cases	doc_1	doc_2	decision	id
	<table><tr><td>**authors**</td><td>• Pasad, A. • Shi, B. • Kamper, H. • Livescu, K.</td></tr><tr><td>**title**</td><td>On the Contributions of Visual and Textual Supervision in Low-Resource Semantic Speech Retrieval</td></tr><tr><td>**publication_date**</td><td>2019-09-15 00:00:00</td></tr><tr><td>**source**</td><td>SupportedSources.CROSSREF</td></tr><tr><td>**journal**</td><td></td></tr><tr><td>**volume**</td><td></td></tr><tr><td>**doi**</td><td>10.21437/interspeech.2019-3051</td></tr><tr><td>**urls**</td><td>• http://dx.doi.org/10.21437/interspeech.2019-3051</td></tr><tr><td>**id**</td><td>id-1606990110734108228</td></tr><tr><td>**abstract**</td><td></td></tr><tr><td>**versions**</td><td></td></tr></table>	<table><tr><td>**authors**</td><td>• Ankita Pasad • Bowen Shi • Herman Kamper • Karen Livescu</td></tr><tr><td>**title**</td><td>On the Contributions of Visual and Textual Supervision in Low-Resource Semantic Speech Retrieval</td></tr><tr><td>**publication_date**</td><td>2019-04-24 17:44:06+00:00</td></tr><tr><td>**source**</td><td>SupportedSources.ARXIV</td></tr><tr><td>**journal**</td><td>None</td></tr><tr><td>**volume**</td><td></td></tr><tr><td>**doi**</td><td></td></tr><tr><td>**urls**</td><td>• http://arxiv.org/pdf/1904.10947v2 • http://arxiv.org/abs/1904.10947v2 • http://arxiv.org/pdf/1904.10947v2</td></tr><tr><td>**id**</td><td>id3492340363340870377</td></tr><tr><td>**abstract**</td><td>Recent work has shown that speech paired with images can be used to learn semantically meaningful speech representations even without any textual supervision. In real-world low-resource settings, however, we often have access to some transcribed speech. We study whether and how visual grounding is useful in the presence of varying amounts of textual supervision. In particular, we consider the task of semantic speech retrieval in a low-resource setting. We use a previously studied data set and task, where models are trained on images with spoken captions and evaluated on human judgments of semantic relevance. We propose a multitask learning approach to leverage both visual and textual modalities, with visual supervision in the form of keyword probabilities from an external tagger. We find that visual grounding is helpful even in the presence of textual supervision, and we analyze this effect over a range of sizes of transcribed data sets. With ~5 hours of transcribed speech, we obtain 23% higher average precision when also using visual supervision.</td></tr><tr><td>**versions**</td><td></td></tr></table>	DUPLICATES	345