Target Speaker Extraction using Discrete Representations from Self-Supervised Models and Language Models

1st Beilong Tang
Duke Kunshan University
Kunshan, China
bt132@duke.edu

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—This document is a model and instructions for LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.**

***Index Terms*—target speaker extraction, speech separation, language models, audio discretization**

I. Introduction

Speech separation, the so-called cocktail party problem [1], focuses on separating each individual speaker's source from a mixture with multiple speakers. This task is easy for humans but not for computers. In real life, speech signals are usually companied with background noise or other speakers' speeches. These corrupted signals may not be optimal for tasks like speaker verification [2], [3], and speech recognition [4], [5], emphasizing the importance of having good and robust speech separation models.

Unlike blind speech separation, which focuses on separating each utterance from a mixture of known speakers, target speaker extraction aims at only extracting the target speaker's voice given another auxiliary information of the target speaker. Due to the development of Deep Neural Network (DNN), many models nowadays are discriminative models. They utilize a masking strategy to minimize the distance between the clean speech and estimated speech directly [6]–[9]. However, these discriminative models may not generalize well to unseen data and might even introduce unwanted distortions [10]. To solve these issues, researchers have proposed generative models. This method aims to learn the underlying distribution of the target speaker's voice and use this knowledge to generate the clean speech of the target speaker from a mixture of voices rather than directly mapping from mixed speech to clean speech. Some generative models, like diffusion models [11] and variational autoencoders (VAE) [12] have been studied. It has been demonstrated that generative models can achieve results comparable to those of discriminative models [11], [13].

Identify applicable funding agency here. If none, delete this.

Discretization of audio has been studied due to the advancement of language models (LM). This approach translates the audio into discrete tokens simulating text vocabularies and uses LMs to model them. This approach simplifies audio generation tasks by transforming the complex regression problems into classification problems [14]. There are currently two approaches for audio discretization, the first one uses neural audio codecs [15]. This approach typically captures the acoustic features of the audio [16]. The second approach utilizes self-supervised Learning (SSL) models like HuBERT [17] and WavLM [18]. SSL models have demonstrated excellent performances on many downstream tasks [19]. These SSL models extract continuous representations containing rich semantic and timbre information from a given speech. As demonstrated in [14], SSL models perform better than audio codec in tasks like speech enhancement and speech separation. Therefore, in this paper, we mainly explore the discretization of SSL models.

Discretization methods have been studied for speech enhancement [20] and blind speech separation [13], [14], however, this approach has been rarely studied for target speaker extraction. In this paper, we present a novel way to do target speaker extraction using discrete tokens and LMs. Inspired by the blind speech separation networks proposed in DASB [14], our model has three stages: encoding, modeling and decoding. For the encoding stage.

A. *Units*

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as "3.5-inch disk drive".
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.

- Do not mix complete spellings and abbreviations of units: "Wb/m^2" or "webers per square meter", not "webers/m^2". Spell out units when they appear in text: ". . . a few henries", not ". . . a few H".
- Use a zero before decimal points: "0.25", not ".25". Use "cm^3", not "cc".)

B. Equations

Number equations consecutively. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \tag{1}$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(1)", not "Eq. (1)" or "equation (1)", except at the beginning of a sentence: "Equation (1) is . . ."

C. LaTeX-Specific Advice

Please use "soft" (e.g., `\eqref{Eq}`) cross references instead of "hard" references (e.g., `(1)`). That will make it possible to combine sections, add equations, or change the order of figures or citations without having to go through the file line by line.

Please don't use the `{eqnarray}` equation environment. Use `{align}` or `{IEEEeqnarray}` instead. The `{eqnarray}` environment leaves unsightly spaces around relation symbols.

Please note that the `{subequations}` environment in LaTeX will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you've discovered a new method of counting.

BibTeX does not work by magic. It doesn't get the bibliographic data from thin air but from .bib files. If you use BibTeX to produce a bibliography you must send the .bib files.

LaTeX can't read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

LaTeX does not have precognitive abilities. If you put a `\label` command before the command that updates the counter it's supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a `\label` command should not go before the caption of a figure or a table.

Do not use `\nonumber` inside the `{array}` environment. It will not stop equation numbers inside `{array}` (there won't be any anyway) and it might stop a wanted equation number in the surrounding equation.

D. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

An excellent style manual for science writers is [21].

E. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

F. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you

to apply a style (in this case, ital
provided by the drop down menu to
the text.

Text heads organize the topics
basis. For example, the p
because all subsequent ma
one topic. If there are
level head (uppercase Ror
conversely, if ther
subheads should b

G. Figures and Ta

a) Positionin
tables at the top a
in the middle of c
across both colun
figures; table hea
figures and tables
abbreviation "Fig.

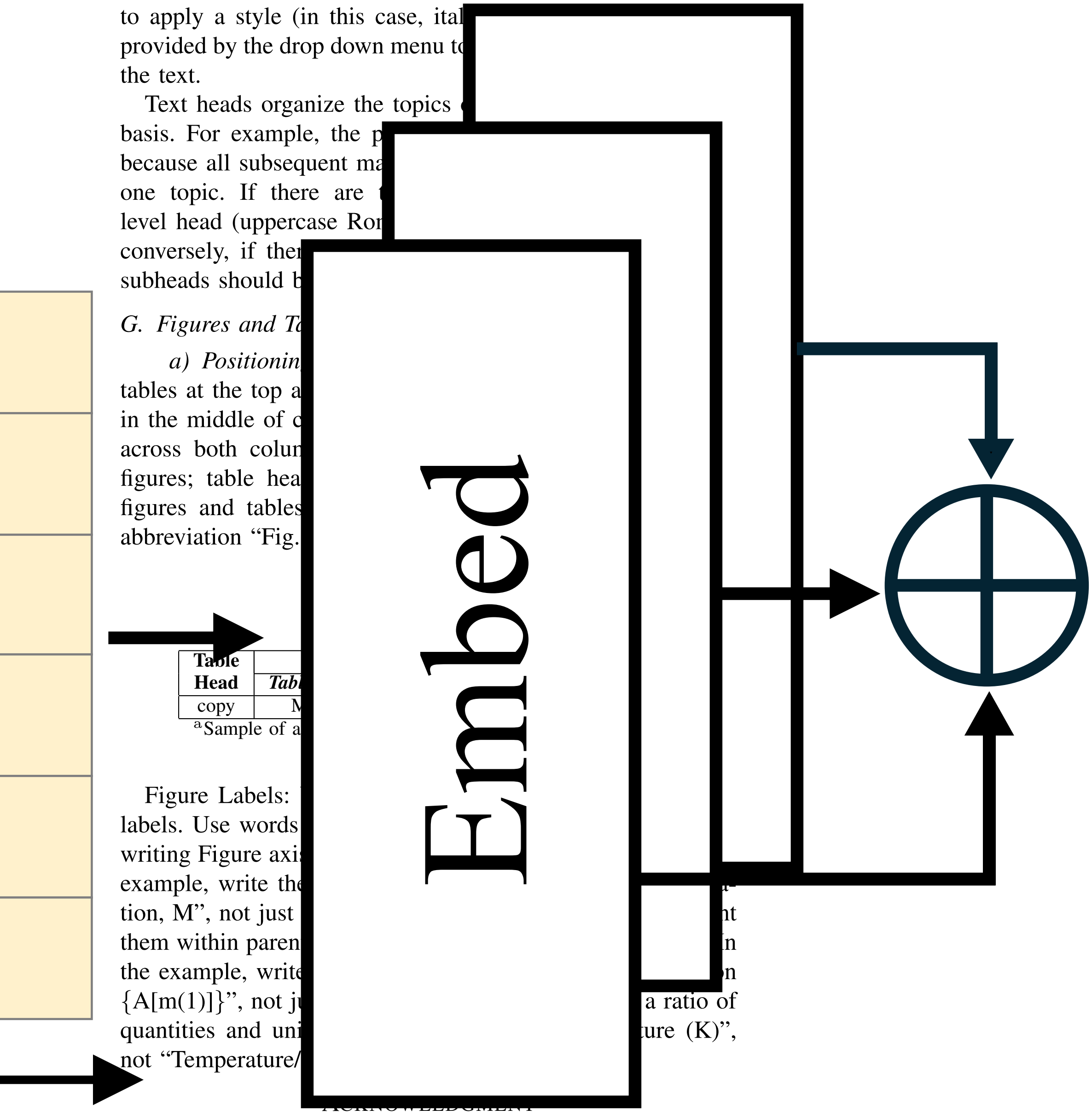


Table Head	*Tabl*
copy	M

[a]Sample of a

Figure Labels:
labels. Use words
writing Figure axi
example, write the
tion, M", not just
them within paren
the example, write
{A[m(1)]}", not ju ... a ratio of
quantities and uni ... ure (K)",
not "Temperature/

ACKNOWLEDGMENT

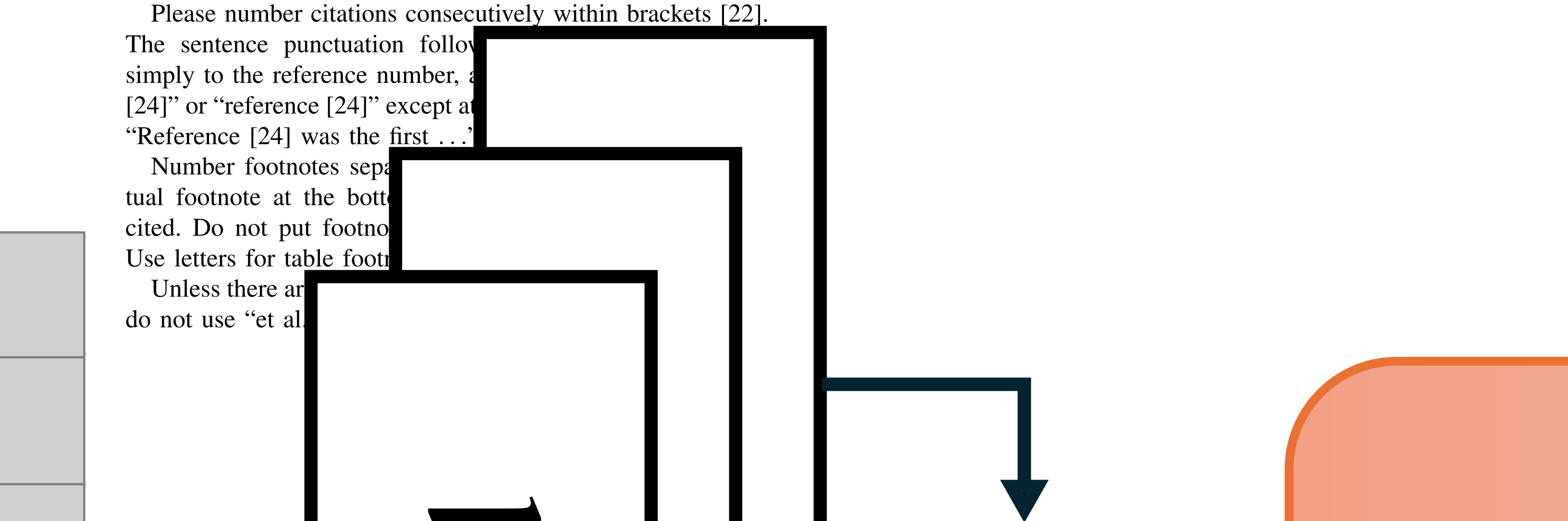
The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks …". Instead, try "R. B. G. thanks…". Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

Please number citations consecutively within brackets [22].
The sentence punctuation follov
simply to the reference number, a
[24]" or "reference [24]" except at
"Reference [24] was the first …"

Number footnotes sepa
tual footnote at the bott
cited. Do not put footno
Use letters for table footr

Unless there ar
do not use "et al.



even if they have been submitted for publication, should be cited as "unpublished" [26]. Papers that have been accepted for publication should be cited as "in press" [27]. Capitalize only the [28] first word [29] in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English [30] citation first, followed by the original foreign-language citation [31].

References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] W. Rao, C. Xu, E. S. Chng, and H. Li, "Target speaker extraction for overlapped multi-talker speaker verification," 2019. [Online]. Available: https://arxiv.org/abs/1902.02546

[3] C. Zhang, M. Yu, C. Weng, and D. Yu, "Towards robust speaker verification with target speaker enhancement," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6693–6697.

[4] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, "Speaker-aware neural network based beamformer for speaker extraction in speech mixtures," in *Interspeech*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:5587779

[5] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5554–5558.

[6] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[7] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "Spex+: A complete time domain speaker extraction network," 2020. [Online]. Available: https://arxiv.org/abs/2005.04686

[8] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 21–25.

[9] B. Zeng, H. Suo, Y. Wan, and M. Li, "Sef-net: Speaker embedding free target speaker extraction network," in *Interspeech*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260912744

[10] P. Wang, K. Tan, and D. Wang, "Bridging the gap between monaural speech enhancement and recognition with distortion-independent acoustic modeling," 2019. [Online]. Available: https://arxiv.org/abs/1903.04567

[11] T. Nguyen, G. Sun, X. Zheng, C. Zhang, and P. C. Woodland, "Conditional diffusion model for target speaker extraction," 2023. [Online]. Available: https://arxiv.org/abs/2310.04791

[12] R. Wang, L. Li, and T. Toda, "Dual-channel target speaker extraction based on conditional variational autoencoder and directional information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1968–1979, 2024.

[13] H. Erdogan, S. Wisdom, X. Chang, Z. Borsos, M. Tagliasacchi, N. Zeghidour, and J. R. Hershey, "Tokensplit: Using discrete speech representations for direct, refined, and transcript-conditioned speech separation and recognition," 2023. [Online]. Available: https://arxiv.org/abs/2308.10415

[14] P. Mousavi, L. D. Libera, J. Duret, A. Ploujnikov, C. Subakan, and M. Ravanelli, "Dasb - discrete audio and speech benchmark," 2024. [Online]. Available: https://arxiv.org/abs/2406.14294

[15] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved rvqgan," *ArXiv*, vol. abs/2306.06546, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259138883

[16] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, "Speechtokenizer: Unified speech tokenizer for speech large language models," 2024. [Online]. Available: https://arxiv.org/abs/2308.16692

[17] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," 2021. [Online]. Available: https://arxiv.org/abs/2106.07447

[18] 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, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[19] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "Superb: Speech processing universal performance benchmark," 2021. [Online]. Available: https://arxiv.org/abs/2105.01051

[20] Z. Wang, X. Zhu, Z. Zhang, Y. Lv, N. Jiang, G. Zhao, and L. Xie, "Selm: Speech enhancement using discrete tokens and language models," 2024. [Online]. Available: https://arxiv.org/abs/2312.09747

[21] M. Young, *The Technical Writer's Handbook*. Mill Valley, CA: University Science, 1989.

[22] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of lipschitz-hankel type involving products of bessel functions," *Phil. Trans. Roy. Soc. London*, vol. A247, pp. 529–551, April 1955.

[23] J. C. Maxwell, *A Treatise on Electricity and Magnetism*, 3rd ed. Oxford: Clarendon, 1892, vol. 2.

[24] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in *Magnetism*, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, vol. III, pp. 271–350.

[25] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2013. [Online]. Available: https://arxiv.org/abs/1312.6114

[26] K. Elissa, "Title of paper if known," unpublished.

[27] R. Nicole, "Title of paper with only first word capitalized," *J. Name Stand. Abbrev.*, in press.

[28] K. Eves and J. Valasek, "Adaptive control for singularly perturbed systems examples," Code Ocean, August 2023. [Online]. Available: https://codeocean.com/capsule/4989235/tree

[29] S. Liu, "Wi-fi energy detection testbed (12mtc)," gitHub repository, 2023. [Online]. Available: https://github.com/liustone99/Wi-Fi-Energy-Detection-Testbed-12MTC

[30] U. D. of Health, S. A. Human Services, and O. o. A. S. Mental Health Services Administration, "Treatment episode data set: discharges (teds-d): concatenated, 2006 to 2009," August 2013.

[31] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," *IEEE Transl. J. Magn. Japan*, vol. 2, pp. 740–741, August 1987, [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].