



# A Chinese Heart Failure Status Speech Database with Universal and Personalised Classification

Yue Pan<sup>1</sup>, Liwei Liu<sup>2</sup>, Changxin Li<sup>2</sup>, Xingyao Wang<sup>3</sup>, Yili Xia<sup>1</sup>, Hanyue Zhang<sup>4</sup>, Ming Chu<sup>4</sup>

<sup>1</sup>School of Information Science and Technology, Southeast University, China

<sup>2</sup>Advanced Computing and Storage Laboratory, 2012 Laboratories, Huawei Technologies Co. Ltd.,

<sup>3</sup>Institute of High Performance Computing, A\*STAR, Singapore

<sup>4</sup>Taizhou School of Clinical Medicine, Nanjing Medical University, China

230228208@seu.edu.cn, liuliwei5@huawei.com, chuming@njmu.edu.cn

## Abstract

Speech is a cost-effective and non-intrusive data source for identifying acute and chronic heart failure (HF). However, there is a lack of research on whether Chinese syllables contain HF-related information, as observed in other well-studied languages. This study presents the first Chinese speech database of HF patients, featuring paired recordings taken before and after hospitalisation. The findings confirm the effectiveness of the Chinese language in HF detection using both standard 'patient-wise' and personalised 'pair-wise' classification approaches, with the latter serving as an ideal speaker-decoupled baseline for future research. Statistical tests and classification results highlight individual differences as key contributors to inaccuracy. Additionally, an adaptive frequency filter (AFF) is proposed for frequency importance analysis. The data and demonstrations are published at [https://github.com/panyue1998/Voice\\_HF](https://github.com/panyue1998/Voice_HF).

**Index Terms:** Heart Failure, Machine Learning, Speech Abnormality Detection

## 1. Introduction

Heart failure (HF) is a progressive condition characterised by a decline in the heart's ability to pump blood effectively, affecting 26 million individuals worldwide [1]. Speech analysis offers a cost-effective and non-intrusive method for detecting HF-related laryngeal edema in its early stages [2], providing an alternative to conventional diagnostic techniques, such as X-ray, echocardiography, and angiography. Previous research has explored various speech tasks for HF detection, including vowels and word articulation [3], sentence reading [1, 3, 4], and short paragraph recitations [3, 5, 6]. These studies have been conducted in multiple languages, such as English [2, 3], Finnish [5, 7], and Portuguese [4], while some supported a mix of different languages [1].

Since HF-related changes in speech features are subtle and can be overshadowed by individual variations, a paired database that captures multiple conditions of the same patient is essential for extracting pathological information. One of the earliest studies in this direction, conducted by Murton et al. [3], analysed speech from ten hospitalised HF patients, revealing detectable trends in phonation and respiration parameters between admission (wet) and discharge (dry) conditions. Similarly, Amir et al. [1] observed vocal alterations between wet and dry states using a mobile application designed for speech analysis. Both studies highlighted intra-patient variability as a potential confounding factor. A subsequent study expanded on [3] by incorporating a larger patient cohort and additional biomarkers, demonstrating a positive correlation between speech features and discharge probability through logistic regression analysis

[2]. However, these studies did not systematically assess the impact of individual differences on classification accuracy, nor did they establish baselines to evaluate such effects.

To address these gaps, this study introduces a pair-wise classification approach alongside the standard patient-wise classification method, leveraging a newly constructed, large-scale Chinese paired speech database. A summary of related studies is shown in Table 1. Our dataset is among the most extensive, particularly in terms of paired speech samples collected before and after medical intervention. While no public dataset is currently available, we intend to release high-level feature data while preserving patient privacy. Previous studies primarily focused on Indo-European and Afroasiatic languages, largely neglecting Sino-Tibetan languages. Although HF detection is generally considered content-independent, linguistic characteristics—such as those specific to Chinese [8, 9]—may influence pathological speech markers. Consequently, features identified in one language may not necessarily apply to another.

Furthermore, prior studies reported varied classification accuracies without thoroughly analysing the sources of model errors. This work hypothesises that individual differences significantly contribute to classification inaccuracies, a claim supported by statistical tests and our proposed pair-wise classification framework, which also serves as a robust baseline for future research. Additionally, we introduce an adaptive frequency filter (AFF) for frequency importance analysis in time-frequency sequential models. The key contributions of this work include:

- Development of the first Chinese speech database for HF detection with paired samples, achieving high classification performance.
- Introduction of a 'pair-wise' classification approach as a speaker-independent baseline, identifying individual variations as a primary source of classification inaccuracy.
- Design of an AFF for frequency importance analysis.

In Section 2, the proposed methods and experimental setup are detailed. The results and discussion are presented in Section 3, while the conclusion and suggestions for future work are provided in Section 4. A summary of related studies is illustrated in Figure 1.

## 2. Methods

### 2.1. Data acquisition

This study involves a total of 127 patients from our partner hospital, admitted for acute HF treatments. The recordings were collected using a standard smartphone, handheld by the staff, at a sampling rate of 22,050 Hz. Each patient participated in two data collection sessions, one before and after hospitalisation. The male-to-female ratio is 0.61:0.39, with an average age

Study	Year	Language	Patients	Data Collection	Data Made Public	Best Result (Accuracy)
[3]	2017	English	10	Paired	No	-
[5]	2021	Finnish	45	Single	No	81.5
[7]	2022	Finnish	45	Single	No	-
[1]	2022	Hebrew, Arabic, Russian	40	Paired	No	-
[6]	2022	English	74	Single	No	93.7
[4]	2023	Portuguese	142	Single	No	91.9
[2]	2023	English	52	Multiple	No	69.0
Ours		Chinese	127	Paired	High-level features/ Full data by request	See results

Table 1: Summary of similar studies

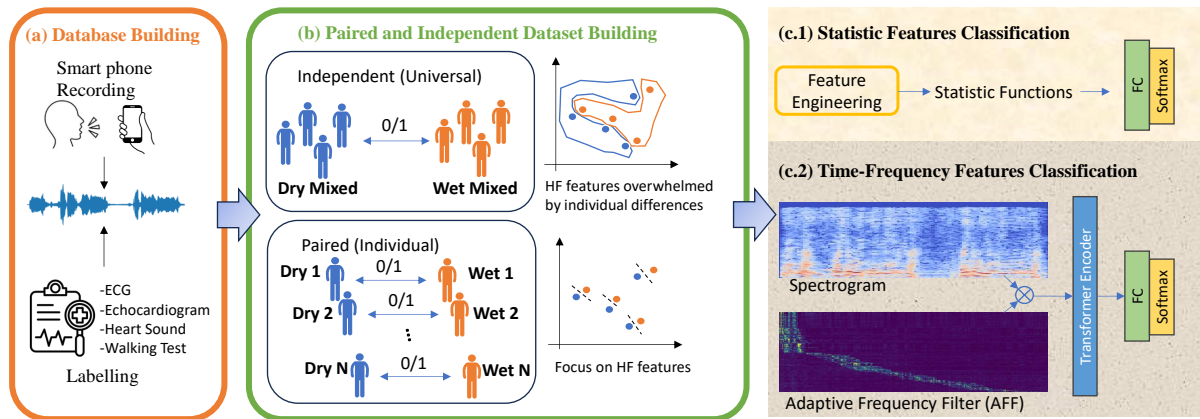


Figure 1: Overall structure of the project.

of 68 and a standard deviation of 13.

Medical professionals assessed the patients’ conditions based on the New York Heart Association (NYHA) functional classification for HF before and after hospitalisation, considering clinical tests, such as electrocardiogram, echocardiography, and walking tests [10]. Patients who exhibited improvement in NYHA levels were considered relevant for the study.

Participants were required to complete four speech tasks, categorised into short and long sentences. In Chinese, consonants exhibit the discrimination between voiced and unvoiced sounds. These are further classified into fully voiced (r), partially voiced (m, n, l, j, w), fully unvoiced (b, d, g, s, x, z), and partially unvoiced (p, t, k, c, q) [9]. Three short sentences were selected to capture both voiced and unvoiced sounds in Chinese. For the long sentence task, participants were asked to count from 1 to 60 in Chinese, ensuring the inclusion of both voiced and unvoiced sounds. Vowels (a, i, u) were present in all four tasks. Details of these tasks are presented in Table 2.

Every recording was manually labelled according to its respective task. Unrelated, incomplete, and erroneous samples, as well as those from patients with other underlying conditions affecting speech, were excluded, leaving a total of 117 patients. The final number of recordings for each task after data cleansing is shown in Table 3.

## 2.2. Feature extraction and selection

This study performed feature extraction using openSMILE [11], version 2.5.0. Two feature sets were utilised: GeMAPS [12] and ComParE 2016 [13]. To identify the most relevant features associated with hospitalisation conditions, paired and independent t-tests were conducted between admission and discharge groups. Table 4 presents the number of selected features where  $p \leq 0.05$  for each task. This process was repeated for each task and both feature sets, considering female, male, and combined groups separately. The selected feature data were then used to train a three-layer fully connected neural network for HF detection, as illustrated in part c.1 in Figure 1.

## 2.3. Pair-wise classification

In this work, we propose the ‘pair-wise’ classification scheme as a baseline for the ideal speaker-decoupled case. In the pair-wise scenario, we aim to simultaneously feed both the wet and dry data points of a single patient into the classifier, which should then determine which is wet and which is dry. In other words, there is an additional by-patient normalisation compared with the raw data. The commonality with the standard scheme is that the training and testing sets are divided according to patient ID, ensuring that the data points in the test set are from patients previously unseen by the classifier, so they are both speaker-independent.

Consider the selected feature vectors of a given patient as  $A^i = [a_1^i, a_2^i, \dots, a_n^i]$  (wet) and  $B^i = [b_1^i, b_2^i, \dots, b_n^i]$  (dry),

task	abbr.	Chinese Pingyin (phonetic symbols)	Consonants	Consonants Type	Average Length of Recordings
Short Sentence 1	pg	shan dong de ping guo you da you tian	s, d, p, g, t	unvoiced	5.1s
Short Sentence 2	mm	ni you yi ge mei li de mei mei	m, n, l	voiced	4.6s
Short Sentence 3	mlh	hao yi duo mei li de mo li hua	m, l	voiced	4.7s
Long Sentence	c	(numbers 1-60)	r, l, j, b, q	both	27.2s

Table 2: *Speech tasks conducted.*

	Tasks	pg	mm	mlh	c
Training (Individuals)	male	46	44	38	47
	female	28	21	23	30
	total	74	65	61	77
Testing (Individuals)	male	22	17	19	20
	female	11	5	6	11
	total	33	22	25	31

Table 3: *Number of individuals used for training and testing in each task. Task name abbreviations are provided in Table 2.*

Feature Set	Sex	T-test	Tasks			
			pg	mm	mlh	c
A	male	ind	2	7	4	3
		pair	8	12	5	6
	female	ind	5	7	2	3
		pair	15	9	11	12
	all	ind	11	10	5	7
		pair	23	20	16	15
B	male	ind	248	241	187	467
		pair	392	378	352	914
	female	ind	332	327	311	467
		pair	562	461	440	996
	all	ind	326	287	212	901
		pair	618	490	434	1483

Table 4: *Number of features selected by paired (pair) and independent (ind) t-tests, where  $p \leq 0.05$ . A: ComParE\_2016, B: eGeMAPSv02. Task name abbreviations are provided in Table 2.*

where  $i$  is the patient ID and  $n$  is the feature ID. For the pair-wise scheme, we first randomly generate a 0/1 label. If the label is 1, a combined vector is formed where the 'wet' vector is subtracted from the 'dry' one, and vice versa:

$$X_{train/test}^i = \begin{cases} A^i - B^i & \text{for label} = 1 \\ B^i - A^i & \text{for label} = 0 \end{cases} \quad (1a)$$

$$(1b)$$

In this way, the result is a comparison within a single patient's data, making it immune to the domain difference caused by inherent individual variations in speech voice. While this scheme may not be as practical as the standard patient-wise scheme in real-world applications—since it would require a known normal state for a given individual—it serves as a useful baseline reference in cases where individual differences are decoupled from pathological features.

The standard patient-wise scheme follows a typical classification approach, where the training and test sets contain a mixture of 'wet' and 'dry' vectors, each treated as a standalone data point. The standard scheme was conducted separately for female, male, and combined groups.

## 2.4. Adaptive frequency filter (AFF)

The AFF (part c.2 in Figure 1) is primarily designed for the visualisation of frequency importance. This technique is inspired by [14]. While [14] operates in the time domain, our AFF is directly applied in the frequency domain. The AFF is a trainable linear projection matrix applied to the time-frequency data before sequential encoding, and in this case, transformer encoders. It has a dimension of  $(d_{freq}, d_{new})$ , where  $d_{freq}$  represents the input dimension of the frequency axis, and  $d_{new}$  is a user-defined target dimension. It is applied to the spectrogram (S) as:

$$\mathbf{S}(T, d_{freq}) \times \mathbf{AFF}(d_{freq}, d_{new}) = \mathbf{F}(T, d_{new}) \quad (2)$$

To obtain the filtered feature map (F), the AFF will have  $d_{new}$  filters, with each filter having trainable attention across all  $d_{freq}$  dimensions. The trainable AFF is initialised as an MFCC filter bank. To ensure that each filter focuses on a relevant frequency band and converges towards a Butterworth-style filter bank, we trim the high attention values outside a specific frequency range around the current highest position after every few epochs.

## 3. Results and discussion

Table 5 presents the patient-wise and pair-wise classification results, as outlined in part c.1 of Figure 1. We report the F1 score, which is calculated as follows:

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (3)$$

Overall, the ComParE\_2016 feature set outperformed eGeMAPSv02, though it comes with a considerably larger feature size. The highest performance was observed in the pair-wise scheme of the 'mm' task using the ComParE\_2016 feature set, achieving an F1 score of 0.964. The pair-wise scheme generally outperformed the patient-wise scheme across nearly all settings, including the average score. This suggests that personal differences substantially influence model accuracy, as expected, with the pair-wise scheme remaining unaffected by inter-patient variations due to its focus on intra-patient comparisons. This assumption is further supported by statistical tests. From Table 4, paired t-tests consistently identified more significant features than independent t-tests, indicating that while there are changes before and after hospitalisation, these differences are smaller than the inherent variations between patients. As a result, the overall difference between HF and normal groups becomes less pronounced.

Although previous studies commonly train separate models for male and female groups, our results show that this approach slightly improves performance (about +4% on average compared with the combined group). However, it cannot completely eliminate the influence of personal differences, as evidenced by the remaining 12% difference compared to the pair-wise baseline.

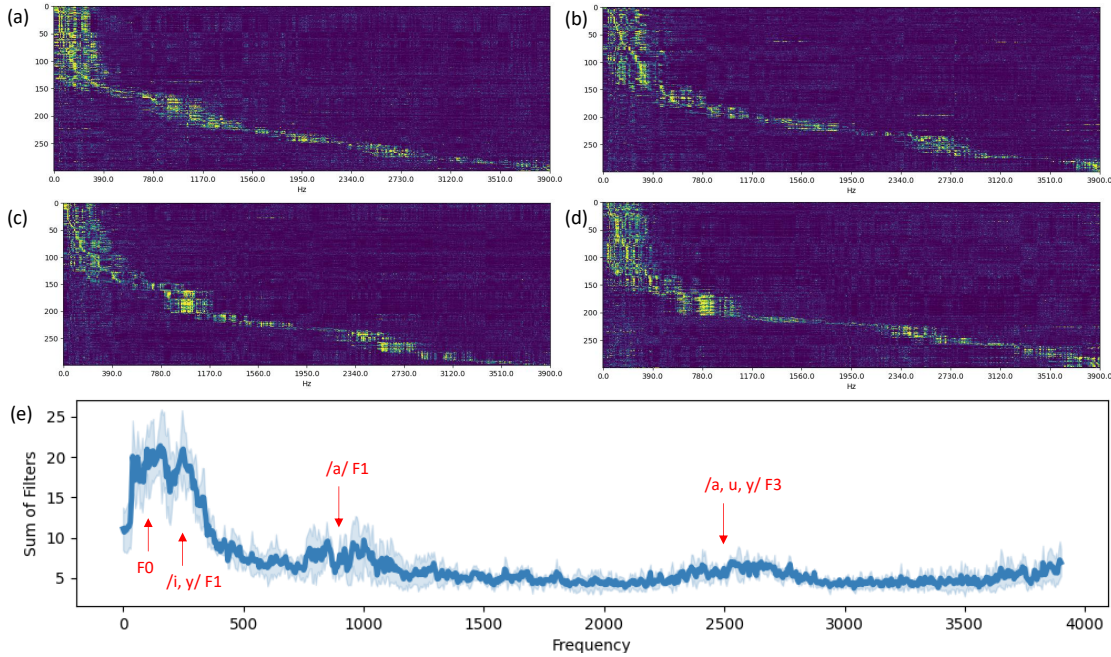


Figure 2: AFF frequency analysis for the tasks: *pg*(a), *mm*(b), *mlh*(c), *c*(d); and the sum of frequency dimensions, average, and standard deviation across the four tasks (e). The positions of the fundamental frequency (F0) and the resonant frequencies of vowels are marked in (e). For task name abbreviations, refer to Table 2.

The frequency analysis of the AFF (part c.2 in Figure 1) is illustrated in Figure 2. Several significant frequency areas

Task	Feature Set	Feature Selection	F1 (%)			Pair-wise
			Female	Male	All	
<b>pg</b>	A	ind	68.1	54.2	54.5	<b>68.6</b>
		pair	59.0	59.0	56.0	<b>61.6</b>
	B	ind	<b>90.9</b>	86.4	72.7	89.6
		pair	<b>90.8</b>	79.5	68.1	88.6
<b>mm</b>	A	ind	58.3	58.7	56.8	<b>68.8</b>
		pair	49.5	55.8	47.7	<b>76.3</b>
	B	ind	89.9	85.2	68.1	<b>96.4</b>
		pair	89.9	67.6	61.2	<b>96.4</b>
<b>mlh</b>	A	ind	48.5	57.8	61.3	<b>71.3</b>
		pair	41.2	59.8	53.8	<b>73.0</b>
	B	ind	91.6	78.9	76.0	<b>91.7</b>
		pair	74.8	76.3	71.8	<b>91.2</b>
<b>c</b>	A	ind	54.2	60.0	62.7	<b>67.3</b>
		pair	54.5	55.0	64.5	<b>61.0</b>
	B	ind	68.1	80.0	74.2	<b>95.4</b>
		pair	72.7	77.5	74.1	<b>96.5</b>
<b>Average</b>			68.9	68.2	64.0	<b>80.9</b>

Table 5: Classification results of fully connected classifiers. A: ComParE\_2016, B: eGeMAPSv02. Task name abbreviations are provided in Table 2.

are identified, with the low-frequency range below 250 Hz corresponding to the fundamental frequency (F0). The AFF also captured several vowel formants, such as /a/ (F1 ~ 800Hz), /i, y/ (F1 ~ 300Hz) and /u/ [8]. Notably, in Figure 2(b), the 800 Hz area is less pronounced for the ‘mm’ task, likely due to the absence of the vowel /a/.

## 4. Conclusion

This study presents the first large-scale Chinese paired speech dataset for HF detection. A ‘pair-wise’ classification scheme, decoupling personal differences, was proposed as an ideal baseline. Two feature sets were employed for feature extraction, and fully connected models were used for classification. The classification results highlight personal differences as a major factor affecting model accuracy, a challenge that cannot be fully addressed by the common practice of training separate models for male and female groups. The frequency analysis using the AFF identified several vowel formants as significant in HF detection.

Future work could improve upon this study in several ways. First, although the ‘pair-wise’ scheme serves as an ideal reference for speaker-irrelevant cases, it may not be practical in real-world applications, as it requires a known normal state, which is not always available. Further improvements in model design would be necessary to reduce the accuracy gap between the standard and pair-wise schemes. Second, while the frequency analysis captured important vowel formants, further refinement is needed to more accurately identify resonant frequencies and their exact positions.

## 5. Ethics

The study adhered to the guidelines of the Declaration of Helsinki and was approved by the Taizhou People’s Hospital (approval number KY 2023-073-01, obtained on 7 June 2023).

## 6. References

- [1] O. Amir, W. T. Abraham, Z. S. Azzam, G. Berger, S. D. Anker, S. P. Pinney, D. Burkhoff, I. D. Shalom, C. Lotan, and E. R. Edelman, "Remote speech analysis in the evaluation of hospitalized patients with acute decompensated heart failure," *Heart Failure*, vol. 10, no. 1, pp. 41–49, 2022.
- [2] O. M. Murton, G. W. Dec, R. E. Hillman, M. D. Majmudar, J. Steiner, J. V. Guttag, and D. D. Mehta, "Acoustic voice and speech biomarkers of treatment status during hospitalization for acute decompensated heart failure," *Applied Sciences*, vol. 13, no. 3, p. 1827, 2023.
- [3] O. M. Murton, R. E. Hillman, D. D. Mehta, M. Semigran, M. Daher, T. Cunningham, K. Verkouw, S. Tabtabai, J. Steiner, G. W. Dec *et al.*, "Acoustic speech analysis of patients with decompensated heart failure: a pilot study," *The Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. EL401–EL407, 2017.
- [4] J. V. Firmino, M. Melo, V. Salemi, K. Bringel, D. Leone, R. Pereira, and M. Rodrigues, "Heart failure recognition using human voice analysis and artificial intelligence," *Evolutionary Intelligence*, pp. 1–13, 2023.
- [5] M. K. Reddy, P. Helkkula, Y. M. Keerthana, K. Kaitue, M. Minkinen, H. Tolppanen, T. Nieminen, and P. Alku, "The automatic detection of heart failure using speech signals," *Computer Speech & Language*, vol. 69, p. 101205, 2021.
- [6] D. Priyasad, A. Partovi, S. Sridharan, M. Kashefpoor, T. Fernando, S. Denman, C. Fookes, J. Tang, and D. Kaye, "Detecting heart failure through voice analysis using self-supervised mode-based memory fusion," in *Proceedings of the 23rd INTERSPEECH Conference*. International Speech Communication Association, 2022, pp. 2848–2852.
- [7] K. R. Mittapalle, H. Pohjalainen, P. Helkkula, K. Kaitue, M. Minkinen, H. Tolppanen, T. Nieminen, and P. Alku, "Glottal flow characteristics in vowels produced by speakers with heart failure," *Speech Communication*, vol. 137, pp. 35–43, 2022.
- [8] H. Liu and M. L. Ng, "Formant characteristics of vowels produced by mandarin esophageal speakers," *Journal of voice*, vol. 23, no. 2, pp. 255–260, 2009.
- [9] L. Li, "Research and implementation of parkinson disease recognition system based on speech recognition," Master's thesis, Chongqing University, 2018.
- [10] S. Giannitsi, M. Bougiakli, A. Bechlioulis, A. Kotsia, L. K. Michalis, and K. K. Naka, "6-minute walking test: a useful tool in the management of heart failure patients," *Therapeutic advances in cardiovascular disease*, vol. 13, 2019.
- [11] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.
- [12] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.
- [13] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *17TH Annual Conference of the International Speech Communication Association (Interspeech 2016)*, Vols 1-5, vol. 8. ISCA, 2016, pp. 2001–2005.
- [14] W. Yang, J. Liu, P. Cao, R. Zhu, Y. Wang, J. K. Liu, F. Wang, and X. Zhang, "Attention guided learnable time-domain filterbanks for speech depression detection," *Neural Networks*, vol. 165, pp. 135–149, 2023.