	doc_1		doc_2		decision	id
cases	authors	• Liu, Y. • Lo, C. • Tian, Z. • Zhang, C. • Zhu, J.	authors	• Jian Zhu • Zuoyu Tian • Yadong Liu • Cong Zhang • Chia-wen Lo	DUPLICATES	16
	title	Bootstrapping meaning through listening: Unsupervised learning of spoken sentence embeddings	title	Bootstrapping meaning through listening: Unsupervised learning of spoken sentence embeddings		
	publication_date	2022-10-23 00:00:00	publication_date	2022-10-23 00:00:00		
	source	SupportedSources.CORE	source	SupportedSources.INTERNET_ARCHIVE		
	journal		journal			
	volume		volume			
	doi	10.48550/arxiv.2210.12857	doi			
	urls	• https://core.ac.uk/download/551606265.pdf	urls	• https://web.archive.org/web/20221102175647/https://arxiv.org/pdf/2210.12857v1.pdf		
	id	id2643774132370098596	id	id4037631304466958798		
	abstract	Inducing semantic representations directly from speech signals is a highly challenging task but has many useful applications in speech mining and spoken language understanding. This study tackles the unsupervised learning of semantic representations for spoken utterances. Through converting speech signals into hidden units generated from acoustic unit discovery, we propose WavEmbed, a multimodal sequential autoencoder that predicts hidden units from a dense representation of speech. Secondly, we also propose S-HuBERT to induce meaning through knowledge distillation, in which a sentence embedding model is first trained on hidden units and passes its knowledge to a speech encoder through contrastive learning. The best performing model achieves a moderate correlation (0.5~0.6) with human judgments, without relying on any labels or transcriptions. Furthermore, these models can also be easily extended to leverage textual transcriptions of speech to learn much better speech embeddings that are strongly correlated with human annotations. Our proposed methods are applicable to the development of purely data-driven systems for speech mining, indexing and search	abstract	Inducing semantic representations directly from speech signals is a highly challenging task but has many useful applications in speech mining and spoken language understanding. This study tackles the unsupervised learning of semantic representations for spoken utterances. Through converting speech signals into hidden units generated from acoustic unit discovery, we propose WavEmbed, a multimodal sequential autoencoder that predicts hidden units from a dense representation of speech. Secondly, we also propose S-HuBERT to induce meaning through knowledge distillation, in which a sentence embedding model is first trained on hidden units and passes its knowledge to a speech encoder through contrastive learning. The best performing model achieves a moderate correlation (0.5~0.6) with human judgments, without relying on any labels or transcriptions. Furthermore, these models can also be easily extended to leverage textual transcriptions of speech to learn much better speech embeddings that are strongly correlated with human annotations. Our proposed methods are applicable to the development of purely data-driven systems for speech mining, indexing and search.		
	versions		versions			