Podcasts

–

a new type of language data

jussi karlgren

about me

- former student at DSV
- linguistics, math, computer science
- worked in industrial language technology research since 1987
- mainly interested in stylistics and non-topical variation
- most recently worked at Spotify (first exposure to recorded speech)

podcasts, are they different?

podcast content is different from other public media

podcast conversations are different from other media

some patterns are similar to what was known before

some are entirely new and as of yet somewhat unformed

most importantly - podcasts are recorded speech!

speech

present as it happens

bound to participants

co-constructive

sequential

writing

persistent and archival

content is separated from creator and audience

reader in control of timing, pace, and order

speech vs writing

Fleeting

speech is made and used in the moment and is [illegible]ed and archival

Personal

speech is used in situations wh[illegible]ker and recipient are present

Elaboration

speech is less plann[illegible] writing; writing needs explicitness to be comprehensible where [illegible] can rely on shared understanding

Abstraction

sp[illegible] situational; writing is general and has unbound variables

Use cases

speech is used for different purposes than writing

no longer as true as it used to be!

differences less notable now

- new technology changes much of that
- speech is now persistent and distributable
- text has lower publication thresholds than before and can be situational and non-archival

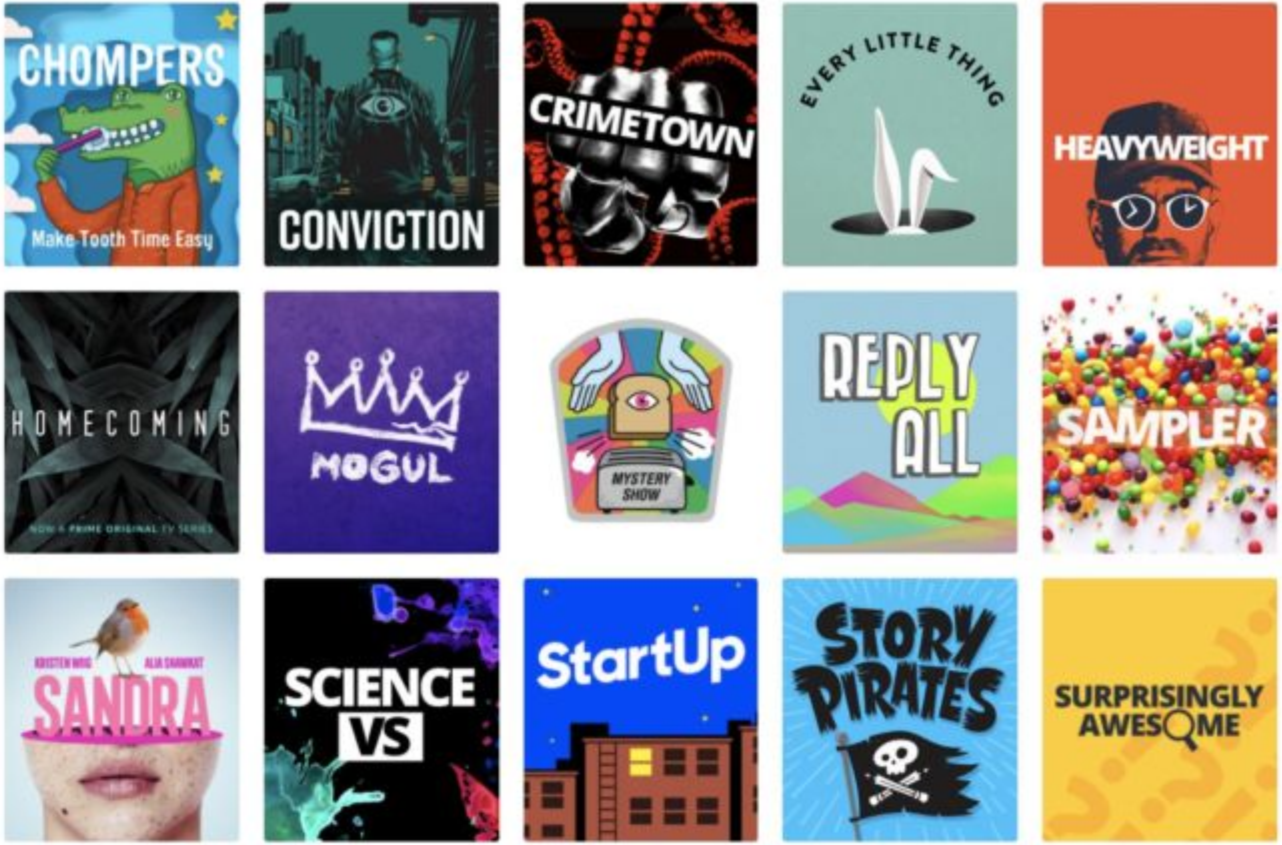
(this does not mean we are returning to a "new orality")

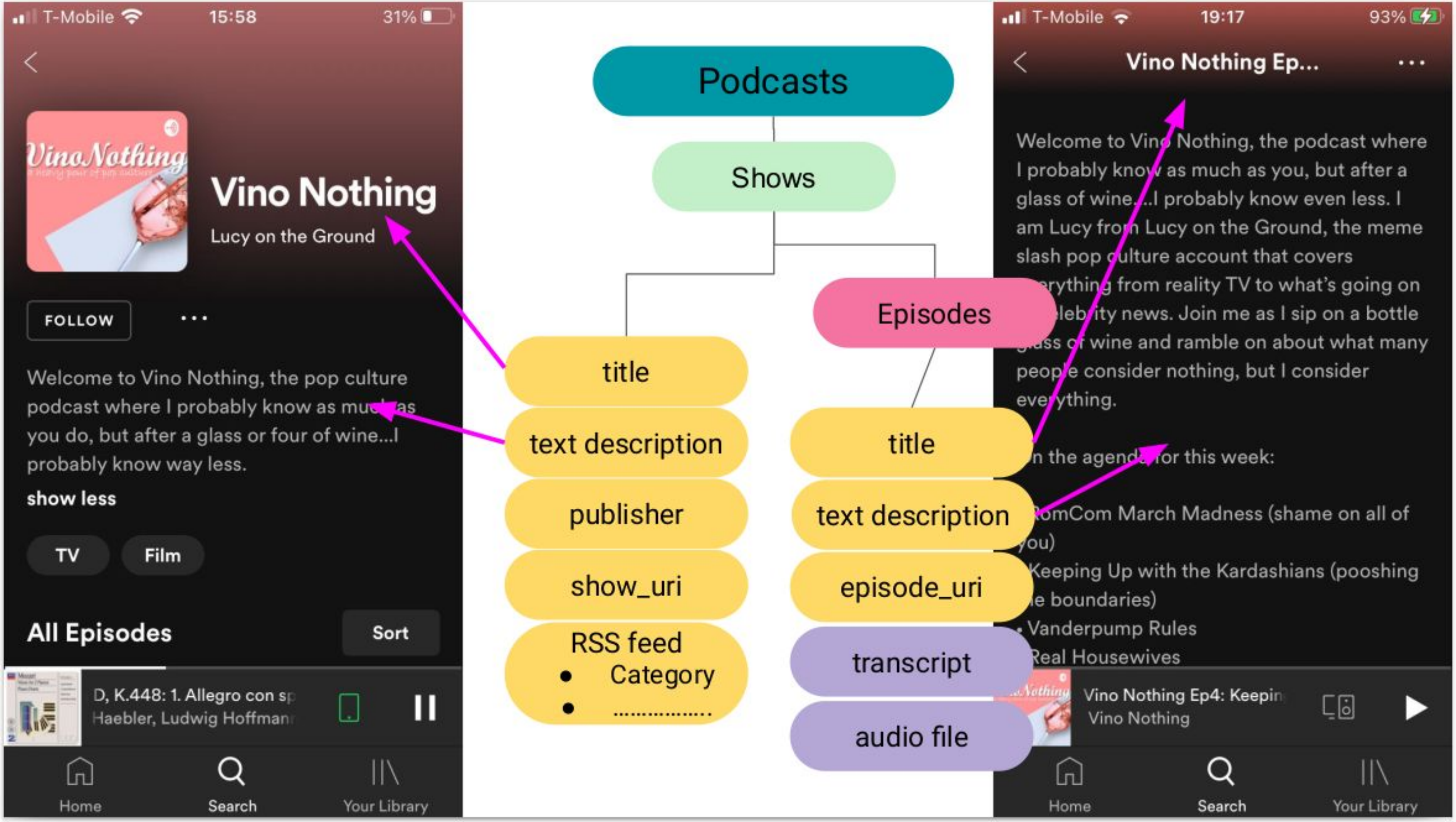
speech is gaining ground

it is likely many important things will be primarily available as recorded speech!

the podcast dataset from Spotify

- a new research resource
- fully-transcribed, large-scale podcast dataset
- 4TB of data - 200k episodes with audio in English and in Portuguese





Welcome to Vino Nothing, the pop culture podcast where I probably know as much as you do, but after a glass or four of wine...I probably know way less.

Welcome to Vino Nothing, the podcast where I probably know as much as you, but after a glass of wine...I probably know even less. I am Lucy from Lucy on the Ground, the meme slash pop culture account that covers everything from reality TV to what's going on celebrity news. Join me as I sip on a bottle glass of wine and ramble on about what many people consider nothing, but I consider everything.

On the agenda for this week:

- RomCom March Madness (shame on all of you)
- Keeping Up with the Kardashians (pooshing the boundaries)
- Vanderpump Rules
- Real Housewives

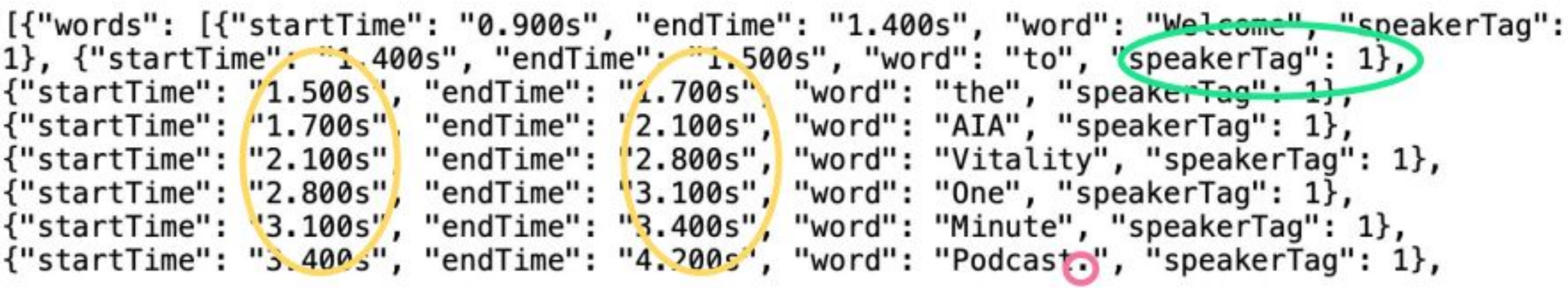
transcripts

3d party

with timing

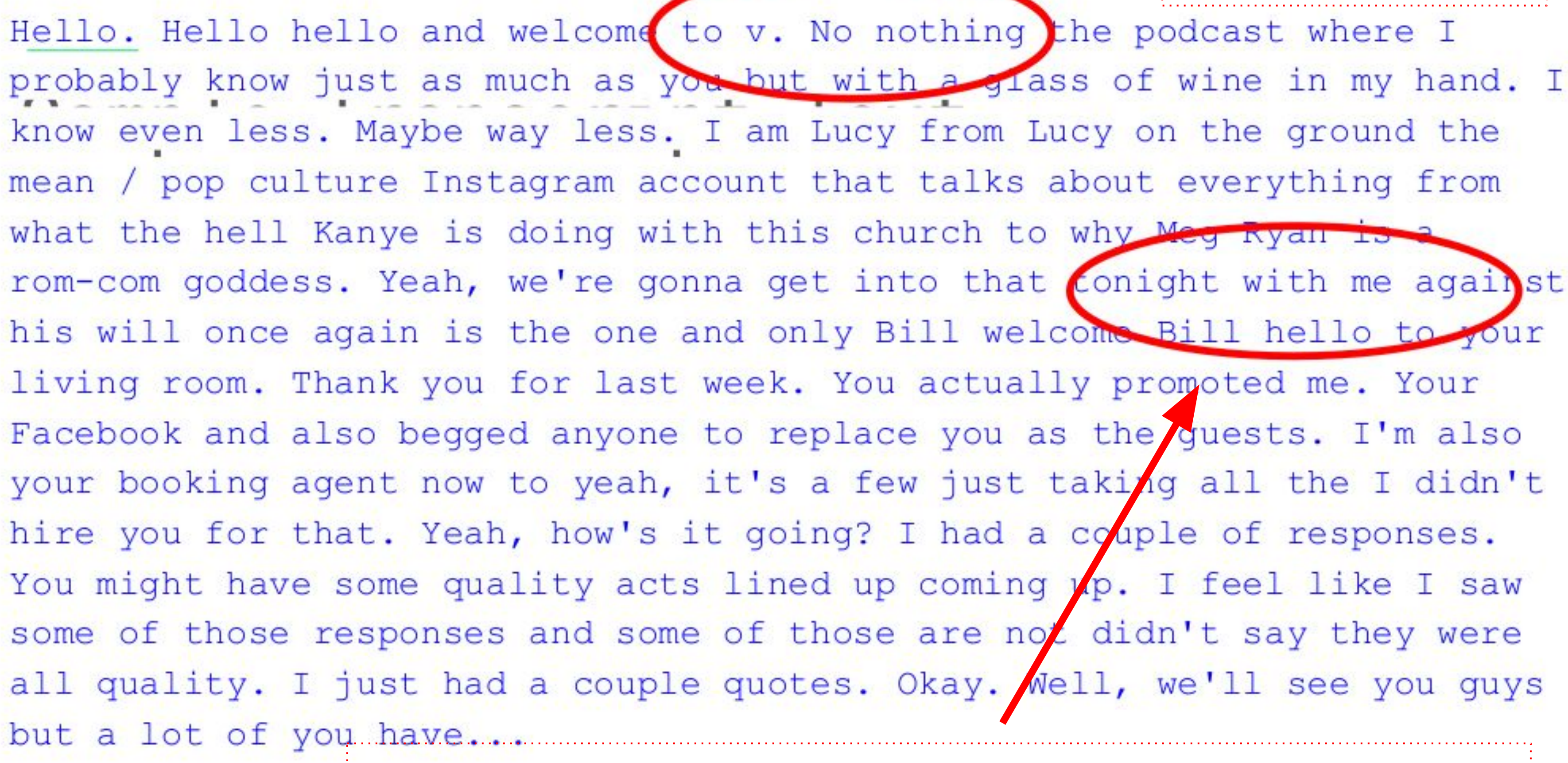
with speaker tags

with inferred punctuation and sentence segmentation



```
[{"words": [{"startTime": "0.900s", "endTime": "1.400s", "word": "Welcome", "speakerTag":
1}, {"startTime": "1.400s", "endTime": "1.500s", "word": "to", "speakerTag": 1},
{"startTime": "1.500s", "endTime": "1.700s", "word": "the", "speakerTag": 1},
{"startTime": "1.700s", "endTime": "2.100s", "word": "AIA", "speakerTag": 1},
{"startTime": "2.100s", "endTime": "2.800s", "word": "Vitality", "speakerTag": 1},
{"startTime": "2.800s", "endTime": "3.100s", "word": "One", "speakerTag": 1},
{"startTime": "3.100s", "endTime": "3.400s", "word": "Minute", "speakerTag": 1},
{"startTime": "3.400s", "endTime": "4.200s", "word": "Podcast.", "speakerTag": 1},
```

"Vino Nothing"



Hello. Hello hello and welcome to v. No nothing the podcast where I probably know just as much as you but with a glass of wine in my hand. I know even less. Maybe way less. I am Lucy from Lucy on the ground the mean / pop culture Instagram account that talks about everything from what the hell Kanye is doing with this church to why Meg Ryan is a rom-com goddess. Yeah, we're gonna get into that tonight with me against his will once again is the one and only Bill welcome Bill hello to your living room. Thank you for last week. You actually promoted me. Your Facebook and also begged anyone to replace you as the guests. I'm also your booking agent now to yeah, it's a few just taking all the I didn't hire you for that. Yeah, how's it going? I had a couple of responses. You might have some quality acts lined up coming up. I feel like I saw some of those responses and some of those are not didn't say they were all quality. I just had a couple quotes. Okay. Well, we'll see you guys but a lot of you have...

… get into that tonight. With me, against his will …

TREC podcast challenge 2020-2021

Shared task:

1. segment retrieval
2. summarisation

AP news (from TREC data set)
Twitter (2017, Harvey storm)
Blogs (Authorship corpus)
Podcast data set
Switchboard corpus of transcribed telephone conversations
Movie scripts (UCSC collection)
Ted talk transcripts

Karlgren (2022). Lexical variation in English language podcasts, editorial media, and social media. NEJLT.

segment search, 1st edition, 2020

given a topic, find 120s segments that are relevant

```
<topic>
<num>41</num>
<query>gmo food labeling</query>
<type>topical</type>
<description>
Some people say we should avoid
foods with genetically-modified
organisms as ingredients. I would
like to learn about GMO food
labeling. What are people saying
about the pros and cons? What's
the difference between the
European and US approaches to GMO
food labeling?
</description>
</topic>
```

```
<topic>
<num>56</num>
<query>gaslighting</query>
<type>known item</type>
<description>On Twitter I
saw someone reference a
podcast interview with a
doctor about gaslighting
within organizational
systems and I would like
to hear it.
</description>
</topic>
```

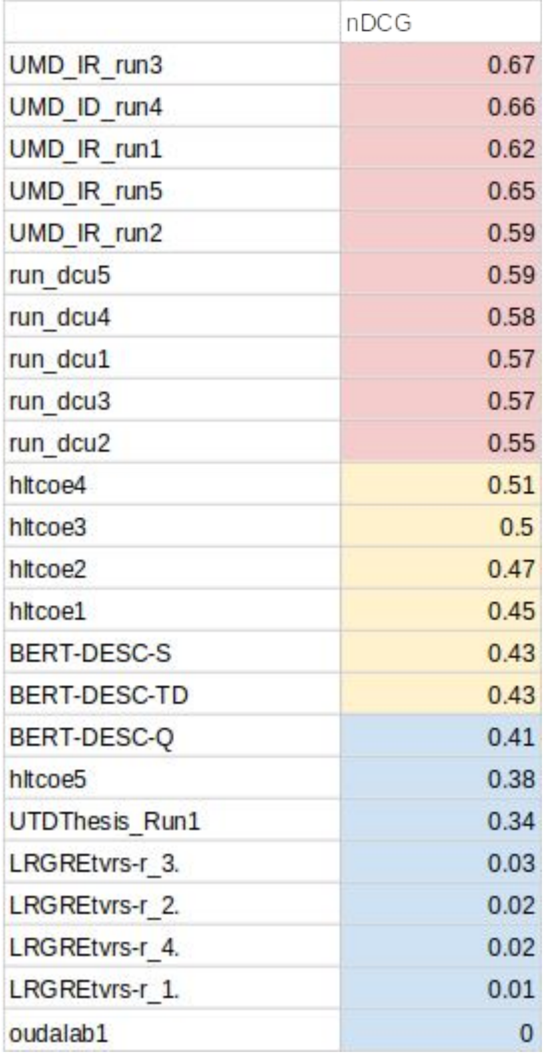
segment search, 2nd edition, 2021

given a topic, find 120s segments that are

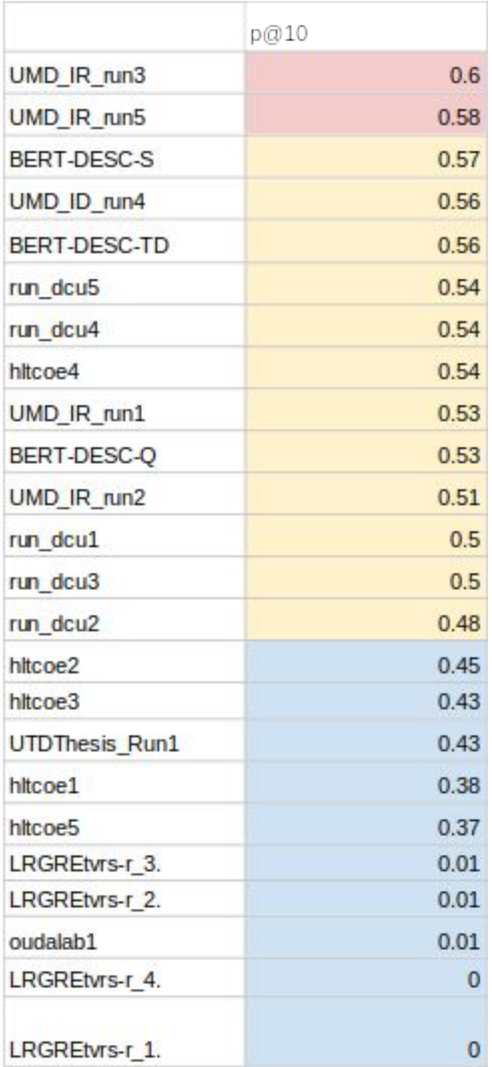
- **relevant** to the topic description

but also rerank them for being

1. **Entertaining**: topically relevant to the topic description AND the topic is presented in a way which the speakers intend to be amusing and entertaining to the listener, rather than informative or evaluative;
2. **Subjective**: topically relevant to the topic description AND the speaker or speakers explicitly and clearly express a polar opinion about the query topic, so that the approval or disapproval of the speaker is evident in the segment;
3. **Discussion**: topically relevant to the topic description AND includes more than one speaker participating with non-trivial topical contribution (e.g. mere grunts, expressions of agreement, or discourse management cues (“go on”, “right”, “well, I don’t know …” etc) are not sufficient).



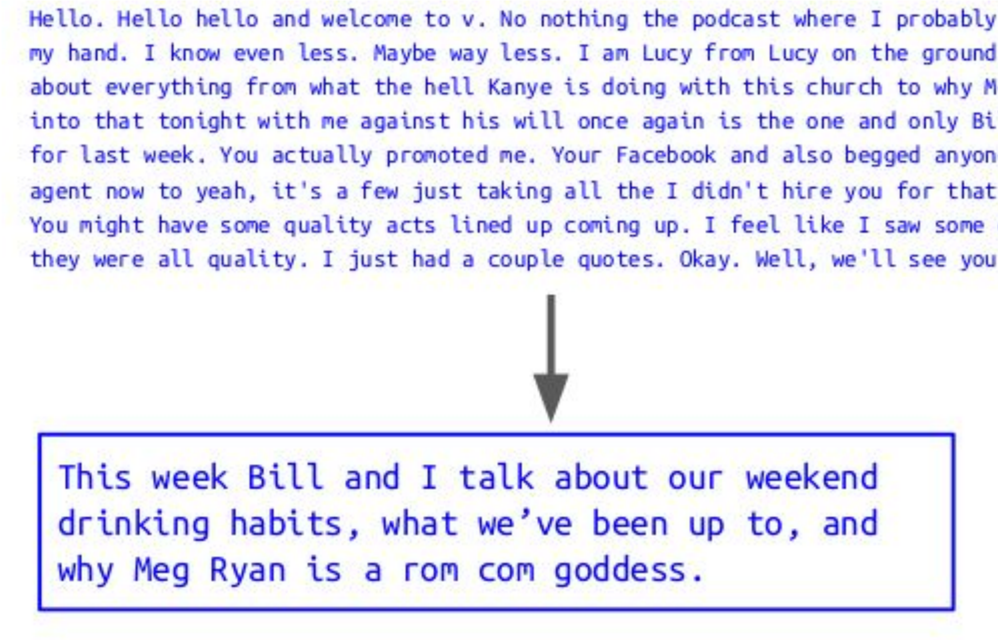
	nDCG
UMD_IR_run3	0.67
UMD_ID_run4	0.66
UMD_IR_run1	0.62
UMD_IR_run5	0.65
UMD_IR_run2	0.59
run_dcu5	0.59
run_dcu4	0.58
run_dcu1	0.57
run_dcu3	0.57
run_dcu2	0.55
hltcoe4	0.51
hltcoe3	0.5
hltcoe2	0.47
hltcoe1	0.45
BERT-DESC-S	0.43
BERT-DESC-TD	0.43
BERT-DESC-Q	0.41
hltcoe5	0.38
UTDThesis_Run1	0.34
LRGREtvrs-r_3.	0.03
LRGREtvrs-r_2.	0.02
LRGREtvrs-r_4.	0.02
LRGREtvrs-r_1.	0.01
oudalab1	0



	p@10
UMD_IR_run3	0.6
UMD_IR_run5	0.58
BERT-DESC-S	0.57
UMD_ID_run4	0.56
BERT-DESC-TD	0.56
run_dcu5	0.54
run_dcu4	0.54
hltcoe4	0.54
UMD_IR_run1	0.53
BERT-DESC-Q	0.53
UMD_IR_run2	0.51
run_dcu1	0.5
run_dcu3	0.5
run_dcu2	0.48
hltcoe2	0.45
hltcoe3	0.43
UTDThesis_Run1	0.43
hltcoe1	0.38
hltcoe5	0.37
LRGREtvrs-r_3.	0.01
LRGREtvrs-r_2.	0.01
oudalab1	0.01
LRGREtvrs-r_4.	0
LRGREtvrs-r_1.	0

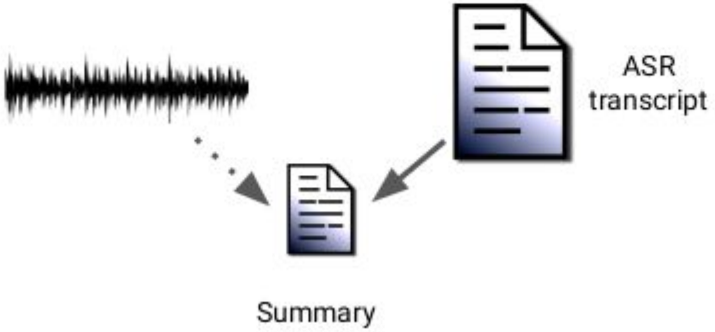
Hard	86	**best drum solo:** I want to hear about great drummers and especially their solos. A segment is relevant if it names the drummer appreciatively and mentions them playing a solo in some song or in some concert.
Easy	62	**descriptions of ponzi schemes and other financial scams:** I want to find stories about Ponzi schemes or similar scams. Segments that claim that something (e.g. bitcoin investing) is a Ponzi scheme or a scam without elaborating on how are not relevant - to be relevant they need to describe the scam and how it works.
Many Entertaining	66	**how to handle failing a job interview:** Any material on how a job interview fails or how a candidate was rejected due to the interview: tips, advice, or personal anecdotes and testimonials are all relevant. If a rejection is not about the interview, even if the segment mentions an interview, it is not relevant.
Few Entertaining	85	**personality disorders:** Discussions about any personality disorder are relevant. Passing claims that someone (e.g. a criminal, a celebrity, the speakers themselves) has a personality disorder without discussing the disorder itself are not.
Many Subjective	67	**pros and cons of ubi:** I want to find arguments for and against universal basic income.
Few Subjective	71	**roman empire:** I am looking to learn something about the history of the Roman Empire
Many Discussion	78	**taboo topics:** I want to understand what topics others consider to be taboo. To be relevant, the segment must mention that a topic is off limits and be clear about what the topic in question is. A mention that some topics are not on is not sufficient. A very general mention that some topics are taboo, e.g. ”sexuality” is partially relevant.
Few Discussion	95	**limericks:** I want to hear limericks. Discussion about limericks is not relevant if a limerick is not included in the segment.

summarisation

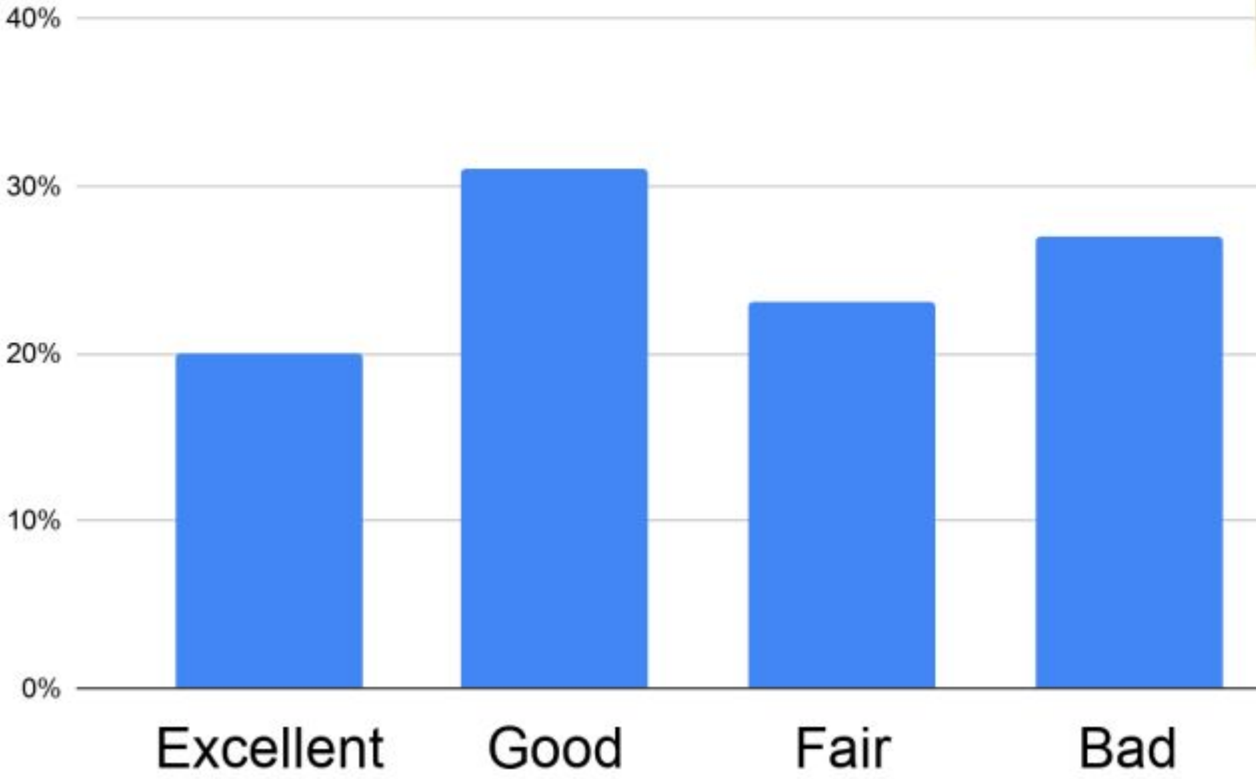


Hello. Hello hello and welcome to v. No nothing the podcast where I probably know just as much as you but with a glass of wine in my hand. I know even less. Maybe way less. I am Lucy from Lucy on the ground the mean / pop culture Instagram account that talks about everything from what the hell Kanye is doing with this church to why Meg Ryan is a rom-com goddess. Yeah, we're gonna get into that tonight with me against his will once again is the one and only Bill welcome Bill hello to your living room. Thank you for last week. You actually promoted me. Your Facebook and also begged anyone to replace you as the guests. I'm also your booking agent now to yeah, it's a few just taking all the I didn't hire you for that. Yeah, how's it going? I had a couple of responses. You might have some quality acts lined up coming up. I feel like I saw some of those responses and some of those are not didn't say they were all quality. I just had a couple quotes. Okay. Well, we'll see you guys but a lot of you have...

This week Bill and I talk about our weekend drinking habits, what we’ve been up to, and why Meg Ryan is a rom com goddess.



Only Half of Creator Generated Descriptions Make Excellent or Good Summaries



	ROUGE R	ROUGE P	ROUGE F
First One Minute	**0.28202**	0.08710	0.11571
TextRank on Sentences	0.16194	0.06532	0.07987
TextRank on Segments	0.16523	0.08305	0.09300
BART Pre-trained on News	0.27193	0.08491	0.11299
BART Fine-tuned on Training	0.21013	**0.20782**	**0.16635**

those data are available for experimentation!

some observations about podcast material

podcasts are happy

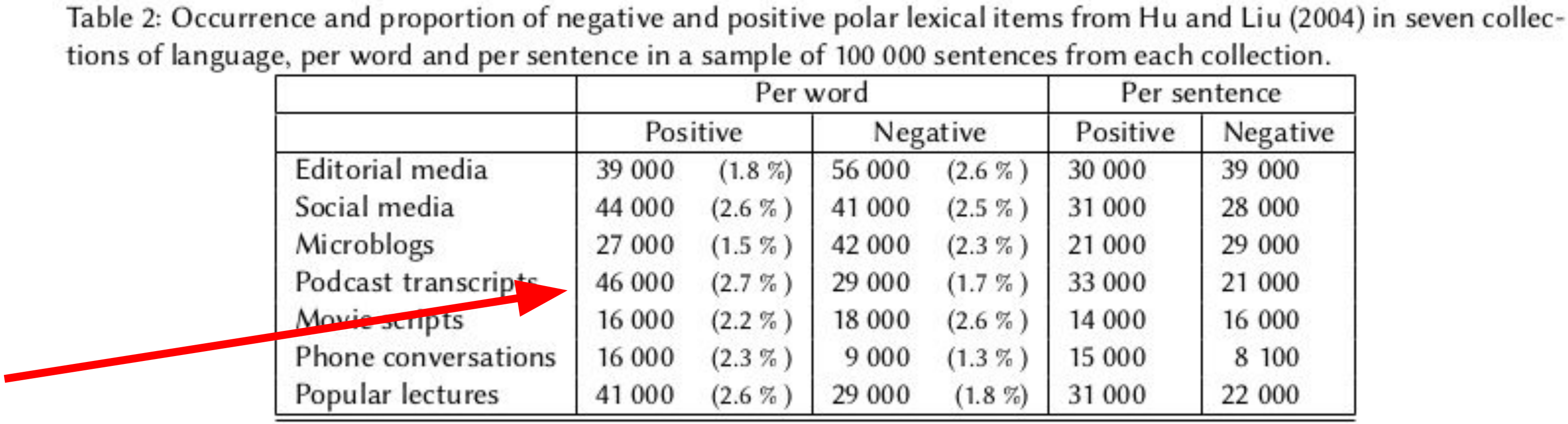
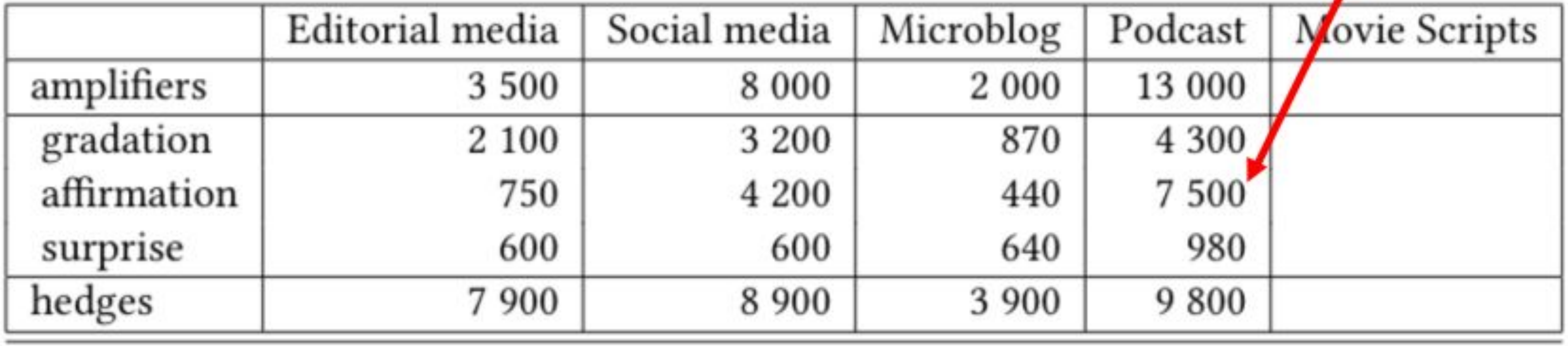


Table 2: Occurrence and proportion of negative and positive polar lexical items from Hu and Liu (2004) in seven collections of language, per word and per sentence in a sample of 100 000 sentences from each collection.

	Per word				Per sentence	
	Positive		Negative		Positive	Negative
Editorial media	39 000	(1.8 %)	56 000	(2.6 %)	30 000	39 000
Social media	44 000	(2.6 %)	41 000	(2.5 %)	31 000	28 000
Microblogs	27 000	(1.5 %)	42 000	(2.3 %)	21 000	29 000
Podcast transcripts	46 000	(2.7 %)	29 000	(1.7 %)	33 000	21 000
Movie scripts	16 000	(2.2 %)	18 000	(2.6 %)	14 000	16 000
Phone conversations	16 000	(2.3 %)	9 000	(1.3 %)	15 000	8 100
Popular lectures	41 000	(2.6 %)	29 000	(1.8 %)	31 000	22 000

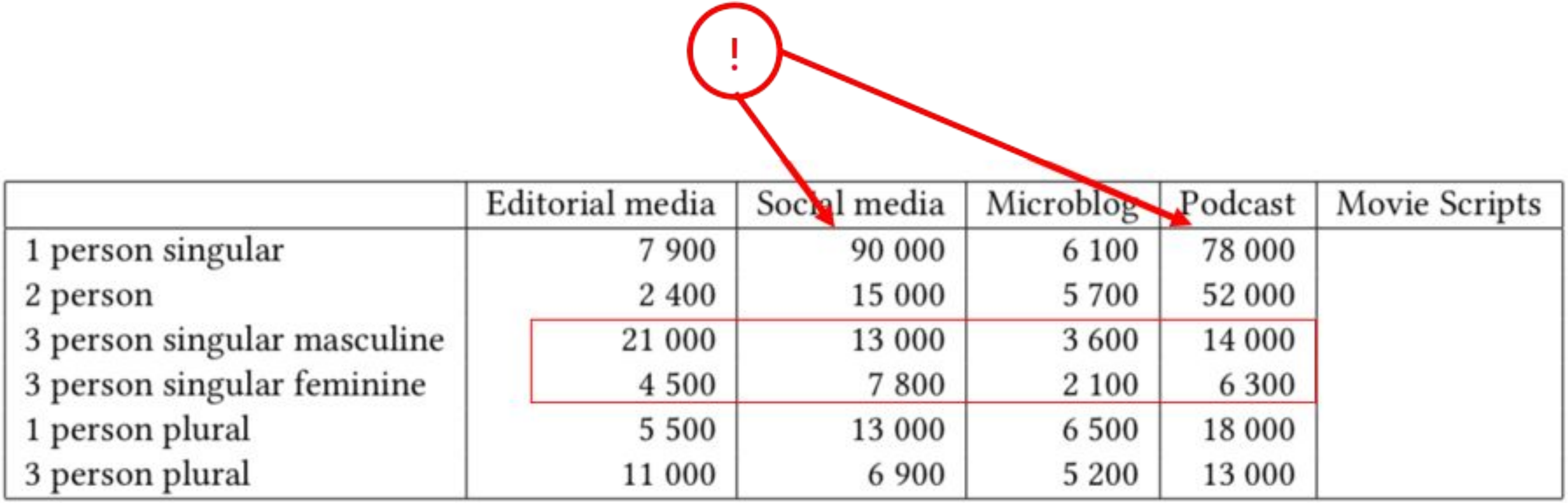
podcasts try to convince



!

	Editorial media	Social media	Microblog	Podcast	Movie Scripts
amplifiers	3 500	8 000	2 000	13 000	
gradation	2 100	3 200	870	4 300	
affirmation	750	4 200	440	7 500	
surprise	600	600	640	980	
hedges	7 900	8 900	3 900	9 800	

more dudes



	Editorial media	Social media	Microblog	Podcast	Movie Scripts
1 person singular	7 900	90 000	6 100	78 000	
2 person	2 400	15 000	5 700	52 000	
3 person singular masculine	21 000	13 000	3 600	14 000	
3 person singular feminine	4 500	7 800	2 100	6 300	
1 person plural	5 500	13 000	6 500	18 000	
3 person plural	11 000	6 900	5 200	13 000	

that was pretty much all done on writing!

how about trying audio?

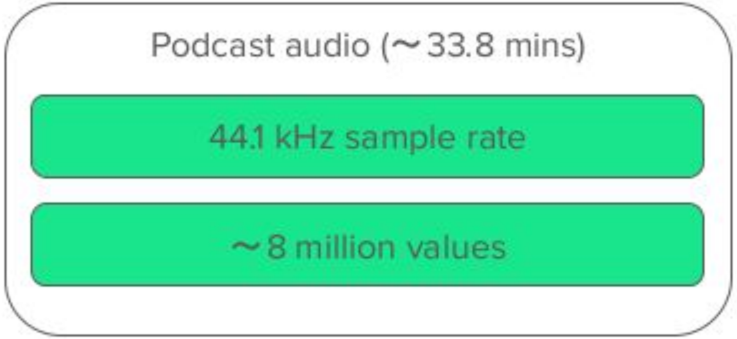
there are precomputed audio data to play with!

Features of Interest

Turn audio data into features for downstream processing for e.g. classifiers

phonetic & linguistic features

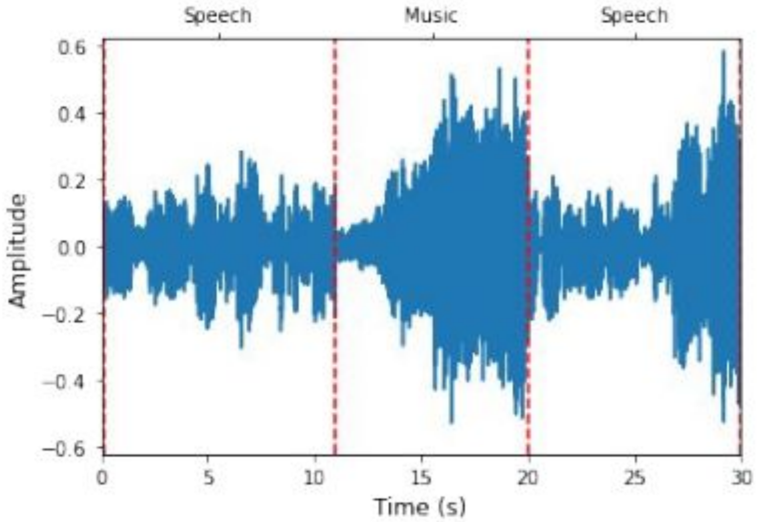
audio events



Podcast audio (∼33.8 mins)

44.1 kHz sample rate

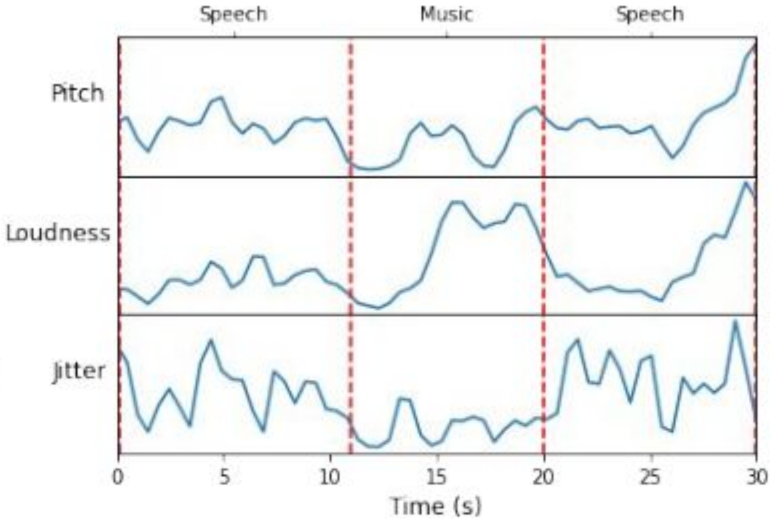
∼8 million values



eGeMAPS

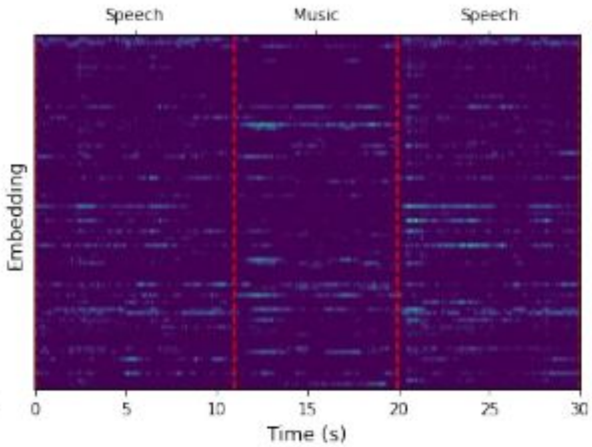
- **Extended Geneva Minimalistic Acoustic Parameter Set**
- Features:
  - time domain (e.g. speech rate),
  - the frequency domain (e.g. pitch),
  - the amplitude domain (e.g. loudness),
  - spectral energy domain (e.g. relative energy in different frequency bands).

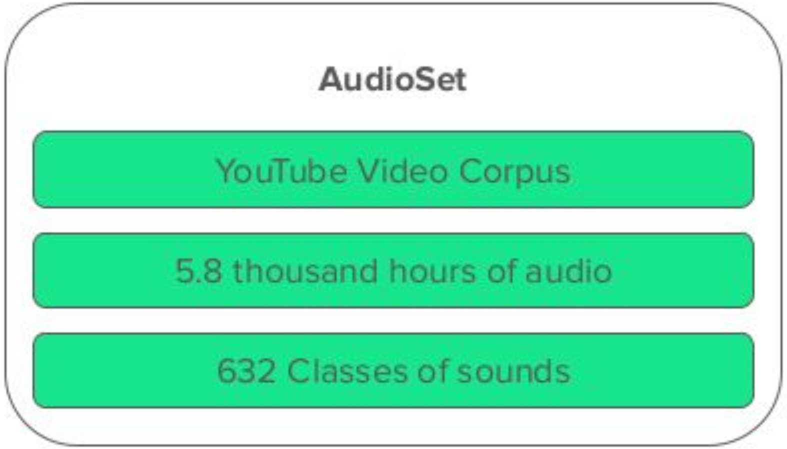
- 25 Low-Level Descriptor Features
- 88 Derived Features ("Functionals") (mean, sdev, peaks,...)



YAMNet

- **Yet Another MobileNet**
-
- 1024-dimensional embeddings
- 521 audio event scores (giggle, chant, moan, duck, helicopter ,...)



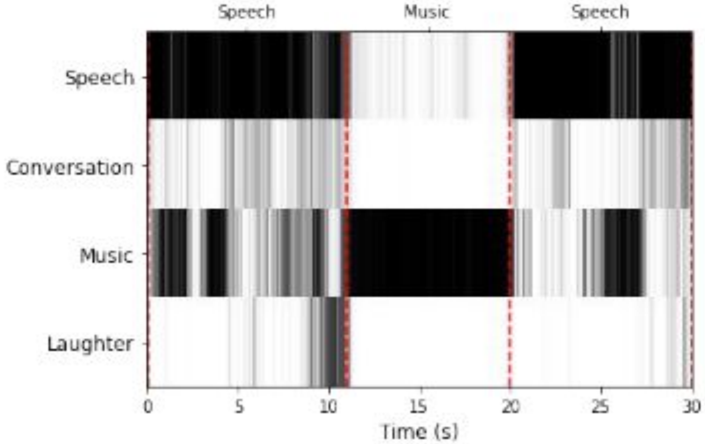


AudioSet

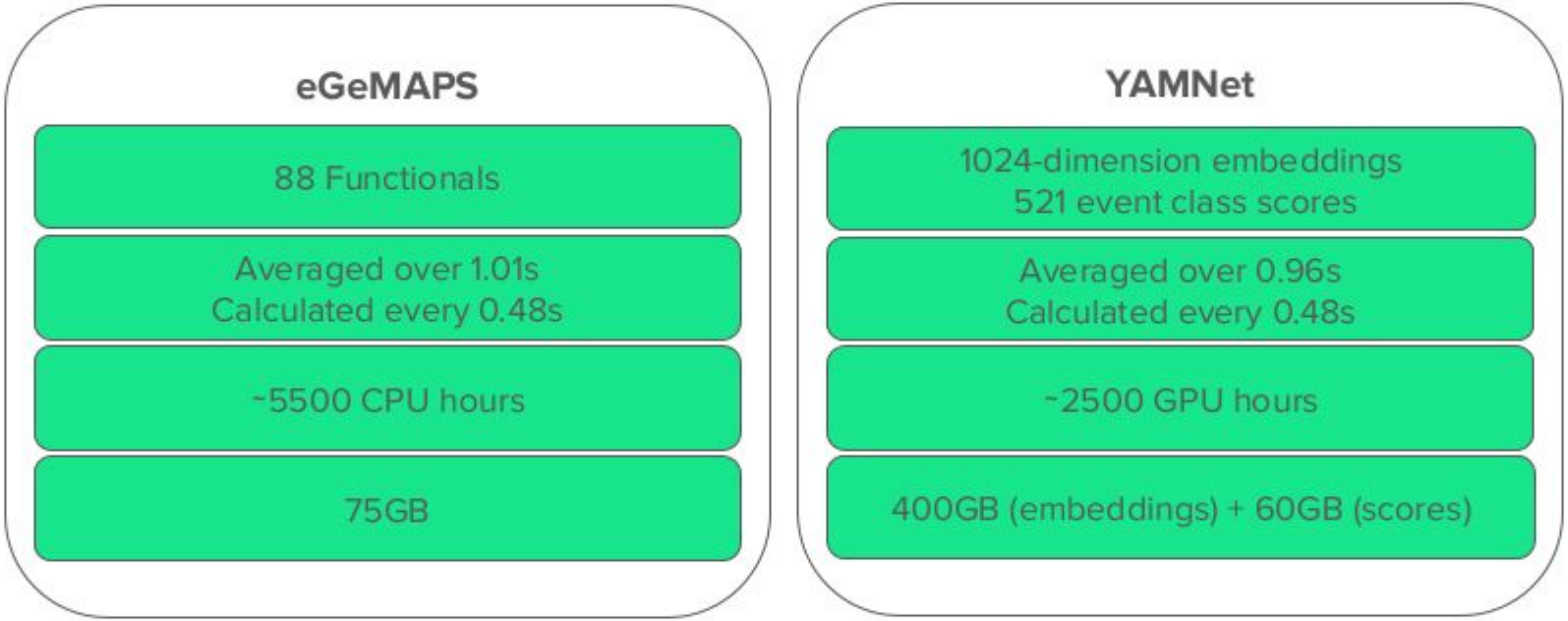
YouTube Video Corpus

5.8 thousand hours of audio

632 Classes of sounds



audio features



eGeMAPS

88 Functionals

Averaged over 1.01s
Calculated every 0.48s

~5500 CPU hours

75GB

YAMNet

1024-dimension embeddings
521 event class scores

Averaged over 0.96s
Calculated every 0.48s

~2500 GPU hours

400GB (embeddings) + 60GB (scores)

these can be used for experimentation!

some ideas for project work

what characterises unscripted unplanned convos?

is there a measure for intensity of conversation?

what are some emergent genres in distributed recorded speech?

what measures for signal richness and informational content could capture the difference between born-text and born-speech material?

how can we assess the quality of something which is distributed as recorded speech?

what is findability for items designed for delight, pleasure, and diversion rather than for topical task?

will publication in recorded speech change narrative practice?

e-objects are not remarkable. how could one build a library and a memory of one's path thru a collection?