Do audio and visual tokenizers capture backchannels?

Benoit Favre
Aix Marseille Univ, CNRS, LIS
Université Grenoble Alpes, CNRS, LIG
first.last@lis-lab.fr

Auriane Boudin
Aix Marseille Université
CNRS, LPL, Marseille, France
first.last@univ-amu.fr

Abstract

Audio and video tokenizers are autoencoders trained to represent the content of recordings as a sequence of vectors. They are prevalently used to interface large language models with non-textual modalities. While they allow advanced applications such as video generation, the envelope of their limitations is not known in the context of multimodal conversation. This work focuses on backchannels, which listeners use to signal to the speaker that they are listening. This feedback is essential to maintain the conversation flow. We evaluate whether a representative set of audio and video tokenizers encode backchannels using linear probing. Results show that although audio tokenizers capture the phenomenon relatively well, backchannels are not linearly separated by video tokenizers. However, joint representations resulting from concatenating representations in both modalities improve accuracy significantly over audio-only representations, suggesting to train multimodal tokenizers.

1 Introduction

Backchannels are an important feature of conversations, allowing a listener to regularly give feedback to the speaker and to show understanding and interest (Schegloff, 1982; Bavelas et al., 2000). They consist in head movements such as nods, smiles, frowning, short verbal unit insertions such as "hmm", "yeah", "okay". They help to regulate conversation flow and communicate engagement, attention or agreement.

Backchannels constitute a complex phenomenon due to significant variability in their frequency, timing, and modality of production. Although numerous opportunities for backchanneling emerge during interaction, only a limited subset is actually realized by interlocutors. Moreover, as an intrinsically multimodal behavior, backchannels can be expressed through auditory, visual, or both modalities. This inherent stochasticity makes them particularly challenging to detect, predict, and generate in dialogue systems. Nonetheless, their production is crucial for supporting and maintaining high-quality conversation. Their absence or inappropriate production can collapse the conversational flow in dialog systems, and decrease dialog naturalness.

Past work has focused on high-level features extracted by specialized models, such as body and head pose estimation, face landmark estimation, or verbalized unit detection in automatic transcripts. Due to the different nature of audio and video features, multimodal models often rely on late fusion, which does not captureT- well cross-modal interactions, and is sensitive to cascading errors (such as lack of detected face).

Recent work on end-to-end generative audio models, such as dGSLM (Nguyen et al., 2022), has shown that indirect modeling of conversation phenomena through self-supervision allows for generating natural-sounding continuations of conversations including speaker identity, turn taking and backchanneling. In particular, dGSLM proposed training a next-token predicting language model to generate discrete units from the HuBERT masked transformer (Hsu et al., 2021) and synthesize audio samples from these units. Since then, a range of audio "tokenizers" have been proposed, which can be fed to large language models (LLMs) in order to account for the audio modality, both as input and output, resulting in promising dialog systems.

Recent developments of video generation models have lead to the emergence of video tokenizers, trained to compress video sequences as a set of discrete tokens that can be fed to LLMs, and detokenized back to a sequence of images as generation output. Like audio tokenizers, video tokenizers can potentially replace the standard video feature extraction stages and generate audio-visual dialog continuations. Yet, the emergence of dialog-related capabilities such as backchannelling through pretraining is not well studied.

Proceedings of the 16th International Workshop on Spoken Dialogue Systems Technology, pages 64–75
February 26–March 1, 2026.

In this paper, we test the capacity of audio and video tokenizers to represent backchannels in the framework of linear probing. Our goal is not to obtain good backchannel detectors, but rather to understand how pretraining captures the phenomenon. Our contributions are: (1) an assessment of activation linear separability with respect to backchannels, from a set of audio and video models, and their combination; (2) a study of linguistic features correlation with backchannel detection success/failure in linear probing; (3) a discussion of the pretraining choices impacting the detection of fine-grained conversational events in both modalities. All data, code and models are made available[1].

2 Related work

Backchannel detection has a long history of research in the audio modality, based on explicit lexical and prosodic features (Noguchi and Den, 1998; Vinciarelli et al., 2008; Al Moubayed et al., 2009; de Kok and Heylen, 2012; Mueller et al., 2015; Ruede et al., 2017; Kholiavin et al., 2020; Amer et al., 2023). Although backchannels have been annotated mainly as part of dialog act annotation efforts on large speech datasets, such as Switchboard or ICSI meeting recordings (Jurafsky et al., 1998; Shriberg et al., 2004), there also exist a number of efforts to annotate backchannels in both the visual and audio modalities (Bertrand et al., 2007; Degutyte and Astell, 2021; Blomsma et al., 2024; Boudin et al., 2021). The larger available multimodal corpora include MPIIGroupInteraction (Muller et al., 2022), Cup of CoFee (Prévot et al., 2016), NOXI (Cafaro et al., 2017), Vyaktitv (Jain et al., 2021), SMYLE, (Boudin et al., 2023), IFADV (Truong et al., 2011), or Chico (Bodur et al., 2021).

Audio tokenizers have been developed in the framework of self-supervision for speech applications such as ASR. They typically start from temporal or spectral representations of the speech signal (such as Mel filterbanks), and encode the speech signal with a variety of neural architectures. Initial approaches include contrastive predictive coding (van den Oord et al., 2018), autoregressive predictive coding (Chung and Glass, 2019), and HuBERT (Hsu et al., 2021) which iteratively trains an encoder to predict masked units originating from k-means clustering of previous iteration representations. While pervasive, theses approaches are being replaced with autoencoders trained to regenerate audio from latent discrete representations. They use VQ-VAEs to learn discrete representations and exploit a number of losses to ensure that frequencies and dynamics are preserved (Kong et al., 2020), they maintain high fidelity with Residual vector quantization (Défossez et al., 2022), and distill higher level representations in order to preserve semantics (Zhang et al., 2023). Examples of influencial tokenizers include VQ-wav2vec (Baevski et al., 2019), SoundStream (Zeghidour et al., 2021), BestRQ (Chiu et al., 2022), Encodec (Défossez et al., 2022), BEATs (Chen et al., 2022), Data2Vec (Baevski et al., 2022), SpeechTokenizer (Zhang et al., 2023), WavTokenizer (Ji et al., 2024).

Video tokenizers aim at producing discrete or continuous representations from a sequence of video frames. They mostly follow the compression paradigm where an auto-encoder is trained to encode video frames as a latent representation which is then decoded to reconstruct the sequence of images. Finite scalar quantization (FSQ) replaces VQ-VAE as it is more stable at training (Mentzer et al., 2023). They include diverse neural architectures and resort to factorizing spatio-temporal relationships to decrease computation costs compared to a full 3D analysis. They might be initialized with 2D encoders from image generation models (Zheng et al., 2024), or are jointly trained on single images and videos in order to benefit from the diversity and quantity of image datasets (Wang et al., 2024). Examples include VideoGPT-Tokenizer (Yan et al., 2021), VideoGPT+ (Maaz et al., 2024), Omni-Tokenizer (Wang et al., 2024), CogVideoX-Tokenize (Yang et al., 2024), OpenSora (Zheng et al., 2024), OpenSora-Plan (Lin et al., 2024), CV-VAE (Zhao et al., 2024), VidTok (Tang et al., 2024), Cosmos-Tokenizer (Agarwal et al., 2025).

Probing of audio and vision models behavior is a very active research area. It consists in analyzing model weights, activations or performances in a particular domain in order to explain observed behavior. In conversation analysis, a number of studies have shown that audio models represent known language structures although they have not been explicitly trained to recognize them (Shah et al., 2021; Martin et al., 2023; Pasad et al., 2023; Ashihara et al., 2023; Ngo and Kim, 2024). Similar patterns have been uncovered in the vision modality (Caron et al., 2021; Vanyan et al., 2023; Kumar et al., 2023;

[1]Will be made public after review

Basaj et al., 2021). Although multimodal scenes correspond to correlated audio and video signals, audio and video tokenizers are trained disjointly. Their representation capabilities are infrequently evaluated, in particular in the context of multimodal conversation.

3 Experimental setup

3.1 Linear probing

Linear probing consists in assessing whether a representation space is linearly separable according to a classification task. Although it is more strict than measuring whether the input contains information about the task, one does not have to find the most effective non-linear model for extracting that information from a potentially infinite set of models. It is important to note that we want to assess whether pretraining results in this linear separability, not whether the information is present and could be used by a more general classifier.

In the following, we train a logistic regression on the output of the tokenizers. Representations are extracted by feeding speech or video corresponding to the evaluated segment to the tokenizers, resulting in a sequence of vectors that are then averaged along the time dimension. Other pooling methods have been proposed but they typically add many parameters or assess a different property than linear separability, such as with RNNs or Echo-state networks (Sun et al., 2024). If we assume that consecutive phenomena occupy different dimensions in the representation space, then average pooling keeps the underlying information in linearly separable form. For tokenizers trained with discrete latent representations, we use the embeddings of the discrete tokens.

3.2 Tokenizers

We have selected four representative audio tokenizers with varying architectures and training data:

- **HuBERT** (Hsu et al., 2021), trained with masked prediction of discrete tokens resulting from a k-means clustering of the underlying acoustic space. In a first training iteration, clusters are generated from MFCCs, and in a second iteration, they are generated from hidden representations of the first iteration model. Variants of HuBERT (base, large) are trained respectively on Librispeech (1k hours) and and Librilight (60k hours).

- **SpeechTokenizer** (Zhang et al., 2023), an autoencoder trained to reconstruct the speech signal. It leverages RVQ-GANs which iteratively quantize the residual of previous quantization stages. It also adds a distillation loss from HuBERT in order to capture high-level information. SpeechTokenizer is trained on Librispeech (1k hours) and Common Voice (31k hours).

- **Wavtokenizer** (Ji et al., 2024), a RQ-GAN autoencoder similar to SpeechTokenizer, but which foregoes the residual quantizer in order to reduce the number of tokens per second. It increases the codebook size and changes the decoder to obtain higher reconstruction fidelity. It is trained on 8k hours of speech.

- **Mimi** (Défossez et al., 2024), is also an autoencoder from the VQ-VAE family, but it conditions the generation of fine-grained residual tokens on higher-level tokens at the same timestamp in order to parallelize processing. Authors do not specify the exact training data but mention training on 7m hours of speech.

Video tokenization is a developing field, therefore we selected a set of tokenizers which were available and sufficiently documented:

- **Cosmos** (Agarwal et al., 2025): A causal autoencoder trained on reconstruction, perceptual, optical flow and Gram-matrix losses. It consists of 3D Haar Wavelet layers followed by residual blocks that perform spatio-temporal factorized 3D convolution and downsampling. The last block uses self-attention in order to account for long-range dependencies. Discrete variants of the model (DV) rely on FSQ for quantization. It is trained on 100M clips from 49 to 121 frames with varying resolution and frame rate, on both still pictures and videos from undisclosed data sources.

- **VidTok** (Tang et al., 2024): A causal VAE with continuous and discrete variants trained with FSQ. The architecture differs from Cosmos in that it includes full 3D convolutions in addition to factorized spatio-temporal convolutions, and an "alpha-blender" module which downsamples temporal resolution via weighted averaging. The model is trained with Latent diffusion losses (Rombach et al., 2021),

including a KL regularizer for continuous variants, and a commitment loss for discrete variants. VidTok is trained on 10M clips of 17 frames, from undisclosed sources.

- **OpenSora** (Zheng et al., 2024): A replication of the Sora work. The model decorrelates spatial and temporal compression, first applying a 2D VAE initialized with the SDXL VAE (Podell et al., 2023), and then applying a 3D VAE to compress in the time dimension. It is trained on 30M instances of 17 frames, (80k hours of video) with progressive introduction of more difficult instances. The variant we work with is version 1.2.

In our experiments, we use the native frame rate of the videos (30 fps) and average the representations of consecutive sets of frames that fit each model's context size.

3.3 The SMYLE dataset

We used a subset of the French multimodal SMYLE corpus (illustrated in Figure 1), consisting of 25 dyads (50 participants) with feedback annotations (Boudin et al., 2024). We selected this corpus for its controlled video recording conditions. SMYLE includes face-to-face interactions across two tasks: a storytelling task, where one participant narrates three types of stories[2] to a listener, followed by 15-minute free conversation.

Figure 1: Illustration of the SMYLE dataset, featuring a screenshot of both participants videos side by side.

Feedback was annotated into generic and specific types following the framework proposed by Bavelas et al. (2000). Here, feedback refers to any reaction from one speaker to the other (excluding responses to explicit questions) and includes vocal, verbal, or gestural cues. *Generic* feedback encompasses brief vocalizations (e.g., "*mh mh*," "*ok*"), nodding, and smiling, typically used to signal understanding and encourage the speaker to continue. In contrast, *specific* feedback involves more expressive and evaluative responses, which may include speech (e.g., completions, repetitions, reformulations, humor, etc.), as well as laughter and various gestures, such as facial expressions, head movements, and hand gestures. This annotated subset totals 13.4 hours of interaction (7.04 hours of storytelling and 6.36 of free conversation) and includes 6,285 instances of generic feedback (3,470 from the storytelling task and 2,815 from the free conversation part). This makes the SMYLE feedback subset twice the size of that in the MPIIGroupInteraction corpus (Muller et al., 2022).

In addition to feedback annotations, the SMYLE corpus includes other manual annotations such as head movements (nods, shakes, tilts, and others), laughter, and orthographic transcriptions. Acoustic features, including pitch (F0) and intensity, were automatically extracted using the OpenSmile toolkit with the eGeMAPS feature set (Eyben et al., 2010, 2016)[3]. Features were computed using a sliding window of 0.08 s with a 0.04-second step, resulting in values every 40 ms. We use these annotations to better understand how tokenizers capture backchannels.

3.4 Backchannels and non-events

Since our work addresses backchannels, and no prior work has addressed the detection of both specific and generic feedback, we concentrate here on detecting generic feedback, which closely aligns with the concept of backchannels. In the following, backchannel events were extracted from the generic feedback annotation of the dataset. In order to stabilize tokenizer behavior, we clipped segments longer than 2000 ms and extended segments shorter than 500 ms while keeping their start time, reducing chance of overlap with a different event. Each selected backchannel event was paired with a random non-event segment of equal duration, within the same video. Backchannels not overlapping with verbal or visual annotations were dropped. This process yielded 6,025 backchannel segments (also referred to as *events*) and 6,025 matched non-backchannel segments (referred to as *non-events*), resulting in a total of 12,050 segments.

Visual, lexical, and prosodic characteristics

[2]Narratives: (1) retelling the content of a video clip; (2) summarizing the plot of a movie, book, or video game; and (3) describing favorite holiday.

[3]Available at https://www.audeering.com/research/opensmile/, using the pipeline at https://github.com/MatthisHoules/opensmile_feature_extractor.

of the segments are summarized in Tables 1 and 2, which present descriptive statistics for both backchannels and non-events. Table 1 shows the proportion of segments that contain different types of head movements, vocal activity (silence, speech, laughter), and interjections (e.g., "yeah," "ok," etc.). Table 2 reports the mean, standard deviation for pitch (F0) and loudness, computed only on segments containing speech. All features reported in these tables—whether visual, lexical, or prosodic—can occur both during main speaker turns and during feedback, which contributes to the task difficulty.

Type	Features	Backchannel	Other
Head	Nod	89.23	16.15
	Shake	1.29	8.94
	Tilt	1.46	5.21
	Other	0.35	1.39
Activity	Silence	98.82	79.25
	Speech	43.62	48.33
	Laughter	0.77	4.87
Interjections	*ouais* / yeah	30.47	27.24
	d'accord / fine	3.13	0.76
	ok / okay	43.30	29.76
	hm / hm hm	11.19	1.43
	oui / yes	17.32	25.88
	non / no	0.27	3.13
	all inter.	43.34	29.79

Table 1: Proportion (%) of visual and lexical features for backchannels and non-event segments. Each value indicates the percentage of segments in which the corresponding feature was annotated. Note that most events affect both the audio and video modality.

Type	Features	Backchannel	Other
F0	Mean	92.65	115.29
	SD	68.05	64.12
Loudness	Mean	0.65	1.41
	SD	0.52	0.72

Table 2: Average values of F0 and loudness for vocalized backchannels and other events. For each segment, the mean and standard deviation (SD) were computed and then averaged across all segments in each class.

4 Results and discussion

For each type of representation, we train a logistic regression to discriminate between backchannel events and non-events. Results are averaged over a 30-fold split of the events. For each split, we randomly sample 5k events for training the regression in order to account for both test set and training set variability. Each linear model is trained via gradient decent implemented in Pytorch, and randomly initialized to account for training variation. We report significance with the two-sided t-test over accuracy means, at a level of 10^{-3}.

4.1 Accuracy results

Audio	Video	Acc.	t-test
HuBERT	-	**0.734**	=
SpeechTokenizer	-	0.731	>
Mimi	-	0.695	>
WavTokenizer	-	0.641	>
Chance	-	0.5	
-	VidTok	**0.626**	>
-	Cosmos	0.574	>
-	OpenSora	0.504	=
-	Chance	0.5	

Table 3: Backchannel detection accuracy of monomodal probes. The t-test column indicates the significance of the difference between the result on the current row and the result on the next row at the 10^{-3} level. For example, HuBERT is not significantly better than SpeechTokenizer.

Table 3 shows the accuracy of the linear probe on audio and video representations. When multiple model variants are available, we select the best variant according to results in Tables 5 and 6. Audio representations lead to higher performance than video representations although the majority of backchannels contain nodding (89%, Table 1), a strong visual cue. HuBERT is better at detecting backchannels than Mimi (trained on several order of magnitude more data), which probably compromises high-level phenomena for better reconstruction accuracy. In the visual modality, VidTok representations are better than Cosmos representations, which might be explained by its use of non-factored 3D convolutions, although no real conclusion can be drawn without controlling their respective training data. The OpenSora probe is not better than chance, which indicates that although the model is able to generate compelling videos from its hidden representations, visual phenomena linked to backchannels are not linearly separable. Figure 2 shows that reconstructed videos sometimes lack precise details which are important for conversation understanding.

Table 4 focuses on multimodal results obtained by training the linear probe on concatenated audio and video representations. Combinations that involve VidTok are systematically significantly bet-

Figure 2: Example of original frame (left) and regenerated frame (right) from Cosmos. Although global picture is faithful, the model failed to capture the correct mouth and eye behavior.

Audio	Video	Acc.	t-test
HuBERT	VidTok	**0.785**	$>$
HuBERT	Cosmos	0.722	$=$
HuBERT	OpenSora	0.705	$=$
Mimi	VidTok	**0.796**	$>$
Mimi	Cosmos	0.728	$=$
Mimi	OpenSora	0.696	$=$
SpeechTokenizer	VidTok	**0.818**	$>$
SpeechTokenizer	Cosmos	0.758	$=$
SpeechTokenizer	OpenSora	0.736	$=$
WavTokenizer	VidTok	**0.766**	$>$
WavTokenizer	Cosmos	0.696	$>$
WavTokenizer	OpenSora	0.638	$=$

Table 4: Multimodal probe accuracy by concatenating monomodal representations. The t-test column indicates significance of the difference to the audio-only model of the pair. $>$ indicates that the result is significantly better at the 10^{-3} level.

ter than corresponding audio probe with accuracy improvements of 3-7 points, suggesting that the probe can account for complementary information in both modalities. VidTok's performance might be related to the quantity of spontaneous conversations in its training data. The best combination is SpeechTokenizer representations concatenated with with VidTok representations with an accuracy of 81.8%. Those results suggest that it might be beneficial for conversation processing to jointly train audio and video tokenizers, so that they can learn the intricate synchronization of fine-grained conversational behavior.

Tables 5 and 6 present probing accuracy for variants of models made available by their authors. In the audio modality, larger models trained or more diverse data, such as HuBERT-large or SpeechTokenizer-snake, lead to higher probe accuracy. In the video modality, higher compression (i.e. 8x16x16 vs 8x8x8) tends to decrease accuracy, continuous representations are better than discrete representations (CV vs DV for Cosmos; kl vs FSQ for VidTok), and more channels or larger

Audio model	Acc.
hubert-large-ll60k	**0.734**
hubert-base-ls960	0.721
WavTokenizer_small_320_24k_4096	**0.641**
wavtokenizer_medium_speech_320_24k	0.633
WavTokenizer_small_600_24k_4096	0.628
wavtokenizer_large_speech_320_24k	0.623
SpeechTokenizer_snake	**0.731**
SpeechTokenizer_hubert_avg	0.712
mimi	0.695

Table 5: Accuracy of audio model variants. The selected variant is denoted in bold. Details on variants are provided in Appendix A.

codebooks are correlated with better accuracy. Interestingly, the Cosmos-1.0 variant (vs 0.1), trained on more data with a larger context size is not able to better capture backchannels in the linear probing sense. It would be interesting to carefully assess those parameters for a range of speech phenomena.

Video model	Acc.
Cosmos-1.0-Tokenizer-CV8x8x8	**0.574**
Cosmos-0.1-Tokenizer-CV8x8x8	0.574
Cosmos-0.1-Tokenizer-CV8x16x16	0.565
Cosmos-0.1-Tokenizer-CV4x8x8	0.543
Cosmos-0.1-Tokenizer-DV8x16x16	0.531
Cosmos-0.1-Tokenizer-DV8x8x8	0.522
Cosmos-1.0-Tokenizer-DV8x16x16	0.514
Cosmos-0.1-Tokenizer-DV4x8x8	0.514
vidtok_kl_causal_488_16chn	**0.626**
vidtok_kl_causal_488_8chn	0.594
vidtok_fsq_causal_488_262144	0.559
vidtok_fsq_causal_488_32768	0.546
vidtok_kl_causal_488_4chn	0.529
vidtok_fsq_causal_41616_262144	0.500
vidtok_fsq_causal_488_4096	0.500
vidtok_kl_causal_41616_4chn	0.500
OpenSora-1.2	0.504

Table 6: Accuracy of video model variants. The selected variant is denoted in bold. Details on variants are provided in Appendix A.

4.2 Correlation with linguistic variables

To better understand model performance and to identify which audio, visual, and lexical characteristics were associated with prediction success, we computed Pearson correlations between a set of visual, lexical, and prosodic features (described in Section 3.3) and a binary success variable (coded as 1 for correct predictions and 0 for incorrect ones). Positive correlations indicate that a feature was more likely to be present, or to take on higher values, when the model made a correct predic-

tion. Negative correlations indicate that the feature tended to occur more often, or with higher values, when the model failed. Correlations were computed separately for *backchannels* and *non-event* predictions to examine whether different cues supported successful classification.

In Tables 7, 8, and 9, we report correlations for the best performing model of each modality: *hubert-large-ll60k* for audio, *vidtok_kl_causal_488_16chn* for video, and the combined *SpeechTokenizer_snake* & *vidtok_kl_causal_488_16chn* model for the multimodal setting.

Features	Backchannel		Other	
	r	p	r	p
Silence	0.015	= 0.25	-0.267	< .001
Speech	**0.346**	< .001	**0.281**	< .001
Laugh	-0.033	< 0.05	0.037	< .005
Inter.	**0.349**	< .001	**0.166**	< .001
ouais	**0.264**	< .001	**0.164**	< .001
d'accord	0.015	= 0.26	-0.025	= 0.06
okay	**0.350**	< .001	**0.165**	< .001
hm	**0.163**	< .001	0.007	= 0.59
oui	**0.197**	< .001	**0.181**	< .001
mean F0	-0.127	< .001	**0.283**	< .001
sd F0	-0.059	< .005	0.040	< .05
mean Loud.	-0.174	< .001	**0.367**	< .001
sd Loud.	-0.052	< .010	**0.210**	< .001

Table 7: Pearson's correlation coefficients and p-values for the **audio model**, separated by backchannel and non-event predictions, with correlations greater than 0.10 highlighted in bold.

Features	Backchannel		Other	
	r	p	r	p
Nod	**0.141**	< .001	-0.154	< .001
Shake	0.021	= 0.10	-0.084	< .001
Tilt	0.026	< .05	-0.076	< .001
Other	-0.015	= 0.26	-0.005	= 0.72
Speech	-0.016	= 0.21	-0.227	< .001
Laugh	0.003	= 0.81	-0.042	< .005

Table 8: Pearson's correlation coefficients and p-values for the **video model**, separated by backchannel and non-event predictions, with correlations greater than 0.10 highlighted in bold.

The results indicate that distinct sets of features contribute to successful *backchannel* and *non-event* predictions. As shown in Table 7, speech activity shows strong positive correlations with prediction success in both cases (r = 0.346, p < .001 for events; r = 0.281, p < .001 for non-events), indicating that the presence of speech in general

Features	Backchannel		Other	
	r	p	r	p
Nod	0.056	< .001	-0.092	< .001
Shake	0.005	= 0.69	0.037	< .005
Tilt	0.015	= 0.25	-0.030	= 0.02
Other	-0.010	= 0.43	0.029	= 0.03
Silence	0.012	= 0.36	-0.187	< .001
Speech	**0.103**	< .001	**0.172**	< .001
Laugh	-0.061	< .001	0.029	< .05
Inter.	**0.106**	< .001	0.093	< .001
ouais	0.068	< .001	0.092	< .001
d'accord	0.016	= 0.22	-0.032	= 0.02
okay	**0.105**	< .001	0.093	< .001
hm	0.057	< .001	-0.017	= 0.19
oui	0.018	= 0.158	**0.110**	< .001
mean F0	-0.127	< .001	**0.164**	< .001
sd F0	-0.059	< .005	0.012	= 0.53
mean Loud.	-0.174	< .001	**0.242**	= 0.53
sd Loud.	-0.052	= 0.01	**0.113**	< .001

Table 9: Pearson's correlation coefficients (r) and associated p-values (p) for the **multimodal model**, separated by backchannel and non-event predictions, with correlations greater than 0.10 highlighted in bold.

facilitates classification. However, interjections provide a more specific lexical cue for **backchannel** prediction, with robust correlations observed for interjections overall (r = 0.349, p < .001) and particularly for *ouais* and *okay*. By contrast, acoustic features (pitch and loudness) show divergent patterns across conditions. For **backchannel** classification, higher values of pitch and loudness are negatively correlated with success, suggesting that increased prosodic prominence tends to mislead the model. In contrast, **non-event** classification benefits more strongly from acoustic cues. Higher mean pitch (r = 0.283, p < .001), greater loudness (r = 0.367, p < .001), and larger loudness variability (r = 0.210, p < .001) are all positively associated with successful classification. These findings are consistent with the production characteristics of the two categories: backchannel events are typically produced as short interjections with low pitch and intensity, whereas non-events involve longer and more complex speech, accompanied by greater prosodic variability that the model can exploit to distinguish them.

The video model exhibits generally weak correlations with prediction success, reflecting its overall low classification performance. For **event** probing, nodding is the only feature exceeding 0.10 (r = 0.141, p < .001; Table 8), consistent with the fact that nods are frequently produced during backchanneling. For **non-events**, speech activity is

negatively correlated with prediction success (r = -0.227, p < .001), and head movements and laughter show weak or negative correlations. Overall, the correlations suggests that the video model fails to distinguish between brief speech produced by listeners during backchannels and the longer, more complex main speakers' speech and gestures. Such confusion likely contributes to the video model's low performance, as it is unable to capture these fine-grained interactional differences.

The correlations observed in the multimodal model (Table 9) generally reflect the same patterns as the unimodal analyses: interjections are most strongly associated with **event** prediction, whereas acoustic features are positively correlated with **non-event** prediction.

4.3 Representation projections

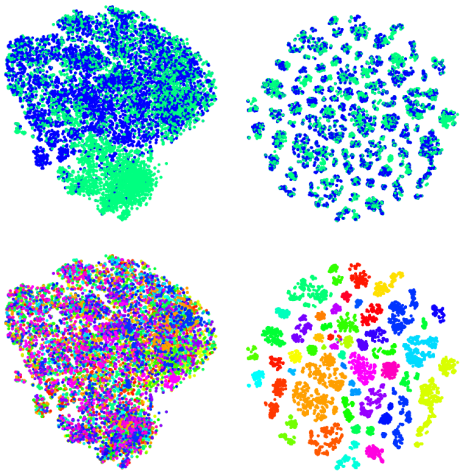

Figure 3: t-SNE projection of event representations for SpeechTokenizer (left) and VidTok (right), colored by event type (top) and speaker (bottom).

From Table 3, there seem to be a performance gap between the two modalities, up to 10 points with the compared models. To tentatively explain this gap, we plotted locality preserving 2D projections of the representations of the 12k events of the dataset using t-SNE. Although this method leads to projections with relatively limited utility, we can observe in Figure 3 that SpeechTokenizer and VidTok representations have very different structure: while the former corresponds to a dense cloud, the later is very clustered. We looked at the t-SNE plot for all models and observed an identical trend, audio and video representations are dissimilar, irrespective of model structure, training losses or data size. Figure 3 shows that while backchannels occupy a distinctive subspace in audio representations, they are scattered among clusters in video representations (top row). When coloring data points with speaker identities, it appears that video representations are dominated by identity information, which might explain the lower performance of the probes. It is the case even though videos are shot in controlled conditions with controlled lighting and uniform background, meaning that those clusters are really related to participant identity or behavior idiosyncrasies. An hypothesis we have is that the absence of distillation from "higher-level" HuBERT-like units is what sets apart video tokenizers. It would be interesting to explore how such component could affect the quality of video tokenizers.

4.4 Limitations

The main limitation of our work is that we do not have enough training data in order to separate speakers in training and test. An order of magnitude larger dataset would be necessary to be able to apply probes to novel participants. Another limitation is that although the SMYLE dataset contains speaker-specific recordings, the interlocutor is slightly audible in some recordings, giving models the opportunity to rely on that information to identify backchannel opportunities in their speech. Finally, resorting to off-the-shelf models, especially when their training recipe is not well documented, precludes definite conclusions on matters related to training data. Further experiments are needed to address those limitations.

5 Conclusion

This study assesses whether backchannels are naturally captured by the self-supervised training of audio and video tokenizers. We observe that the accuracy of linear probes trained from their representations is higher for audio than video models. Analysis uncovers that video representations are much more centered on speakers/participants than audio representations.

In future work, we will explore whether LLMs trained on top of those tokenizers can effectively better extract backchannels from the representation space that linear probes cannot untangle. We also plan on training audio-visual tokenizers to better model fine-grained behavior.

References

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. 2025. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*.

Sames Al Moubayed, Malek Baklouti, Mohamed Chetouani, Thierry Dutoit, Ammar Mahdhaoui, J-C Martin, Stanislav Ondas, Catherine Pelachaud, Jérôme Urbain, and Mehmet Yilmaz. 2009. Generating robot/agent backchannels during a storytelling experiment. In *2009 IEEE International Conference on Robotics and Automation*, pages 3749–3754. IEEE.

Ahmed Youssef Ali Amer, Chirag Bhuvaneshwara, Gowtham Krishna Addluri, Mohammed Maqsood Shaik, Vedant Bonde, and Philippe Muller. 2023. Backchannel detection and agreement estimation from video with transformer networks. *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Takanori Ashihara, Takafumi Moriya, Kohei Matsuura, Tomohiro Tanaka, Yusuke Ijima, Taichi Asami, Marc Delcroix, and Yukinori Honma. 2023. Speechglue: How well can self-supervised speech models capture linguistic knowledge? In *Interspeech*.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. data2vec: A general framework for self-supervised learning in speech, vision and language. *ArXiv*, abs/2202.03555.

Alexei Baevski, Steffen Schneider, and Michael Auli. 2019. vq-wav2vec: Self-supervised learning of discrete speech representations. *ArXiv*, abs/1910.05453.

Dominika Basaj, Witold Oleszkiewicz, Igor Sieradzki, Michał Górszczak, Barbara Rychalska, Tomasz Trzcinski, and Bartosz Zieliński. 2021. Explaining self-supervised image representations with visual probing. In *International Joint Conference on Artificial Intelligence*.

Janet Beavin Bavelas, Linda Coates, and Trudy Johnson. 2000. Listeners as co-narrators. *Journal of personality and social psychology*, 79(6):941–952.

Roxane Bertrand, Gaëlle Ferré, Philippe Blache, Robert Espesser, and Stéphane Rauzy. 2007. Backchannels revisited from a multimodal perspective. In *Auditory-visual speech processing*, pages 1–5.

Peter Blomsma, Julija Vaitonyté, Gabriel Skantze, and Marc Swerts. 2024. Backchannel behavior is idiosyncratic. *Language and Cognition*, 16(4):1158–1181.

Kübra Bodur, Mitja Nikolaus, Fatima Kassim, Laurent Prévot, and Abdellah Fourtassi. 2021. Chico: A multimodal corpus for the study of child conversation. *Companion Publication of the 2021 International Conference on Multimodal Interaction*.

Auriane Boudin, Roxane Bertrand, Stéphane Rauzy, Matthis Houlès, Thierry Legou, Magalie Ochs, and Philippe Blache. 2023. Smyle: A new multimodal resource of talk-in-interaction including neurophysiological signal. In *Companion Publication of the 25th International Conference on Multimodal Interaction*, pages 344–352.

Auriane Boudin, Roxane Bertrand, Stéphane Rauzy, Magalie Ochs, and Philippe Blache. 2021. A multimodal model for predicting conversational feedbacks. In *International conference on text, speech, and dialogue*, pages 537–549. Springer.

Auriane Boudin, Stéphane Rauzy, Roxane Bertrand, Magalie Ochs, and Philippe Blache. 2024. The distracted ear: How listeners shape conversational dynamics. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15872–15887, Torino, Italia. ELRA and ICCL.

Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel F. Valstar. 2017. The noxi database: multimodal recordings of mediated novice-expert interactions. *Proceedings of the 19th ACM International Conference on Multimodal Interaction*.

Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel C. Tompkins, Zhuo Chen, and Furu Wei. 2022. Beats: Audio pre-training with acoustic tokenizers. *ArXiv*, abs/2212.09058.

Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, and Yonghui Wu. 2022. Self-supervised learning with random-projection quantizer for speech recognition. In *International Conference on Machine Learning*.

Yu-An Chung and James R. Glass. 2019. Generative pre-training for speech with autoregressive predictive coding. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3497–3501.

IA de Kok and Dirk KJ Heylen. 2012. A survey on evaluation metrics for backchannel prediction models. In *Interdisciplinary Workshop on Feedback Behaviors in Dialog, Stevenson, Washington, USA: Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog*, pages 15–18. University of Texas.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

Ziedune Degutyte and Arlene Astell. 2021. The role of eye gaze in regulating turn taking in conversations: a systematized review of methods and findings. *Frontiers in Psychology*, 12:616471.

Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. 2016. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1459–1462, New York, NY, USA. Association for Computing Machinery.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Vidit Jain, Maitree Leekha, Rajiv Ratn Shah, and Jainendra Shukla. 2021. Exploring semi-supervised learning for predicting listener backchannels. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, et al. 2024. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. *arXiv preprint arXiv:2408.16532*.

Dan Jurafsky, Rebecca A. Bates, Noah Coccaro, Rachel Martin, Marie W. Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul A. Taylor, and Carol Van Ess-Dykema. 1998. Switchboard discourse language modeling project (final report). In *LVCSR Workshop*.

Pavel Kholiavin, Anna Mamushina, Daniil Kocharov, and Tatiana Kachkovskaia. 2020. Automatic detection of backchannels in russian dialogue speech. In *International Conference on Speech and Computer*, pages 204–213. Springer.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *ArXiv*, abs/2010.05646.

Akash Kumar, Ashlesha Kumar, Vibhav Vineet, and Yogesh Singh Rawat. 2023. A large-scale analysis on self-supervised video representation learning. *arXiv preprint arXiv:2306.06010*.

Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, Tanghui Jia, Junwu Zhang, Zhenyu Tang, Yatian Pang, Bin She, Cen Yan, Zhiheng Hu, Xiao wen Dong, Lin Chen, Zhang Pan, Xing Zhou, Shaoling Dong, Yonghong Tian, and Li Yuan. 2024. Open-sora plan: Open-source large video generation model. *ArXiv*, abs/2412.00131.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman H. Khan, and Fahad Shahbaz Khan. 2024. Videogpt+: Integrating image and video encoders for enhanced video understanding. *ArXiv*, abs/2406.09418.

Kinan Martin, Jon Gauthier, Canaan Breiss, and Roger Philip Levy. 2023. Probing self-supervised speech models for phonetic and phonemic information: a case study in aspiration. In *Interspeech*.

Fabian Mentzer, David C. Minnen, Eirikur Agustsson, and Michael Tschannen. 2023. Finite scalar quantization: Vq-vae made simple. *ArXiv*, abs/2309.15505.

Markus Mueller, David Leuschner, Lars Briem, Maria Schmidt, Kevin Kilgour, Sebastian Stueker, and Alex Waibel. 2015. Using neural networks for data-driven backchannel prediction: A survey on input features and training techniques. In *Human-Computer Interaction: Interaction Technologies: 17th International Conference, HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proceedings, Part II 17*, pages 329–340. Springer.

Philippe Muller, Michael Dietz, Dominik Schiller, Dominike Thomas, Hali Lindsay, Patrick Gebhard, Elisabeth Andr'e, and Andreas Bulling. 2022. Multimediate'22: Backchannel detection and agreement estimation in group interactions. *Proceedings of the 30th ACM International Conference on Multimedia*.

Jerry Ngo and Yoon Kim. 2024. What do language models hear? probing for auditory representations in language models. In *Annual Meeting of the Association for Computational Linguistics*.

Tu Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Mamdouh Elkahky, Paden Tomasello, Robin Algayres, Benoît Sagot, Abdel rahman Mohamed, and Emmanuel Dupoux. 2022. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266.

Hiroaki Noguchi and Yasuharu Den. 1998. Prosody-based detection of the context of backchannel responses. *5th International Conference on Spoken Language Processing (ICSLP 1998)*.

Ankita Pasad, Chung-Ming Chien, Shane Settle, and Karen Livescu. 2023. What do self-supervised speech models know about words? *Transactions*

of the Association for Computational Linguistics, 12:372–391.

Dustin Podell, Zion English, Kyle Lacey, A. Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *ArXiv*, abs/2307.01952.

Laurent Prévot, Jan Gorisch, and Roxane Bertrand. 2016. A cup of cofee: A large collection of feedback utterances provided with communicative function annotations. In *International Conference on Language Resources and Evaluation*.

Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685.

Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. 2017. Enhancing backchannel prediction using word embeddings. In *Interspeech*, pages 879–883.

Emanuel A Schegloff. 1982. Discourse as an interactional achievement: Some uses of 'uh huh'and other things that come between sentences. *Analyzing discourse: Text and talk*, 71:71–93.

Jui Shah, Yaman Kumar Singla, Changyou Chen, and Rajiv Ratn Shah. 2021. What all do audio transformer models hear? probing acoustic representations for language delivery and its structure. *ArXiv*, abs/2101.00387.

Elizabeth Shriberg, Rajdip Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The icsi meeting recorder dialog act (mrda) corpus. In *SIGDIAL Workshop*.

Chenxi Sun, Moxian Song, Derun Cai, Bao Feng Zhang, linda Qiao, and Hongyan Li. 2024. A systematic review of echo state networks from design to application. *IEEE Transactions on Artificial Intelligence*, 5:23–37.

Anni Tang, Tianyu He, Junliang Guo, Xinle Cheng, Li Song, and Jiang Bian. 2024. Vidtok: A versatile and open-source video tokenizer. *arXiv preprint arXiv:2412.13061*.

Khiet P. Truong, Ronald Poppe, Iwan de Kok, and Dirk K. J. Heylen. 2011. A multimodal analysis of vocal and visual backchannels in spontaneous dialogs. In *Interspeech*.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748.

Ani Vanyan, Alvard Barseghyan, Hakob Tamazyan, Vahan Huroyan, Hrant Khachatrian, and Martin Danelljan. 2023. Analyzing local representations of self-supervised vision transformers. *ArXiv*, abs/2401.00463.

Alessandro Vinciarelli, Maja Pantic, Hervé Bourlard, and Alex Pentland. 2008. Social signals, their function, and automatic analysis: a survey. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 61–68.

Junke Wang, Yi Jiang, Zehuan Yuan, Binyue Peng, Zuxuan Wu, and Yu-Gang Jiang. 2024. Omnitokenizer: A joint image-video tokenizer for visual generation. *ArXiv*, abs/2406.09399.

Wilson Yan, Yunzhi Zhang, P. Abbeel, and A. Srinivas. 2021. Videogpt: Video generation using vq-vae and transformers. *ArXiv*, abs/2104.10157.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. 2024. Cogvideox: Text-to-video diffusion models with an expert transformer. *ArXiv*, abs/2408.06072.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2023. Speechtokenizer: Unified speech tokenizer for speech large language models. *arXiv preprint arXiv:2308.16692*.

Sijie Zhao, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Muyao Niu, Xiaoyu Li, Wenbo Hu, and Ying Shan. 2024. Cv-vae: A compatible video vae for latent generative video models. *ArXiv*, abs/2405.20279.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. 2024. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*.

A Model variants

We use author-provided variants for WavTokenizer, SpeechTokenizer, Cosmos, VidTok:

- WavTokenizer[4]: `small` is trained on LibriTTS, `medium` is trained on 10k hours of speech, audio and music, and `large` is trained on 80k hours of the same; `600` yields 40 units per second while `320` yields 75 units per seconds.
- SpeechTokenizer[5]: `hubert_avg` is trained on LibriSpeech and adopts average representation across all HuBERT layers as semantic

[4] https://github.com/jishengpeng/WavTokenizer
[5] https://github.com/ZhangXInFD/SpeechTokenizer

teacher, while snake LibriSpeech and Common Voice, with Snake activation, average representation across all HuBERT layers.

- Cosmos[6]: variants depend on the type of latent representation (CV for continuous, DV for discrete) and compression $t \times x \times y$ where t is temporal, and $x \times y$ is spatial resolution; 0.1 models are trained on instances of 17 frames while 1.0-8x8x8 is trained on 49 frames and 1.0-8x16x16 is trained on 121 frames.
- VidTok[7]: variants are categorized according to the training regularizer/quantizer (kl, Kullback-Leibler for continuous latents, and fsq, Finite Scalar Quantization for discrete latents), the compression ratio (txy, temporal and spatial, 41616 meaning 4x16x16), and the size of the latent space (in channels for continuous latents and codebook size for discrete latents).

[6] https://huggingface.co/nvidia/Cosmos-1.0-Tokenizer-DV8x16x16

[7] https://github.com/microsoft/vidtok