



Transcribing and coding voice answers obtained in web surveys: comparing three leading automatic speech recognition tools

AAPOR, Los Angeles (USA), May 2026

Melanie Revilla, Carlos Ochoa, **Jan Karem Höhne**, & Mick P. Couper

Acknowledgments:

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 849165).

We thank Maria Paula Acuña Pardo for her hard work on coding the questions, and Ixchel Perez Duran for her help in preparing and analyzing the survey.

Using voice answers in web surveys

Researchers

Technical problems

Selection bias

Data protection & ethical issues

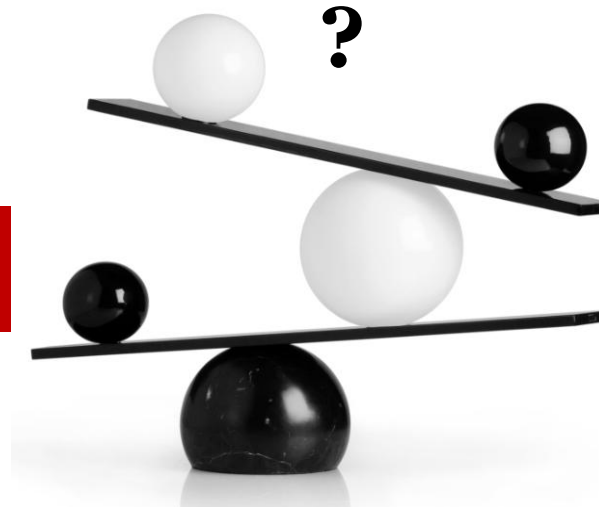
Transcription & coding

Disadvantages

Participants

Privacy issues

Context not conducive to speaking



Longer/richer answers

More spontaneous

Emotions, tone of voice, etc.

Advantages

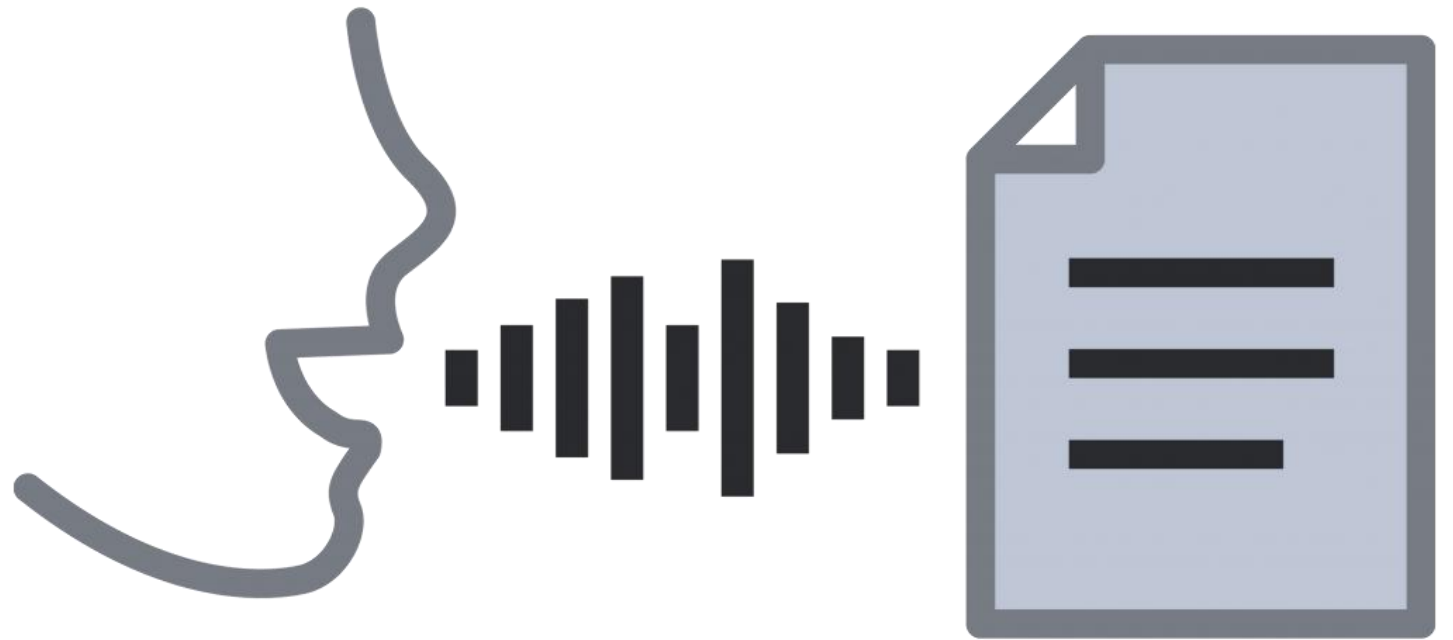
Reduced time/effort to provide information

No need for writing skills

Researchers

Participants

This study



Transcription

RQ1: How do 3 leading automatic speech recognition (ASR) tools perform across various dimensions?



Google Cloud
Speech-to-Text



Whisper
OpenAI



Coding

RQ2: How similar or different are the codes of transcribed responses generated by a human and the OpenAI GPT-4o model?



VS



GPT-4o

Data collection & sample

Data collection

Opt-in online panel, February/March 2024

netquest
a Bilendi company

Tool

Population of
interest

Quotas

Sample used

Data collection & sample

Data collection

Opt-in online panel, February/March 2024



Tool

[WebdataVoice](#) (Revilla et al., 2022)

Population of
interest

Quotas

Sample used

Data collection & sample

Data collection

Opt-in online panel, February/March 2024



Tool

[WebdataVoice](#) (Revilla et al., 2022)

**Population of
interest**

Adult online population living in Spain

Quotas

Sample used

Data collection & sample

Data collection

Opt-in online panel, February/March 2024



Tool

[WebdataVoice](#) (Revilla et al., 2022)

Population of
interest

Adult online population living in Spain

Quotas

Gender & age + education

Sample used

Data collection & sample

Data collection

Opt-in online panel, February/March 2024



Tool

[WebdataVoice](#) (Revilla et al., 2022)

Population of
interest

Adult online population living in Spain

Quotas

Gender & age + education

Sample used

Those with at least one transcription = 859 panellists

Survey about perceptions of nursing homes in Spain

80+ questions



2 requests for voice



Focus on first one

WHY TRANSP

Explain why you think that nursing homes provide [*no information at all/very little/some/a lot of/a huge **amount of information***] about the implementation of their services.

Please give as much detail as you can. In your answer, mention if you think there is a difference among **public** and **private** nursing homes.

Transcription

Block 1 (each ASR tool)

- 🗣️ Transcription provided
- 🔍 Length
- 🗣️ Clarity
- 🗣️ Problems
- 🗣️ Valid

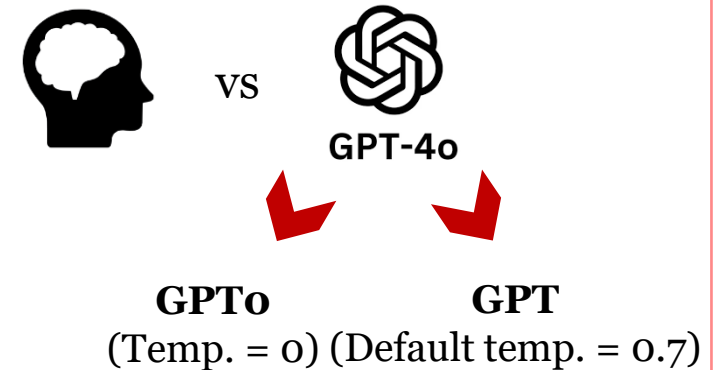
Block 2 (pairs)

- 🔍 Different words
- 🗣️ Similar meaning

Coding

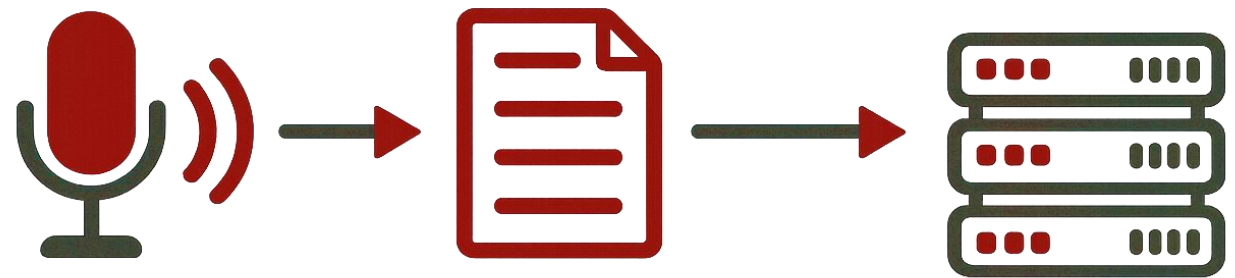
Subset of indicators of *RQ1*

Compare:



⚠️ No “true value”

Main results



Transcription: Comparing the 3 ASR tools

Block 1	
Provided	% Provided
Length	# Characters # Words
Clarity	% Clear % Very clear
Problems	% 1+ Missing % 1+ Added % 1+ Wrong % No problem
Valid	% Valid

N^{“Provided”} = 859. N^{“Other indicators”} = 636 for Google, 857 for Whisper, 773 for Vosk.
Superscripts indicate a significant difference (p < 0.05).

Transcription: Comparing the 3 ASR tools

Block 1		Google (a)
Provided	% Provided	74 ^{bc}
Length	# Characters	285 ^c
	# Words	51 ^c
Clarity	% Clear	5 ^{bc}
	% Very clear	94 ^{bc}
Problems	% 1+ Missing	9 ^{bc}
	% 1+ Added	7 ^{bc}
	% 1+ Wrong	10 ^{bc}
	% No problem	77 ^{bc}
Valid	% Valid	97 ^{bc}

N_{“Provided”} = 859. N_{“Other indicators”} = 636 for Google, 857 for Whisper, 773 for Vosk.
Superscripts indicate a significant difference ($p < 0.05$).

G
O
O
G
L
E

- Only 74% have a transcription
- However, when there is a transcription, it performs strongly:
 - 94% very clear
 - 77% with no detected problem

Transcription: Comparing the 3 ASR tools

Block 1		Google (a)	Whisper (b)
Provided	% Provided	74 ^{bc}	100 ^{ac}
Length	# Characters	285 ^c	310 ^c
	# Words	51 ^c	55 ^c
Clarity	% Clear	5 ^{bc}	22 ^{ac}
	% Very clear	94 ^{bc}	73 ^{ac}
Problems	% 1+ Missing	9 ^{bc}	19 ^{ac}
	% 1+ Added	7 ^{bc}	28 ^{ac}
	% 1+ Wrong	10 ^{bc}	5 ^{ac}
	% No problem	77 ^{bc}	57 ^{ac}
Valid	% Valid	97 ^{bc}	80 ^{ac}

N_{“Provided”} = 859. N_{“Other indicators”} = 636 for Google, 857 for Whisper, 773 for Vosk.
Superscripts indicate a significant difference ($p < 0.05$).

G
O
O
G
L
E

- Only 74% have a transcription
- However, when there is a transcription, it performs strongly:
 - 94% very clear
 - 77% with no detected problem

W
H
I
S
P
E
R

- Almost all have a transcription
- But high levels of missing (19%) and added words (28%), and low rates of valid answers (80%)
 - Hallucinations?

Transcription: Comparing the 3 ASR tools

Block 1		Google (a)	Whisper (b)	Vosk (c)
Provided	% Provided	74 ^{bc}	100 ^{ac}	90 ^{ab}
Length	# Characters	285 ^c	310 ^c	386 ^{ab}
	# Words	51 ^c	55 ^c	69 ^{ab}
Clarity	% Clear	5 ^{bc}	22 ^{ac}	72 ^{ab}
	% Very clear	94 ^{bc}	73 ^{ac}	11 ^{ab}
Problems	% 1+ Missing	9 ^{bc}	19 ^{ac}	14 ^{ab}
	% 1+ Added	7 ^{bc}	28 ^{ac}	13 ^{ab}
	% 1+ Wrong	10 ^{bc}	5 ^{ac}	90 ^{ab}
	% No problem	77 ^{bc}	57 ^{ac}	8 ^{ab}
Valid	% Valid	97 ^{bc}	80 ^{ac}	94 ^{ab}

N_{“Provided”} = 859. N_{“Other indicators”} = 636 for Google, 857 for Whisper, 773 for Vosk.
Superscripts indicate a significant difference ($p < 0.05$).

G
O
O
G
L
E

- Only 74% have a transcription
- However, when there is a transcription, it performs strongly:
 - 94% very clear
 - 77% with no detected problem

W
H
I
S
P
E
R

- Almost all have a transcription
- But high levels of missing (19%) and added words (28%), and low rates of valid answers (80%)
 - Hallucinations?

V
O
S
K

- Intermediate level of transcriptions (90%) → Closer to true answer rate?
- Longest length (386 characters)
- Punctuation issues → Low clarity (11%)

Transcription: Comparing pairs of tools

Block 2	
Diff. words	# Perc. diff words
Similar meaning	% Partly similar
	% Very similar

Superscripts indicate a significant difference ($p < 0.05$).

Transcription: Comparing pairs of tools

Block 2		G-W (a)	W-V (b)	V-G (c)
Diff. words	# Perc. diff words	19 ^{bc}	33 ^{ac}	24 ^{bc}
Similar meaning	% Partly similar	10 ^{bc}	26 ^a	22 ^a
	% Very similar	77 ^b	59 ^{ac}	73 ^b

Superscripts indicate a significant difference ($p < 0.05$).

G
-
W

Google & Whisper have the **smallest** word discrepancies and the **highest** level of similarity

Transcription: Comparing pairs of tools

Block 2		G-W (a)	W-V (b)	V-G (c)
Diff. words	# Perc. diff words	19 ^{bc}	33 ^{ac}	24 ^{bc}
Similar meaning	% Partly similar	10 ^{bc}	26 ^a	22 ^a
	% Very similar	77 ^b	59 ^{ac}	73 ^b

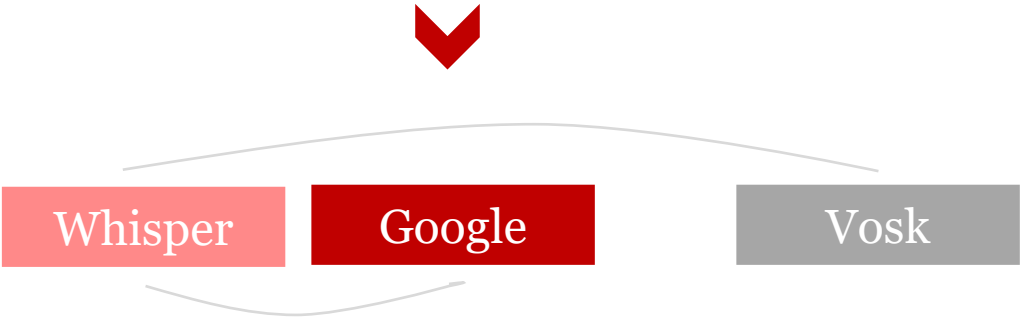
Superscripts indicate a significant difference (p < 0.05).

G
-
W

Google & Whisper have the **smallest** word discrepancies and the **highest** level of similarity

W
-
V

Whisper & Vosk exhibit the **largest** word discrepancies, and show the **lowest** level of similarity



Coding: Comparing human and GPT-4o coding

Block 1	
Clarity	% Clear % Very clear
Problems	% 1+ Missing % 1+ Added % 1+ Wrong % No problem
Valid	% Valid

Superscripts indicate a significant difference ($p < 0.05$), within ASR tool.

Coding: Comparing human and GPT-4o coding

Block 1		Google		Whisper		Vosk	
		(b) GPTo	(c) GPT	(b) GPTo	(c) GPT	(b) GPTo	(c) GPT
Clarity	% Clear	31 ^a	30 ^a	24	22	59 ^a	58 ^a
	% Very clear	66 ^a	66 ^a	68 ^a	69 ^a	7 ^a	7 ^a
Problems	% 1+ Missing	41 ^a	43 ^a	28 ^a	27 ^a	98 ^a	97 ^a
	% 1+ Added	36 ^{ac}	38 ^{ab}	28	29	78 ^a	78 ^a
	% 1+ Wrong	33 ^a	34 ^a	11 ^a	11 ^a	91	91
	% No problem	44 ^{ac}	41 ^{ab}	56	55	2 ^a	2 ^a
Valid	% Valid	85 ^a	84 ^a	66 ^a	66 ^a	75 ^a	74 ^a

Superscripts indicate a significant difference ($p < 0.05$), within ASR tool.

G
P
T
o
-
G
P
T

Few significant differences between GPT codings, and only for Google

→ Setting the temperature to 0 does not substantially affect the results

Coding: Comparing human and GPT-4o coding

Block 1		Google			Whisper			Vosk		
		(a) Human	(b) GPTo	(c) GPT	(a) Human	(b) GPTo	(c) GPT	(a) Human	(b) GPTo	(c) GPT
Clarity	% Clear	5 ^{bc}	31 ^a	30 ^a	22	24	22	72 ^{bc}	59 ^a	58 ^a
	% Very clear	94 ^{bc}	66 ^a	66 ^a	73 ^{bc}	68 ^a	69 ^a	11 ^{bc}	7 ^a	7 ^a
Problems	% 1+ Missing	9 ^{bc}	41 ^a	43 ^a	19 ^{bc}	28 ^a	27 ^a	14 ^{bc}	98 ^a	97 ^a
	% 1+ Added	7 ^{bc}	36 ^{ac}	38 ^{ab}	28	28	29	13 ^{bc}	78 ^a	78 ^a
	% 1+ Wrong	10 ^{bc}	33 ^a	34 ^a	5 ^{bc}	11 ^a	11 ^a	90	91	91
	% No problem	77 ^{bc}	44 ^{ac}	41 ^{ab}	57	56	55	8 ^{bc}	2 ^a	2 ^a
Valid	% Valid	97 ^{bc}	85 ^a	84 ^a	80 ^{bc}	66 ^a	66 ^a	94 ^{bc}	75 ^a	74 ^a

Superscripts indicate a significant difference ($p < 0.05$), within ASR tool.

G
P
T
o
-
G
P
T

Few significant differences between GPT codings, and only for Google

→ Setting the temperature to 0 does not substantially affect the results

H
-
G
P
T

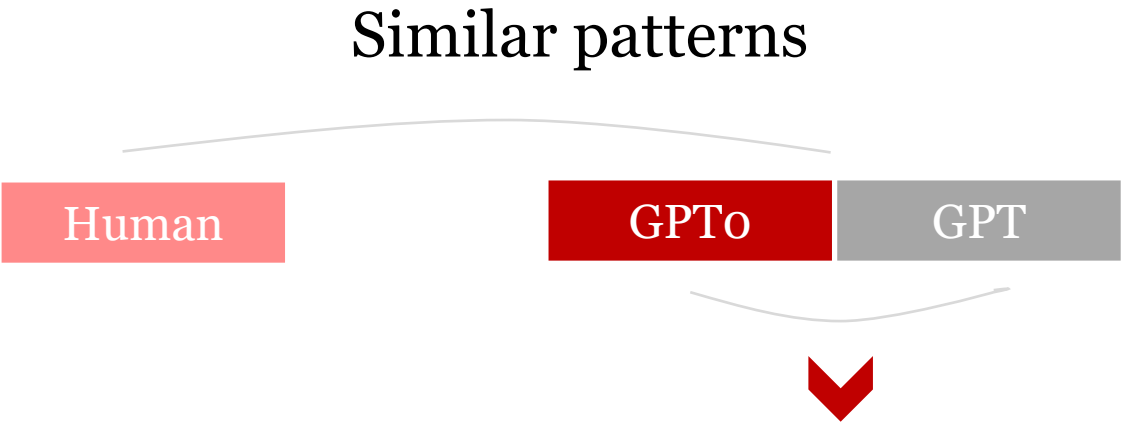
Significant and substantial differences between human and GPT codings across all 3 ASR tools

However, they do not always alter the order of performance across ASR tools

Coding: Comparing human and GPT-4o coding

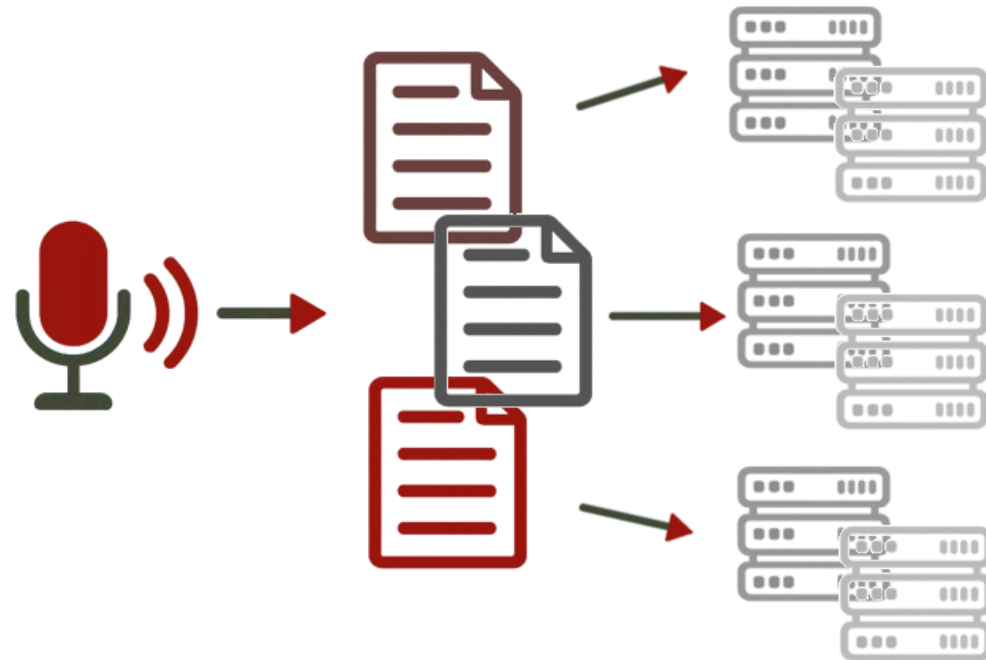
Block 2		Google-Whisper			Whisper-Vosk			Vosk-Google		
		(a) Human	(b) GPTo	(c) GPT	(a) Human	(b) GPTo	(c) GPT	(a) Human	(b) GPTo	(c) GPT
Similar meanings	% Partly similar	10 ^{bc}	6 ^a	7 ^a	26 ^{bc}	9 ^a	11 ^a	22 ^{bc}	12 ^a	13 ^a
	% Very similar	77	75	75	59 ^{bc}	65 ^a	64 ^a	73	76	75

Superscripts indicate a significant difference ($p < 0.05$), within a pair of tools.



In general, GPT codings suggest a lower quality of the transcriptions

Conclusions



Summary of results

Each of the
ASR tools has
distinct
merits and
limits

+

Human and
LLM-based
coding also
differ
significantly

=

Use **several** ASR
tools & implement
human + LLM-based
coding



Google sometimes fails to provide
transcriptions but shows high clarity

Whisper almost always provides
transcriptions but has hallucinations

Vosk has the longest length but clarity
issues and high rates of incorrect words



Setting the temperature
to 0 does not affect much



Additional information at
minimal added cost

Conclusions

Performance of ASR tools and GPT models **evolves rapidly** + **vary** with factors like ...

- ... language and background noise for the ASR tools
- ... specific prompts or settings for GPT



Researchers should be prepared to **adjust** to these changes and potentially incorporate newly emerging tools

Key recommendations likely to remain effective across contexts and over time

1. Carefully select the ASR tools when transcribing voice responses
2. Use multiple transcriptions to mitigate each ASR tool's weaknesses
3. Test different settings for each tool
4. Document the choices made to ensure transparency and replicability

Thanks!

Questions?

Paper accepted in JSSAM:
<https://doi.org/10.1093/jssam/smafo28>



melanie.revilla@upf.edu



<https://www.upf.edu/web/webdataopp>