Assessing Subjective Probabilistic Expectations in Household Surveys with Audio Records

Javier J. Alonso, Laura Crespo and **Nicolás Forteza**

Banco de España

The views expressed here do not necessarily reflect the views of the Bank of Spain or the Euro-System

Overview

Introduction I

In survey methodology...

- Audio records have been used to measure response times, speech rate, pitch or pausing to gain insights about the question-answer sequence and the cognitive process of interviewees (Conrad et al., 2013; Bergmann and Bristle, 2020; Benkí et al., 2011; Schaeffer et al., 2008).
- However, listening audio records is an intensive and difficult task, which requires personnel resources during long periods of time and applying common checklists to extract, interpret and judge similarly the information recorded.

Introduction II

At the same time ...

- Interviewers play a crucial role when asking questions and entering respondents' answers: the interviewer effect (Kish, 1962; Mangione et al., 1992; Ongena and Dijkstra, 2007; Vandenplas et al., 2017).
- One way to minimize such effects is done by the standardization of interviewing protocols referring to the interaction with the respondent, the reading of the questions, their probing behavior or their explanations to respondents about the survey and interview procedures (Fowler and Mangione, 1990).

What we do in this work:

1. Develop a state-of-the-art, open source audio transcription machine learning pipeline.
2. Apply it to the question on subjective probabilistic expectations on price changes of households's main residence in the next 12 months of the Spanish Survey of Household Finance (EFF by its Spanish acronym).
3. In economics, eliciting expectations plays a crucial role for understanding how agents take key decisions (D'Acunto and Weber, 2024).

What we do in this work:

This is a descriptive work:

1. Using the generated transcriptions of audio records, we extract specific indicators that inform us about how interviewers comply with standarized protocols and neutral probing.
2. We also generate additional indicators about the reactions or difficulties that respondents might encounter when answering this type of questions.
3. Future survey practitioners could use these insights to construct better their subjective expectations question!

Background

- We use data from 2020 and 2022 waves of the Spanish Survey of Household Finances (EFF by its Spanish acronym).
- The EFF is a F2F longitudinal survey conducted by the Spanish central bank, the Bank of Spain, that since 2002 provides detailed information on households' assets, debt, income and spending.
- 2020 wave was conducted in CATI mode.
- Population frame: Spanish Population Register and wealth tax files information for oversampling wealthy households.
- The EFF collects audio records for several questions along the interview (upon the respondent's consent), including questions on probabilistic expectations.

Subjective Probabilistic Expectations Question in the EFF

We are interested in knowing how you think the price of your home will evolve in the next 12 months: distribute ten points among the following five possibilities, assigning more points to the scenarios you think are more likely (assign 0 if a scenario looks impossible):

- *Drop over 6 %*
- *Drop in between 2% and 6%*
- *Approximately stable (drops or increases of no more than 2%)*
- *Increase in between 2% and 6%*
- *Increase larger than 6 %*
- *Don't know*
- *No answer*

Background

- Interviewers need to follow standardized protocols and methods to minimize interviewer effects or systematic biases:
  1. They are instructed to read literally the statement of the question.
  2. Explanations are provided if needed using the clarifications stated in the CAPI screen right below the main formulations.
  3. They are are instructed to read again the question if the respondent did not understand or could not follow the whole explanation.
- An automatic prompt appears on the CAPI screen if the answers entered in the computer by the interviewers do not add up to ten.
- A showcard or helpcard is given to households to better perform the exercise.

Data I - Sample Overview

Table: Summary statistics of audios length (in seconds) in our sample

	N	Mean	Median	Max	Min	p_25	p_75
EFF 2020	5770	75.11	66.56	262.10	6.00	50.28	91.96
EFF 2022	5695	70.23	62.54	302.60	7.20	47.55	84.36

Data II - Machine Learning Pipeline

How we transcribe the audios?:

- We use the Whisper-large-v3 model (Radford et al., 2022) from OpenAI: open source pre-trained speech-to-text AI model.
- We use a Python 3.9 local environment for confidentiality purposes, with a Nvidia 3080RTX 10Gb RAM GPU.
    1. Apply a speech enhancement model, following the methodology of Defossez et al. (2020), to remove background noises and room reverb.
    2. To remove hallucinations, we iterate over the detected hallucinated transcriptions with different chunk length configurations.
    3. We also apply a voice activity detection (VAD) model with Pyannote Bredin et al. (2019) for the characterization of silence (no speaking) parts of the conversation between the household and the interviewer.

Data III - Extracted Audio Characteristics

1. The interviewer formulates literally the question: using semantic similarity metric of encoded Sentence-Transformer model transcribed audios vs. the literal question.
2. Whether the showcard is used by the respondent.
3. Whether the interviewer reminds the household to add 10 points in total among the allocated.
4. Speech rate (n. words / conversation length).
5. Whether there's induced behavior of the interview: the interviewer probes a one option answer.
6. Whether the household does not understand the exercise and does not know what to answer, causing the interviewer to have to give additional explanations.
7. Whether the household is expressing some doubt about during the conversation.
8. Total duration time in seconds of silence in the conversation to proxy for cognitive process of household.

2, 3, 5, 6, and 7 by means of regular expressions. Regex Patterns

Data IV - Output Validation

We validate 120 audios' transcriptions and extracted characteristics by comparing them with a human manual annotation:

1. We use the Word Error Rate for the audio transcription (Wang et al., 2021). We obtain a 24%, compared to a 12% of the Whisper original paper.
2. For the Voice Activity Detection, we use the Detection Error Rate, achieving a 13% compared to the original 8% of the paper.
3. We use the Cohen's Kappa for all binary indicators. We obtain around a 90% of agreement between the extracted binary indicators and the human annotated ones.

Results I - Description of Extracted Variables

Table: Means of binary audio indicators by respondents' characteristics

Variables	Observations		Reminder sums 10 (%)		Int. shows card 10 (%)		Doubt (%)		Understand (%)		Induce (%)	
	2020	2022	2020	2022	2020	2022	2020	2022	2020	2022	2020	2022
Female	2280	2395	52.28	43.30	90.04	79.00	23.51	25.76	2.46	2.38	19.74	17.58
Male	3490	3300	51.38	43.15	89.97	80.27	13.18	15.39	0.89	1.21	14.76	14.36
Primary	773	673	47.22	36.40	87.06	73.70	25.87	30.46	3.75	3.27	22.77	20.65
Secondary	2048	2076	51.95	43.83	90.48	79.53	18.99	21.87	1.76	1.73	18.70	18.88
Tertiary	2910	2919	53.23	44.36	90.72	81.47	13.61	15.69	0.65	1.27	13.71	12.37
Under 35	311	367	52.41	40.33	87.78	80.65	17.04	13.08	2.25	1.36	17.36	11.99
35-45	986	925	54.26	42.49	90.67	80.54	13.69	17.08	0.91	1.62	14.71	12.86
46-55	1299	1358	53.43	42.34	91.15	80.04	14.32	18.11	1.15	1.55	15.86	14.21
56-65	1329	1342	52.22	45.31	90.97	80.77	18.21	19.75	1.58	1.49	17.01	15.35
66-75	1022	926	51.17	44.60	89.14	81.21	19.77	23.22	1.37	1.84	17.91	18.25
Over 75	823	777	45.69	41.70	87.73	74.26	21.63	24.84	2.55	2.45	18.35	21.11
Non-owner	1104	1262	52.63	38.83	91.58	81.38	20.38	22.58	2.36	2.61	18.84	16.48
Owner	4666	4433	51.52	44.46	89.63	79.27	16.52	18.95	1.31	1.44	16.22	15.50
Total	5770	5695	51.73	43.21	90.00	79.74	17.26	19.75	1.51	1.70	16.72	15.72

Results II - Description of Extracted Variables

Table: Mean and standard deviations for continuous audio indicators by respondents' characteristics

Variables	Observations		Verbatim Reading				Silent time				Speech rate			
	2020	2022	2020		2022		2020		2022		2020		2022	
			μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Female	2280	2395	0.69	0.21	0.70	0.16	30.65	20.41	18.86	14.84	2.48	0.57	2.76	0.58
Male	3490	3300	0.69	0.21	0.71	0.15	29.72	19.32	17.93	13.72	2.38	0.56	2.70	0.57
Primary	773	673	0.65	0.23	0.67	0.18	30.51	19.41	18.22	13.99	2.47	0.53	2.82	0.63
Secondary	2048	2076	0.69	0.21	0.70	0.16	30.95	20.57	19.07	15.23	2.44	0.55	2.75	0.56
Tertiary	2910	2919	0.70	0.20	0.72	0.15	29.35	19.21	17.79	13.49	2.39	0.59	2.69	0.57
Under 35	311	367	0.71	0.20	0.73	0.12	32.39	19.86	19.48	14.14	2.42	0.84	2.69	0.58
35-45	986	925	0.72	0.19	0.72	0.15	30.95	20.63	19.53	14.80	2.44	0.59	2.71	0.57
46-55	1299	1358	0.71	0.20	0.71	0.15	29.57	18.77	19.07	15.81	2.41	0.54	2.73	0.58
56-65	1329	1342	0.70	0.20	0.71	0.15	29.45	19.12	17.53	13.07	2.42	0.51	2.75	0.59
66-75	1022	926	0.66	0.22	0.69	0.17	29.05	19.09	16.81	12.82	2.38	0.52	2.75	0.55
Over 75	823	777	0.64	0.23	0.67	0.18	31.32	21.82	18.16	13.83	2.43	0.60	2.70	0.56
Non-owner	1104	1262	0.69	0.20	0.70	0.16	33.72	21.22	19.33	14.48	2.42	0.66	2.72	0.58
Owner	4666	4433	0.69	0.21	0.71	0.16	29.23	19.30	18.03	14.12	2.42	0.54	2.73	0.57
Total	5770	5695	0.69	0.21	0.71	0.16	30.09	19.76	18.32	14.21	2.42	0.57	2.73	0.57

Results III

Using the reported household answers, we calculate:

1. Expectations concentration as:

$$EC_i = \sum_{j=1}^{5} a^2 \tag{1}$$

   where a is the number of points to each slot.

2. Expectations concentration in just one scenario (the respondent allocates the 10 points to one scenario signalling total lack of uncertainty about housing price growth): the so-called ”bunching.

Results IV

Table: Factors Determining Expectations Concentration

	Expectations Concentration	Bunching
Verbatim Reading	-0.022***	-0.120***
	(0.004)	(0.018)
Int. shows Card	0.009	0.114**
	(0.005)	(0.053)
Reminder to Sum 10	-0.118***	-0.820***
	(0.016)	(0.135)
Speech Rate	0.049***	0.271***
	(0.005)	(0.033)
Induces	0.065***	0.493***
	(0.020)	(0.127)
Understanding	0.071***	0.176*
	(0.014)	(0.106)
Doubt	0.045***	0.147*
	(0.008)	(0.086)
Silence (secs.)	-0.095***	-0.734***
	(0.006)	(0.073)
Wave	Yes	Yes
Interviewer	Yes	Yes
Observations	10,970	11,383
R^2	0.241	
Pseudo R^2	0.363	0.167

Conclusion

- First paper to extract indicators about the elicitation process of subjective probabilistic expectations questions from automatic transcriptions of audio records collected in a F2F household survey.
- After controlling for household, wave and interviewer characteristics, there are significant associations between:
  - Interview protocol compliance and expectations concentration: the more the compliance (literal reading of verbatim, fluent and paused speech rate, etc.), the less concentrated are the expectations.
  - Household cognition and expectations allocation: the lower the understanding or levels of doubt, the greater the concentration. In addition, the more the household pauses to think about the answer, the lower the concentration.
- All these variables and correlations have the expected sign and pattern along multiple household dimensions (education and age levels specially).

Thank you

References I

Benkí, José, Jessica Broome, Frederick Conrad, Robert Groves and Frauke Kreuter. (2011). “Effects of speech rate, pitch, and pausing on survey participation decisions”. In *American Association for Public Opinion Research Annual Meeting, Phoenix, AZ*.

Bergmann, Michael, and Johanna Bristle. (2020). “Reading fast, reading slow: The effect of interviewers’ speed in reading introductory texts on response behavior”. *Journal of survey statistics and methodology*, 8, pp. 325–351. https://doi.org/10.1093/jssam/smy027

Bredin, Hervé, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz and Marie-Philippe Gill. (2019). “Pyannote.audio: Neural building blocks for speaker diarization”. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7124–7128. https://api.semanticscholar.org/CorpusID:207779942

References II

Conrad, Frederick G, Jessica S Broome, José R Benkí, Frauke Kreuter, Robert M Groves, David Vannette and Colleen McClain. (2013). "Interviewer speech and the success of survey invitations". *Journal of the Royal Statistical Society Series A: Statistics in Society*, 176 (1), pp. 191–210.

Defossez, Alexandre, Gabriel Synnaeve and Yossi Adi. (2020). "Real time speech enhancement in the waveform domain". In *Interspeech*.

D'Acunto, Francesco, and Michael Weber. (2024). "Why survey-based subjective expectations are meaningful and important". Working Paper, 32199, National Bureau of Economic Research. https://doi.org/10.3386/w32199

Fowler, F.J., and T.W. Mangione. (1990). *Standardized Survey Interviewing: Minimizing Interviewer-Related Error*. Applied Social Research Methods. SAGE Publications. https://books.google.es/books?id=gYyD-WUs35oC

References III

Kish, Leslie. (1962). “Studies of interviewer variance for attitudinal variables”. *Journal of the American Statistical Association*, 57 (297), pp. 92–115. http://www.jstor.org/stable/2282442

Mangione, Thomas W., Floyd J. Fowler and Thomas A. Louis. (1992). “Question characteristics and interviewer effects”. *Journal of Official Statistics*, 8 (3), p. 293. https://login.ezproxy-bde.greendata.es/login?url=https://www.proquest.com/scholarly-journals/question-characteristics-interviewer-effects/docview/1266807067/se-2

Ongena, Yfke P., and Wil Dijkstra. (2007). “A model of cognitive processes and conversational principles in survey interview interaction”. *Applied Cognitive Psychology*, 21 (2), pp. 145–163. https://doi.org/https://doi.org/10.1002/acp.1334

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey and Ilya Sutskever. (2022). “Robust speech recognition via large-scale weak supervision”. https://doi.org/10.48550/ARXIV.2212.04356

References IV

Schaeffer, Nora Cate, Jennifer Dykema, Dana Garbarski and Douglas W Maynard. (2008). “Verbal and paralinguistic behaviors in cognitive assessments in a survey interview”. In *Proceedings of the American Statistical Association, Section on Survey Research Methods. Alexandria, VA: American Statistical Association*.

Vandenplas, Caroline, Geert Loosveldt, Koen Beullens and Katrijn Denies. (2017). “Are Interviewer Effects on Interview Speed Related to Interviewer Effects on Straight-Lining Tendency in the European Social Survey? An Interviewer-Related Analysis”. *Journal of Survey Statistics and Methodology*, 6 (4), pp. 516–538. https://doi.org/10.1093/jssam/smx034

Wang, Changhan, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Miguel Pino and Emmanuel Dupoux. (2021). “Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation”. *ArXiv*, abs/2101.00390. https://api.semanticscholar.org/CorpusID:230433640

Table: Regular expressions used to search in the audios transcriptions

Variables	Regex Pattern
Reminder to Sum 10	`(?:sumar 10), (?:total 10), (?:sumar diez), (?:total diez)`
Int. shows Card	`(?:cartón), (?:Cartón),(?:carton), (?:Carton)`
Doubt	`(?:no lo sé), (?:no sé),(?:No lo sé), (?:No sé),` `(?:que se yo), (?:Que sé yo), (?:ni idea), (?:Ni idea)`
Understanding	`(?:no lo entiendo), (?:no entiendo), (?:No lo entiendo),` `(?:No entiendo)`
Induce	`(?:Entonces\b[^.]*?\bpongo\b\.?), (?:entonces\b[^.]*?\bpongo\b\.?),` `(?:Entonces\b[^.]*?\bponemos\b\.?), (?:entonces\b[^.]*?\bponemos\\b\.?),` `(?:Entonces\b[^.]*?\bpondría\b\.?), (?:entonces\b[^.]*?\bpondría\b\.?),` `(?:Ponemos\b[^.]*?\bentonces\b\.?), (?:ponemos\b[^.]*?\bentonces\b\.?),` `(?:Pongo\b[^.]*?\bentonces\b\.?), (?:pongo\b[^.]*?\bentonces\b\.?),` `(?:Pondría\b[^.]*?\bentonces\b\.?), (?:pondría\b[^.]*?\bentonces\b\.?)`