一个具有通用和个性化分类的中国心力衰竭状态语音数据库

Yue Pan[1], *Liwei Liu*[2], *Changxin Li*[2], *Xingyao Wang*[3], *Yili Xia*[1], *Hanyue Zhang*[4], *Ming Chu*[4]

[1]School of Information Science and Technology, Southeast University, China

[2]Advanced Computing and Storage Laboratory, 2012 Laboratories, Huawei Technologies Co. Ltd.,

[3]Institute of High Performance Computing, A*STAR, Singapore

[4]Taizhou School of Clinial Medicine, Nanjing Medical University, China

23022828@seu.edu.cn, liulwei5@huawei.com, chuming@njmu.edu.cn

Abstract

语音是一种成本效益高且非侵入性的数据来源，用于识别急性心力衰竭（HF）和慢性心力衰竭。然而，在汉语音节是否包含如其他研究较为充分的语言中所观察到的心力衰竭相关的信息方面，缺乏相关的研究。本研究介绍了首个心力衰竭患者的中文语音数据库，其中包括住院前后配对的录音。研究结果证实了使用标准“患者为中心”和个性化的“成对”分类方法进行心力衰竭检测的有效性，后者作为未来研究的理想说话人解耦基线。统计测试和分类结果显示个体差异是导致不准确性的关键因素。此外，提出了一种自适应频率滤波器（AFF）用于频率重要性分析。数据和演示发布于 https://github.com/panyue1998/Voice_HF。

Index Terms: 心力衰竭，机器学习，语音异常检测

1. 介绍

心力衰竭（HF）是一种渐进性疾病，其特征是心脏泵血能力下降，影响全球 2600 万人 [1]。语音分析提供了一种成本低廉且非侵入性的方法，在早期阶段检测与 HF 相关的喉水肿 [2]，为传统的诊断技术（如 X 射线、超声心动图和血管造影）提供了替代方案。以往的研究探索了用于 HF 检测的各种语音任务，包括元音和单词发音 [3]、句子阅读 [1, 3, 4] 和短段落背诵 [3, 5, 6]。这些研究已经用多种语言进行，如英语 [2, 3]、芬兰语 [5, 7] 和葡萄牙语 [4]，而有些则支持混合使用不同的语言 [1]。

由于与 HF 相关的语音特征变化细微，可能被个体差异所掩盖，因此捕捉同一患者多种状况的配对数据库对于提取病理信息至关重要。在这方面最早的几项研究之一由 Murton 等人 [3] 进行，分析了 10 名住院 HF 患者的语音，在入院（湿状态）和出院（干状态）条件下的发声和呼吸参数中发现了可检测的趋势。类似地，Amir 等人 [1] 使用一款专为语音分析设计的移动应用程序观察到在湿状态和干状态之间的声带变化。这两项研究都强调了患者内变异作为一个潜在的混淆因素。随后的一项研究通过纳入更多的患者群体和额外的生物标志物扩展了 [3]，并通过逻辑回归分析 [2] 展示了语音特征与出院概率之间存在正相关关系。然而，这些研究并没有系统地评估个体差异对分类准确性的影响力，也没有建立基准来评估这种影响。

为了解决这些差距，本研究引入了一种成对分类方法，同时使用标准的患者分类方法，并利用了一个新构建的大规模中文配对语音数据库。相关研究的总结见表 1。我们的数据集是最全面的之一，特别是在收集医疗干预前后配对语音样本方面。尽管目前没有公开的数据集可用，我们打算在保护患者隐私的同时发布高级特征数据。以往的研究主要集中在印欧语系和闪含语系语言上，很大程度上忽视了汉藏语系语言。虽然 HF 检测通常被认为是内容无关的，但语言特性——比如中文 [8, 9] 的特有之处——可能会影响病理语音标记。因此，在一种语言中识别出的特征不一定适用于另一种语言。

此外，先前的研究报告了不同的分类准确率，但没有彻底分析模型错误的来源。本研究假设个体差异显著导致了分类不准确性，这一主张得到了统计测试的支持，并通过我们提出的配对分类框架进行验证，该框架也为未来研究提供了一个

稳健的基准。另外，我们引入了一种自适应频率滤波器（AFF），用于时频序列模型中的频率重要性分析。本工作的主要贡献包括：

- 开发了首个用于 HF 检测的中文语音数据库，并实现了高分类性能。
- 介绍了一种作为说话人独立基线的“成对”分类方法，将个体差异识别为分类不准确的主要来源。
- 频率重要性分析的 AFF 设计。

在第 2 节中，详细介绍了所提出的方法和实验设置。结果与讨论部分见第 3 节，而结论及未来工作建议则在第 4 节提供。相关研究的总结如图 1 所示。

2. 方法

2.1. 数据采集

本研究共涉及我们合作医院的 127 名患者，因急性心力衰竭治疗而入院。录音使用标准智能手机由工作人员手持录制，采样率为 22,050 赫兹。每位患者参加了两次数据收集会话，一次在住院前，另一次在住院后。男女比例为 0.61:0.39，平均年龄为 68 岁，标准差为 13。

医疗专业人员根据纽约心脏协会（NYHA）心力衰竭的功能分类，在住院前后评估了患者的状况，考虑了临床测试，如心电图、超声心动图和行走测试 [10]。在 NYHA 水平上显示出改善的患者被认为与研究相关。

参与者需要完成四个演讲任务，这些任务被分为短句和长句两类。在汉语中，辅音表现出有声与无声的区别。这些进一步分类为完全有声（r）、部分有声（m, n, l, j, w）、完全无声（b, d, g, s, x, z）以及部分无声（p, t, k, c, q）[9]。选择了三个短句来捕捉汉语中的有声和无声音素。对于长句任务，参与者被要求用中文从 1 数到 60，以确保包含有声和无声的音素。元音（a, i, u）出现在所有四个任务中。这些任务的详细信息见表 2。

每个录音都根据其相应的任务手动标注。无关的、不完整的和错误的样本，以及来自其他基础疾病影响言语的患者的样本被排除在外，共剩下 117 名患者。数据清洗后每项任务最终的录音数量如表 3 所示。

2.2. 特征提取与选择

本研究使用 openSMILE[11]，版本 2.5.0 进行特征提取。使用了两个特征集：GeMAPS[12] 和 ComParE 2016[13]。为了识别与住院条件相关的最相关特征，在入院组和出院组之间进行了配对和独立 t 检验。表 4 显示了每个任务中所选特征的数量 $p \leq 0.05$。此过程针对每个任务以及两个特征集重复进行，分别考虑女性、男性和合并组。所选特征数据随后被用于训练一个三层全连接神经网络以检测 HF，如图 1 中的 c.1 部分所示。

2.3. 成对分类

在本工作中，我们提出了“成对”分类方案作为理想情况下说话者无关的基线。在成对场景中，我们的目标是同时将单个患者的湿润和干燥数据点输入分类器，然后由分类器确定哪个是湿润的，哪个是干燥的。换句话说，与原始数据相比，有一个额外的按患者归一化处理。与标准方案的共同之处在于训练集和测试集根据患者 ID 划分，确保测试集中的数据点来自分类器之前未见过的患者，因此它们都是说话者无关的。

考虑给定患者的选定特征向量为 $A^i = [a_1^i, a_2^i, ..., a_n^i]$（湿）和 $B^i = [b_1^i, b_2^i, ..., b_n^i]$（干），其中 i 是患者 ID，n 是特征 ID。对于成对方案，我们首先随机生成一个 0/1 标签。如果标签为 1，则形成一个组合向量，其中’ 湿’ 向量从’ 干’ 向量中减去，反之亦然：

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

这样，结果是在单个患者的数据内进行比较，使其免受由于语音固有个体差异导致的领域差异的影响。虽然这个方案在实际应用中可能不如标准的基于患者的方案实用——因为它需要知道给定个体的正常状态——但在个体差异与病理特征解耦的情况下，它作为一个有用的基线参考。

标准的以患者为单位的方案遵循典型的分类

Study	Year	Language	Patients	数据 收集	数据 公开发布	最佳结果 (准确性)
[3]	2017	English	10	Paired	No	-
[5]	2021	Finnish	45	Single	No	81.5
[7]	2022	Finnish	45	Single	No	-
[1]	2022	希伯来语、阿拉伯语、俄语	40	Paired	No	-
[6]	2022	英语	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	高级特征/ 按请求提供全部数据	See results

表 1: 相似研究的总结

task	abbr.	中文 拼音 (拼音符号)	Consonants	辅音类型 类型	记录的平均长度
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

表 2: 演讲任务已完成。

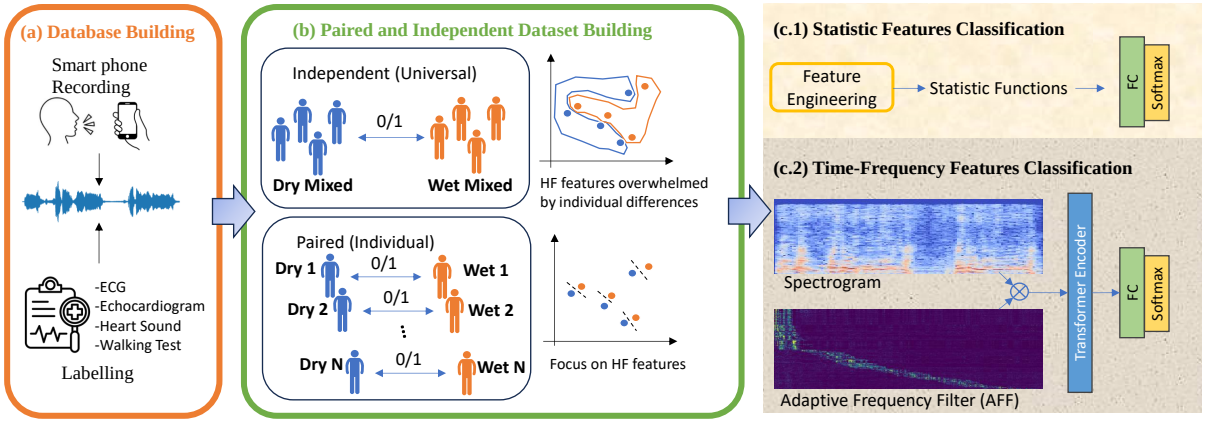

图 1: 项目总体结构。

	Tasks	页码	毫米	每隆小时	c
训练（个体）	male	46	44	38	47
	female	28	21	23	30
	total	74	65	61	77
测试（个体）	male	22	17	19	20
	female	11	5	6	11
	total	33	22	25	31

表 3: 用于训练和测试的个体数量在每个任务中。任务名称缩写见表 2。

方法，其中训练集和测试集包含“湿”和“干”向量的混合，每个都作为独立的数据点处理。标准方案分别对女性、男性和合并组进行了单独实施。

2.4. 自适应频率滤波器（AFF）

AFF（图 1 中的 c.2 部分）主要用于频率重要性的可视化。该技术受到 [14] 的启发。虽然 [14] 在时域中运行，但我们的 AFF 直接应用于频域。AFF 是一个可训练的线性投影矩阵，在顺序编码之前应用于时频数据，在这种情况下是变压器编码器。它的维度为 (d_{freq}, d_{new})，其中 d_{freq} 表示频率轴的输入维度，d_{new} 是用户定义的目标维度。它应用于谱图 (S) 如下：

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

特征集	性别	T 检验	任务			
			页码	毫米	每隆小时	c
A	男性	ind	2	7	4	3
		pair	8	12	5	6
	女性	ind	5	7	2	3
		pair	15	9	11	12
	全部	ind	11	10	5	7
		pair	23	20	16	15
B	男性	ind	248	241	187	467
		pair	392	378	352	914
	女性	ind	332	327	311	467
		pair	562	461	440	996
	全部	ind	326	287	212	901
		pair	618	490	434	1483

表 4: 所选特征的数量由配对（成对）和独立（独立）t 检验决定，其中 $p \leq 0.05$。*A*: *ComParE_2016*, *B*: *eGeMAPSv02*。任务名称缩写见表 2。

为了获得过滤后的特征图（F），AFF 将有 d_{new} 个滤波器，每个滤波器在所有 d_{freq} 维上具有可训练的注意力。可训练的 AFF 被初始化为 MFCC 滤波器组。为了确保每个滤波器专注于相关的频率带并收敛到 Butterworth 样式的滤波器组，我们在每次迭代后的特定频率范围外修剪当前最高位置之外的高注意力值。

3. 结果与讨论

表 5 展示了按患者和成对分类的结果，如图 1 的 c.1 部分所述。我们报告 F1 分数，其计算方法如下：

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

总体而言，ComParE_2016 特征集的表现优于 eGeMAPSv02，尽管其特征规模显著更大。在使用 ComParE_2016 特征集的“mm”任务配对方案中观察到了最高性能，F1 分数达到了 0.964。

任务	特征集	特征选择	F1 (%) Female	Male	All	
页码	A	指示	68.1	54.2	54.5	**68.6**
		对	59.0	59.0	56.0	**61.6**
	B	指数	**90.9**	86.4	72.7	89.6
		对	**90.8**	79.5	68.1	88.6
毫米	A	指标	58.3	58.7	56.8	**68.8**
		对儿	49.5	55.8	47.7	**76.3**
	B	指标	89.9	85.2	68.1	**96.4**
		对儿	89.9	67.6	61.2	**96.4**
每隆小时	A	指标	48.5	57.8	61.3	**71.3**
		对	41.2	59.8	53.8	**73.0**
	B	指数	91.6	78.9	76.0	**91.7**
		对	74.8	76.3	71.8	**91.2**
c	A	指数	54.2	60.0	62.7	**67.3**
		对	54.5	55.0	64.5	**61.0**
	B	指标	68.1	80.0	74.2	**95.4**
		对	72.7	77.5	74.1	**96.5**
平均值			68.9	68.2	64.0	**80.9**

表 5: 全连接分类器的分类结果。*A: ComParE_2016*, *B: eGeMAPSv02*。任务名称缩写见表 *2*。

一般来说，配对方案的整体表现优于患者方案，在几乎所有设置下都是如此，包括平均分数。这表明个人差异显著影响模型准确性，正如预期的那样，由于配对方案专注于患者内部比较，因此不受患者间变异的影响。这一假设得到了统计检验的支持。从表 4 中可以看出，配对 t 检验一致识别出比独立 t 检验更多的显著特征，这表明尽管住院前后存在变化，但这些差异小于患者之间的固有变异。因此，HF 组与正常组之间的总体差异变得不那么明显。

尽管之前的研究所通常为男性和女性群体分别训练模型，但我们的结果显示这种方法仅能略微提高性能（平均而言比混合组提高了约 4%）。然而，它并不能完全消除个人差异的影响，这可以从与配对基准相比仍存在的 12%差异中看出。

AFF 的频率分析（图 1 的 c.2 部分）如图 2 所示。识别出几个重要的频率区域，其中低于 250 Hz 的低频范围对应于基频（F0）。AFF 还捕获了几个元音共振峰，例如 /a/ (F1$\sim 800Hz$)，/i, y/ (F1$\sim 300Hz$) 和 /u/[8]。值得注意的是，在图 2 (b) 中，“mm”任务的 800 Hz 区域不太明显，这可能是因为缺少元音/a/。

4. 结论

本研究提出了用于 HF 检测的第一个大规模中文配对语音数据集。提出了一种“成对”分类方案，解耦个人差异，作为理想的基线。采用了两套特征集进行特征提取，并使用全连接模型进行分类。分类结果突出了个人差异是影响模型准确性的主要因素，这一挑战无法通过为男性和女性群体分别训练模型的常见做法完全解决。使用 AFF 进行频率分析时，识别出几个元音共振峰在 HF 检测中具有重要意义。

未来的工作可以从几个方面改进这项研究。首先，尽管“成对”方案在与说话人无关的情况下作为理想参考是合适的，但它可能在实际应用中并不实用，因为它需要一个已知的正常状态，而这并非总是可用的。为了缩小标准方案和成对方案之间的准确度差距，还需要进一步改进模型设计。其次，虽然频率分析捕获了重要的元音共振峰，但

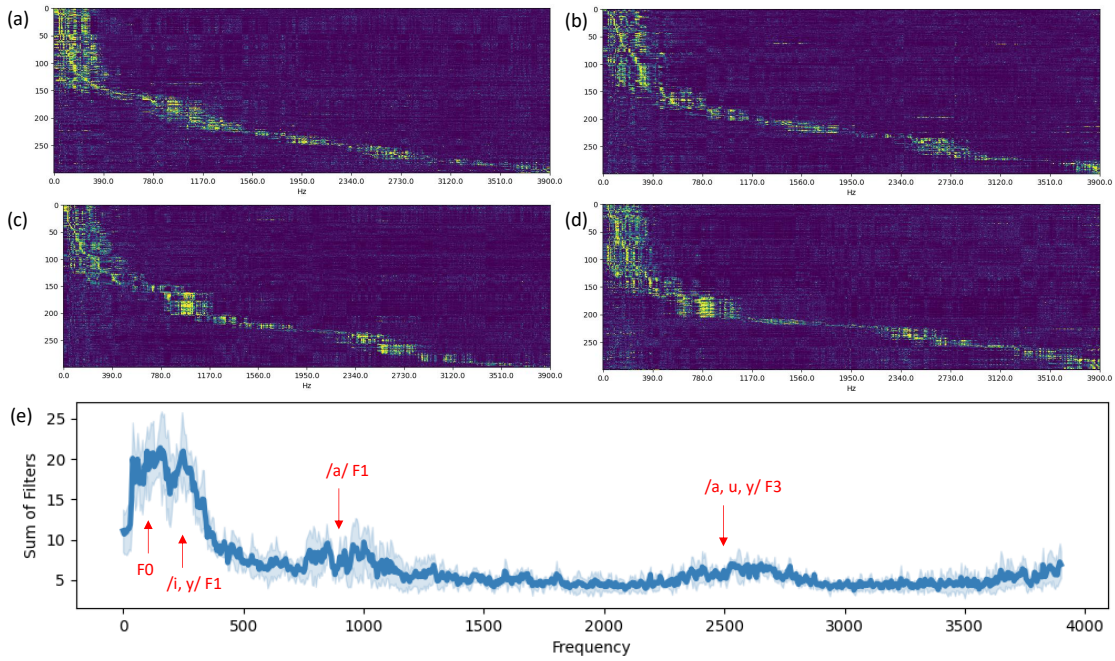

图 2: *AFF* 频率分析任务：*pg(a)*，*mm(b)*，*mlh(c)*，*c(d)*；以及四个任务的频率维度总和、平均值和标准差 *(e)*。基频 *(F0)* 和元音共振频率的位置在 *(e)* 中标记。关于任务名称缩写，请参见表 *2*。

还需要进一步完善以更准确地识别共振频率及其确切位置。

5. 伦理学

该研究遵循赫尔辛基宣言的指导原则，并获得了台州市人民医院的批准（批准号 KY 2023-073-01，于 2023 年 6 月 7 日获得）。

6. References

[1] O. Amir, W. T. Abraham, Z. S. Azzam, G. Berger, S. D. Anker, S. P. Pinney, D. Burkhoff, I. D. Shallom, 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. Minkkinen, 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. Minkkinen, 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.