, 1991, vol.27(13), pp.1159–1161, doi:10.1049/el:19910723.
- [15] Z.K. Peng, G. Meng, F.L. Chu, Z.Q. Lang, W.M. Zhang, Y. Yang, "Polynomial Chirplet Transform With Application to Instantaneous Frequency Estimation," *IEEE Transactions on Instrumentation and Measurement*, 2011, vol.60(9), pp.3222–3229, doi:10.1109/TIM.2011.2124770.
- [16] G. Yu, Y. Zhou, "General linear chirplet transform," *Mechanical Systems and Signal Processing*, 2016, vol.70–71, pp.958–973, doi:10.1016/j.ymssp.2015.09.004.
- [17] H.D. Do, D.T. Chau, S.T. Tran, "Speech feature extraction using linear Chirplet transform and its applications," *Journal of Information and Telecommunication*, 2023, vol.7(3), pp.376–391, doi:10.1080/24751839.2023.2207267.
- [18] H. Do-Duc, D. Chau-Thanh, S. Tran-Thai, "A New Algorithm for Speech Feature Extraction Using Polynomial Chirplet Transform," *Circuits, Systems, and Signal Processing*, 2024, vol.43, pp.2320–2340, doi:10.1007/s00034-023-02561-6.
- [19] S.P. Mishra, P. Warule, S. Deb, "Chirplet transform based time frequency analysis of speech signal for automated speech emotion recognition," *Speech Communication*, 2023, vol.155, pp.102986, doi:10.1016/j.specom.2023.102986.
- [20] Z. Ge, X. Xu, H. Guo, T. Wang, Z. Yang, "Speaker recognition using isomorphic graph attention network based pooling on self-supervised representation," *Applied Acoustics*, 2024, vol.219, pp.109929, doi:10.1016/j.apacoust.2024.109929.
- [21] F. Liu, C. Ye, C. Peng, X. Geng, M. Jiang, L. Zhang, F. Zhang, "Residual-optimized general linear chirplet transform: A method for time-frequency feature extraction," *Signal Processing*, 2025, vol.234, pp.110006, doi:10.1016/j.sigpro.2025.110006.
- [22] M.S. Aliaskar, Sh.A. Jomartova, A.T. Mazakova, T. Zh. Mazakov, A.D. Mailbybaeva, N.T. Isimov, K.B. Begalieva, A.T. Dosanalieva, "Using chirlets to identify a person from audio recordings," *KazUTB Bulletin*, 2025, vol.2(27), pp.70–78, doi:10.58805/kazutb.v.2.27-1027.
- [23] M. Aliaskar, T. Mazakov, A. Mazakova, S. Jomartova, T. Shormanov, "Human voice identification based on the detection of fundamental harmonics," *IEEE 7th International Energy Conference (ENERGYCON)*, 2022, pp.1–4, doi:10.1109/ENERGYCON53164.2022.9830471.
- [24] T.S. Shormanov, A.T. Mazakova, M.S. Aliaskar, Sh.A. Dzhomartova, T. Zh. Mazakov, "Application of neural networks for voice identification taking into account the Kazakh accent," *Vestnik KazUTB*, 2024, vol.4(25), pp.60–68, doi:10.58805/kazutb.v.4.25-720.