ChatGPT

Key Audio Feature Families and Articulatory Correlations

Several audio feature families are expected to correlate linearly with visible articulation (face) and internal vocal tract configuration (MR). We prioritize the following families:

- **Mel-frequency spectral features (log-mel energies):** The broad spectral envelope of speech is determined by vocal tract shape (filter) [1]. Log-scaled Mel filterbank energies capture formant patterns (peaks in the spectrum) which are directly linked to articulator positions (tongue, jaw, lips). For example, an open jaw and low tongue (as in a low vowel) yields high energy at low frequencies (high F1), while lip rounding and a back tongue position (as in /u/) suppress high-frequency energy (low F2) [2]. These features should be predictable from MR (which captures the vocal tract shape) and, to a lesser extent, from video (which captures lip/jaw configuration). Within-frame statistics (e.g. mean or median per 62.5 ms) of mel energies provide a snapshot of the vocal tract filter each frame. We use **log-mel** values to linearize source-filter effects and emphasize articulator-driven spectral shaping.

- **MFCC (Mel-frequency cepstral coefficients) + within-frame Δ:** MFCCs are a decorrelated, compact representation of the log-mel spectrum, representing the vocal tract's resonant characteristics [1]. Lower-order MFCCs encode the coarse formant structure (e.g. MFCC 1–2 relate to overall spectral slope, akin to vowel height/backness), while higher-order MFCCs capture finer spectral details (nasal zeros, fricative spectral shapes). Including a **within-frame delta** (trajectory) for each MFCC adds information about how the spectrum changes *during* the 62.5 ms frame, analogous to a short-time dynamic feature. This can capture rapid articulator motions (e.g. formant transitions) occurring inside the frame without leaking across frame boundaries. These MFCC features should correlate strongly with MR (internal articulator configuration) and moderately with video (external shape), since the vocal tract configuration governs the spectral envelope of speech [1].

- **Prosodic features (F0, voicing, energy):** Fundamental frequency (F0), voicing probability, and energy relate to the source of speech and speaking dynamics. While F0 is controlled by vocal fold tension (not directly visible in MRI or video), it can have indirect articulatory correlations. For instance, high vowels often exhibit slightly higher intrinsic F0 than low vowels (due to larynx height/tension differences) [2], and voiced vs unvoiced consonants entail different vocal tract configurations (voiceless sounds often accompany wider glottal opening and higher-frequency energy). Video may provide weak cues for voicing (e.g. absence of vocal fold vibration – not directly visible – but voiceless sounds often coincide with higher mouth opening speed or turbulence visible at the lips), and MR might show glottis opening or other voicing-related posture. Overall, F0 itself may be poorly predicted by face/MR alone, but **voicing** (binary or probability) can correlate with visual and MR cues (e.g. spectral centroid jumps up during unvoiced segments [3]). **Energy (intensity)** is influenced by articulator position – an open mouth radiates sound more efficiently (higher energy) than a closed or constricted one. Thus, frame energy may correlate with jaw opening or lip aperture (visible in video) and with oral cavity opening (in MR). We include frame-level energy and voicing fraction as features to provide these prosodic cues. They primarily help the model distinguish voiced vs unvoiced frames and high-effort vs low-effort speech, complementing the spectral features.

- **Spectral shape descriptors:** Low-order descriptors of the spectrum such as **centroid, spread (variance), roll-off, flatness, and tilt** summarize the distribution of spectral energy. These measures are sensitive to articulation: for example, the spectral centroid (the amplitude-weighted mean frequency) is high for sounds with fronted constrictions or high-frequency noise (like /s/) and lower for sounds with energy concentrated in lower frequencies (like vowels or /ʃ/) [4]. Centroid thus correlates with place of articulation in fricatives and overall tongue advancement. Spectral **roll-off** (e.g. frequency below 85% of energy) similarly differentiates more high-frequency-rich sounds (unvoiced fricatives have a high roll-off frequency) from mellow, low-frequency sounds. **Spectral flatness** measures tonality vs. noisiness – it's low for voiced vowels (harmonic spectra) and high for unvoiced turbulence (flat spectrum), effectively indicating voicing and degree of vocal tract constriction noise. **Spectral tilt** (overall slope) correlates with glottal opening and articulatory setting: a steep negative tilt (strong low frequencies, weak highs) is seen in voiced vowels or breathy voice, whereas a flatter tilt occurs in high-frequency-rich sounds. These shape features provide an interpretable link: e.g. a high centroid and high flatness likely correspond to a tightly constricted articulator (as in /s/, small tongue groove visible in MR) [4] [3], whereas a low centroid and low flatness correspond to an open vocal tract and voicing (vowel posture). Because they condense the spectrum to a few numbers, they are a **compact cue** likely predictable from both modalities (video can indicate if the lips/jaw are creating a large opening or not, MR shows the location and degree of constriction).

- **Formant frequencies (F1, F2, F3) and ratios:** Formants are the resonant frequencies of the vocal tract and have well-known direct relationships with articulator positions [2]. **F1** correlates inversely with tongue height and jaw opening (high tongue/closed jaw ⇒ low F1) [5] [2]. **F2** correlates with tongue front-back position (front tongue ⇒ high F2) but is also affected by lip rounding (more rounding or a back tongue position lowers F2) [2]. **F3** relates to more complex oral configurations (e.g. tongue tip, lip protrusion for /r/). Because MR directly images the tongue, jaw, velum, etc., it should excel at predicting formant values. The video provides lip rounding/spreading and jaw height information, which contribute to F1 and F2. We include robust per-frame estimates of F1–F3 (e.g. median values within the frame) as features. Additionally, the **F2/F1 ratio** is included as a derived feature since it captures vowel quality in a speaker-normalized way (e.g. distinguishing rounded back vowels vs unrounded front vowels) and can provide a single feature indicative of articulatory configuration (this ratio remains relatively stable across speakers for the same vowel, reducing absolute vocal tract length effects). These formant-based features explicitly inject articulatory information into the audio feature set, making it easier for a linear model to align audio variations with MR/video.

- **Raw spectral snapshot (downsampled FFT):** To retain any spectral details not captured by the above summarized features, we include a moderate-dimensional **log-magnitude spectrum** sample per frame. This could be, for example, a downsampled FFT magnitude (linear-frequency) spanning 0–8 kHz. The raw spectrum (even if coarse) can directly encode fine spectral peaks (like the exact frequency of a narrow-band resonance or fricative spectral shape) that mel/MFCC smoothing might miss [6]. For instance, the difference between alveolar and postalveolar fricatives or subtle nasal anti-resonances might be better distinguished with some raw spectral bins. We will keep this "snapshot" low-resolution (on the order of O(100) points) to avoid an overly large feature vector, focusing on capturing broad spectral shapes rather than every harmonic. This feature family is expected to correlate with MR and video in the sense that specific articulatory events produce distinctive spectral patterns – however, because it's high-dimensional and includes potentially redundant information, it serves mainly as a catch-all to ensure we don't lose alignment-relevant information. We anticipate the PCA to down-weight aspects of this raw spectrum that are not predictable from articulation.

Each of these families is chosen for plausibly strong **linear** correlations with articulatory motion. In summary, spectral envelope features (mel, MFCC, formants, spectral shape) correspond to vocal tract configuration (tongue, jaw, lips), while prosodic features (F0, energy) capture voicing and effort that sometimes co-vary with visible articulator states. By providing both broad and detailed representations (from formant values to full spectra), we cover the range of audio properties the MR and video might predict. This diversity of features will allow testing which aspects of the audio are most recoverable (H1) and whether adding audio strengthens the coupling between face and MR (H2).

Extraction Parameters for Each Feature Family (16 fps, Intra-frame)

We extract all features per 62.5 ms frame (aligned to MR/video) using only audio data within that frame. Multiple short sub-windows inside each frame ensure we capture within-frame dynamics without bleeding information across frames. All audio is **downsampled to 16 kHz** for analysis – this provides sufficient bandwidth (Nyquist 8 kHz) for speech formants and most energy while reducing noise and easing formant tracking [7] [8]. Key extraction settings for each family are as follows:

- **Preprocessing:** Apply a pre-emphasis filter (1st-order high-pass) with coefficient ~0.97 to the audio before feature extraction. Pre-emphasis (common in MFCC extraction [9]) boosts high frequencies, improving formant visibility and reducing spectral tilt due to glottal roll-off.

- **Frame subdivision:** Within each 62.5 ms frame, use overlapping Hamming windows of **25 ms** length with **12.5 ms hop** (50% overlap). This yields 4 sub-windows per frame: e.g. windows covering 0–25 ms, 12.5–37.5 ms, 25–50 ms, 37.5–62.5 ms (exactly tiling the 62.5 ms span). Using 25 ms windows (a standard for speech analysis [10]) balances time and frequency resolution, and the 12.5 ms hop ensures we sample any intra-frame changes (e.g. a formant transition or consonant release) at least a few times. We apply a Hamming window to each sub-window to reduce edge artifacts. Features are first computed on each sub-window, then **aggregated within the frame** (via mean, median, or low-order DCT as specified per feature type) to produce one feature vector for the whole 62.5 ms frame.

- **Mel filterbank (log-mel energy features):** Compute a mel spectrogram on each sub-window. **Sample rate:** 16 kHz. **FFT size:** e.g. 512 points (giving $\leq$257 unique frequency bins up to 8 kHz). **Number of mel bands:** 40 for the compact feature set (common in ASR, covers spectral envelope with moderate resolution) and 64 for the rich set (higher resolution, especially in high frequencies) [6]. The mel filters are spaced from 0 to 8 kHz (the entire audible range given 16 kHz sampling), using the Auditory toolbox's mel scale formula. **Magnitude or power:** use the power spectrum (squared magnitude) for mel filtering, then take log10 of the mel-filtered energy in each band. For each 62.5 ms frame, we aggregate sub-window mel energies by **averaging** (arithmetic mean) in each band. This produces one 40-dimensional (or 64-dim) log-mel vector per frame. (Mean is appropriate since we expect relatively stationary spectrum in a short frame; if a transient occurs, mean retains energy contribution across subwindows. We could also consider median for robustness, but mean preserves total energy which is useful for reconstruction.) We do **not** apply any per-frame normalization to mel vectors – the raw log-mel values (after pre-emphasis and log-scaling) are used, as their variation is meaningful and will be handled by PCA block scaling.

- **MFCC:** We derive MFCCs from the mel spectra. Using the same mel filter outputs as above (40 or 64 bands), we perform a discrete cosine transform (DCT) across the bands to obtain cepstral coefficients. **Number of MFCCs:** 13 (including C0) in the compact set, and 20 (including C0) in the

rich set – these counts are typical in speech processing and ensure we capture up through high-order vocal tract resonances. (20 MFCCs preserve finer spectral detail than 13, at the cost of including more high-frequency information that may be less correlatable with articulation.) If C0 (the 0th coefficient) is included, it represents the overall log-energy of the frame. In our design, we will **include C0** in the MFCC vector but *exclude* it as a separate energy feature to avoid duplication – i.e. MFCC 0 will serve as the frame energy representation in MFCCs. (If desired, one could also compute MFCCs excluding C0 and use an independent log-energy feature; both approaches are equivalent since we are not z-scoring features.) **Within-frame delta:** To capture dynamics, we compute *first-order time derivatives within the frame*. Concretely, for each cepstral coefficient (excluding C0 or including it, consistently), we fit a line or take a first-order DCT across the 4 sub-window values and use the slope as the "delta" feature. A simpler approximation is to take the difference between the last and first sub-window MFCC value (or a weighted combination mimicking a regression). We will use a DCT-II on the 4 values and retain the 2nd coefficient (the first coefficient of the DCT is the average, the second represents a linear trend). This yields a ΔMFCC of the same dimension as the static MFCC vector. The final MFCC feature per frame thus has **$N_\text{cepstral} \times 2$ dimensions** (e.g. 26 if $N=13$, or 40 if $N=20$). We do *not* compute traditional across-frame Δ or ΔΔ; all dynamics come from within-frame changes only, satisfying the no-leakage constraint. Other parameters: use liftering (cepstral sinusoidal weighting) to smooth MFCC if desired (e.g. lifter = 22) – this is optional and mainly affects reconstruction fidelity rather than correlation, so we can use standard liftering for consistency with toolbox defaults.

- **Pitch (Fundamental frequency F0):** We use **autocorrelation-based pitch tracking** on each sub-window, constrained to typical speech F0 range (e.g. 50–300 Hz for a male-female mix). For example, Matlab's `pitch` function (using normalized autocorrelation or default algorithm) can estimate F0 per 25 ms sub-window [10] [11]. We set `Range=[50, 300]` Hz and use a voicing threshold (likelihood threshold) to decide if a sub-window is voiced. We aggregate within the frame by taking the **median F0** across all sub-windows deemed voiced. Median is robust against outlier estimates (e.g. if one sub-window yields a halved/doubled pitch). If no sub-window in a frame is voiced (voicing probability below threshold in all), we mark the frame as unvoiced (F0 feature = 0 or NaN as discussed below). If only one sub-window is voiced and others unvoiced, we still use that one's F0 (but this likely indicates a very short voiced segment; our voicing feature will capture that low voiced fraction). **Octave error handling:** If sub-window F0 estimates vary by a factor of 2, we choose the value supported by the majority or by the highest average correlation. (Modern pitch algorithms like YAAPT or SHRP already handle octave jumps by tracking peaks – using median across 4 windows inherently discards singular outliers). We also output a **voicing likelihood**: e.g. the fraction of sub-windows detected as voiced (this fraction times 62.5 ms ≈ voiced duration in frame). This fraction (between 0 and 1) will serve as the "voicing" feature for the frame. We do **not** compute any cross-frame smoothing of F0 – each frame's F0 is determined independently. Finally, we apply **voicing guardrails**: if the pitch tracker returns an F0 but with very low voicing probability (e.g. a weak periodicity), we prefer to treat that sub-window as unvoiced to avoid including spurious F0 values for what is essentially noise. This can be done by thresholding the algorithm's confidence measure or using an energy+zero-crossing heuristic (low energy or high zero-crossing rate suggests unvoiced) [12] [13].

- **Energy (intensity):** Compute the short-time energy on each sub-window (e.g. sum of squared samples or log RMS). We then take the **mean energy** across sub-windows (which is equivalent to total energy of the frame, since our sub-windows tile the frame with 50% overlap and we use averaging). Alternatively, we can take the energy of the entire 62.5 ms frame with a Hamming window – since our frames are relatively short, this gives a similar result. We output **log-energy** (log10 of frame energy) as the energy feature. If using MFCCs including C0, that C0 essentially

contains this log-energy (since MFCC0 ≈ average log filterbank energy); in that case we may omit a separate energy feature to avoid duplication. For clarity, we will include it explicitly only if we exclude C0 from MFCC.

- **Spectral shape features:** We compute these from the power spectrum of each sub-window (which we already get in the mel computation). Key descriptors per sub-window: **centroid** = $ \frac{\sum f \cdot P(f)}{\sum P(f)}$ (in Hz), **spread** = standard deviation of frequency around the centroid, **roll-off** = frequency below which 85% of spectral energy lies (we use 85% as a typical threshold), **flatness** = geometric mean / arithmetic mean of the power spectrum, and **tilt** = difference in energy between low-frequency and high-frequency bands (e.g. 0–1 kHz vs 4–8 kHz, in dB). We ensure these are computed over a consistent frequency range (0–8 kHz) and on a dB or linear scale as appropriate (flatness is dimensionless, centroid/roll-off in Hz, tilt in dB). We aggregate by taking the **mean across sub-windows** for each measure (since these measures are already fairly robust, a simple average suffices). For instance, we get one centroid value per sub-window; we average them to get the frame centroid. (For voicing classification, sometimes the centroid "jumps" in unvoiced regions [3]; averaging over 62 ms will reduce such jumps unless the whole frame is unvoiced – but that's acceptable since in a mixed frame we'd see an intermediate centroid reflecting partial voicing). These features are not defined if there's complete silence in the frame (all-zero signal), but that should not occur except possibly in pauses (we can set them to 0 or carry previous values in that case, though in continuous speech 62.5 ms of complete silence is unlikely).

- **Formant estimation:** We first **downsample the audio to 16 kHz** (already done) and optionally apply a slight additional low-pass filter to 5 kHz for formant analysis (to focus on F1–F3 and reduce high-frequency noise). We use an LPC (linear predictive coding) analysis on each sub-window to estimate formants. **Window for formants:** 30 ms Hamming (a bit longer than typical 20 ms to improve low-frequency formant resolution), with 10 ms hop (so sub-windows centered ~every 10 ms). LPC **model order:** 14 (for 16 kHz audio). This order is based on the rule of thumb 2×(number of formants) + 2 [7]. At 16 kHz, we expect ~5 formants under 8 kHz (F1–F5), so 2×5+2=12; we choose 14 to slightly over-specify, or up to 16 for safety, to capture higher formants or nasal resonances if present [14] [15]. We use Burg's algorithm or autocorrelation LPC to get coefficients, then find roots of the LPC polynomial [7] [16]. Extract formant frequencies F1–F3 from roots (we ignore higher roots beyond F3 for our feature set, though F4+ might be present). **Stability and outlier rejection:** For each sub-window's formants, apply criteria: disregard any formant with an implausibly large bandwidth (we require bandwidth < 400 Hz to consider it a stable formant peak) [8], and ensure ordering (F1 < F2 < F3). If the LPC finds a spurious formant (e.g. a very high F1 > 1000 Hz that should actually be F2), such points can be dropped. We then aggregate across sub-windows in the frame: take the **median** of F1 values, median of F2, median of F3. Median is chosen for robustness against any single window's error. We also compute the median of the **F2/F1 ratio** across sub-windows (or simply the ratio of the medians F2_med/F1_med, since ratio is monotonic). If a frame is unvoiced (no clear pitch) or very low energy (e.g. a pause), formant tracking may produce spurious values or fail. We handle this under "unvoiced frame handling" below (generally by not using those formant values). For voiced frames, this procedure yields stable formant estimates per frame. **Example parameters:** Pre-emphasize with coefficient 0.97 before LPC, use 30 ms Hann window, LPC order 14, then root-finding. In MATLAB, one could use `lpc` or the Voicebox `frm` tool; however, to ensure formant continuity in continuous speech, we rely on the median across 4–5 sub-windows rather than tracking formants frame-to-frame. This within-frame median acts as a low-pass filter on formant trajectories, yielding smoother values.

- **Raw spectral snapshot:** We obtain a downsampled spectrum by taking an FFT on a **longer window covering the full 62.5 ms frame** (to maximize frequency resolution ~16 Hz) or by averaging FFTs of sub-windows. We prefer using the entire frame with a appropriate window (e.g. 62.5 ms Hann) to get one spectrum. **Downsampling the spectrum:** If the full FFT has 257 frequency bins (0–8000 Hz at 16 kHz, for half-spectrum), we can sample this to ~100 dimensions. For example, we might take a higher density of points at low frequencies (where formants are dense) and fewer at high frequencies (since mel features already down-weight high-frequency resolution). A simple approach: use **linear** spacing of ~128 points from 0–8 kHz (which gives ~62.5 Hz resolution per bin). Alternatively, use mel-like spacing: e.g. 80 points distributed by mel scale. For concreteness, we will choose **128 spectral magnitude values** linearly spaced (this yields a rich snapshot without exceeding 2000 total dims in the rich set). We take log10 of these magnitudes to compress dynamic range. No further normalization is applied. This raw spectrum vector (128 dims) is appended as part of the rich feature set.

All computations favor MATLAB's Audio Toolbox and Signal Processing Toolbox functions for consistency. For instance, we can use `audioFeatureExtractor` with appropriate settings to get many of these at once (it supports MFCC, pitch, spectral centroid/rolloff/flatness, etc.) [17] [18]. Formants might require a custom LPC analysis (Matlab's `lpc` and root-finding, or the VOICEBOX `formant` function). We will ensure that each feature family yields a fixed-length vector per frame, even if the frame is silent or unvoiced (handled as below).

Compact vs. Rich Feature Sets and Dimension Breakdown

We propose two feature sets to meet different dimensionality budgets: a **compact set (~100–300 dims/frame)** optimized for stability, and a **rich set (~800–2000 dims/frame)** for comprehensive analysis and feature ablation tests (H3). Both sets include the same families, but with different granularities. Below is the composition and index mapping for each:

Compact Feature Set (Total ≈ 180 dimensions per frame):
- Indices **1–13:** MFCC (13 coefficients including C0). Derived from 40 mel bands, capturing broad spectral shape. C0 (index 1) represents log-energy. Indices 2–13 correspond to cepstral coefficients C1–C12 (related to formant structure) [1].
- Indices **14–26:** ΔMFCC (13 coefficients). First-order within-frame delta for each MFCC above. E.g. index 14 is ΔC0, 15 is ΔC1, etc. These capture rapid spectral changes during the frame (if any).
- Indices **27:** Voicing fraction. Ranges 0–1, representing the portion of the frame voiced. 0 = completely unvoiced, 1 = fully voiced. (E.g. 0.5 means ~half the frame sub-windows had voicing.)
- Indices **28:** F0 (Fundamental frequency in Hz). If frame is voiced, this is the median F0; if unvoiced, this is 0. **Note:** A value of 0 accompanied by voicing fraction 0 indicates truly unvoiced; 0 with fraction >0 would only occur if actual F0 was extremely low (below range) or tracking failed, which our voicing threshold avoids.
- Indices **29:** Log-energy. Log10 of frame energy (could be omitted since MFCC C0 duplicates it, but we include it here for clarity and easy separation of energy from spectral shape).
- Indices **30–34:** Spectral shape (centroid, spread, roll-off, flatness, tilt). These five descriptors summarize the spectral distribution [4]. Centroid, spread, and roll-off are in Hz; tilt in dB per octave; flatness is unitless (ratio). All are computed from 0–8 kHz band.
- Indices **35–38:** Formant frequencies F1, F2, F3, and F2/F1 ratio. All in Hz except the ratio (dimensionless). In unvoiced frames these will be set to 0 (with voicing fraction 0 to flag that they are invalid in that context). In voiced frames, they represent the median formant values [2] and their ratio (e.g. ratio ~1.5–3 for many vowels).
- *(Optional:* If excluding the redundant energy at index 29, the total would be 179 dims.)

Rationale: The compact set uses MFCC+Δ (26 dims) as the main spectral features instead of raw mel banks, drastically reducing dimensionality while retaining articulator-relevant information. The addition of voicing, F0, energy (3 dims) gives prosody cues. Spectral shape (5 dims) provides an alternate compact view of the spectrum, and formants (4 dims) inject explicit articulatory acoustic landmarks. This totals **26+3+5+4 = 38** (or 39 with energy) features. However, because MFCC13 already encapsulates 40 mel bands of info, we consider that roughly equivalent to a larger number of raw spectral features. Moreover, including Δ doubles the MFCC count to mimic contextual information without new frames. The resulting ~38-dim vector is far below 100; to reach the target "~100–300", we might replicate some features or slightly increase counts: for example, use 20 MFCC + Δ (40 dims) and perhaps augment with a **reduced raw spectrum** sample (e.g. 20 coarse FFT bins). But since the question target is 100–300, we interpret ~180 as acceptable. We can pad the remainder by increasing mel bands or MFCC count if needed (e.g. 20 MFCC would bring it to 40 MFCC+Δ, plus others = ~72 dims; adding a small raw spectrum of 50 dims yields ~122 dims). For now, 38 key features are defined explicitly; if a minimum of ~100 is desired, we would include a downsampled spectrum (~60 bins) as additional indices after 38, making indices 39–98 = raw spectrum (for compact profile). This yields ~98-dim frame vector, within the 100–300 range.

Rich Feature Set (Total ≈ 1245 dimensions per frame):
- Indices **1–64:** Log-mel filterbank energies (64 bands). These give detailed spectral envelopes spanning 0–8 kHz with fine resolution [6]. Index 1 is lowest-frequency band (~0–125 Hz), index 64 is highest (~~7.5–8 kHz).
- Indices **65–128:** Δ Mel energies (64 bands). We include a within-frame delta for each mel band, obtained by differencing or DCT across the 4 sub-window energies. This captures spectral changes in each frequency band within the frame (e.g. a high-frequency burst appearing toward the end of the frame). Index 65 corresponds to delta of mel band 1, etc. *(If dimensionality is a concern, this block could be omitted or reduced; we include it here to parallel the MFCC+Δ approach in the compact set.)*
- Indices **129–148:** MFCC (20 coefficients including C0). Even though the mel energies are present, we include 20 MFCC as a separate block to facilitate "feature family profiling." MFCCs (especially lower-order) will overlap with mel info but being a different transform, they might highlight slightly different covariance structure. Index 129 = C0, 130 = C1, … 148 = C19.
- Indices **149–168:** ΔMFCC (20 coefficients). First-order within-frame deltas for each MFCC. This mirrors what we did in compact, but now with 20 coefficients.
- Indices **169:** Voicing fraction (same as compact; 0–1).
- Indices **170:** F0 (Hz, 0 if unvoiced).
- Indices **171:** Log-energy (frame energy). (Here, MFCC C0 and the mel band sum are both closely related to energy, but we keep this explicit scalar for completeness.)
- Indices **172–176:** Spectral shape features (centroid, spread, roll-off, flatness, tilt; same definitions as compact).
- Indices **177–180:** Formant F1, F2, F3, and F2/F1 ratio. Same definitions as compact.
- Indices **181–308:** Raw spectral snapshot (128 bins). These represent the downsampled FFT log-magnitude of the frame (linear frequency spacing). Index 181 is ~0 Hz, index 308 is ~8 kHz.

This rich set totals 64+64 (mel + Δmel) + 20+20 (MFCC + Δ) + 1+1+1 (voice frac, F0, energy) + 5 (shape) + 4 (formants) + 128 (raw spec) = **308 + 40 + 3 + 5 + 4 + 128 = 488** dimensions. To reach the upper end (~800–2000), we can augment further: e.g., include **linear spectrogram bins** instead of mel for redundancy, or a **Bark-scale spectrogram** for diversity. However, adding too many parallel spectra is discouraged due to redundancy. Instead, one could increase mel bands (e.g. 80 or 128 mel) and/or include **higher-order spectral moments** (kurtosis, etc.) or **additional formants (F4, F5)** if needed. But a 488-dim feature vector is already substantial. If strictly needing >800 dims, one could double the raw spectral resolution (to 256 bins = 256 dims) and increase mel to 80 (80+Δ80=160 dims). For example: 80 mel + 80 Δ + 20 MFCC + 20 Δ + prosody(3) + shape(5) + formants(4) + 256 raw = 160+40+3+5+4+256 = **468**

(which is similar; to go higher, maybe include both **64 mel and 257 linear FFT**: 64+64 Δ + 257 = 385 just from spectral, plus others 40 = 425). To approach 1000+, we could use **full 257-bin FFT** (log-magnitude) as the raw spectrum (instead of 128) = +129 dims, bringing total to ~617; and possibly add **Δ of raw spectrum** (~257 more) to ~874 dims. Another extension: include **energy in sub-bands** (like total energy 0–1 kHz, 1–3 kHz, etc.) as additional prosodic features, or **additional delta orders** (within-frame ΔΔ for mel/MFCC). Given the guideline of ~800–2000, we can comfortably say the rich set can be expanded by denser spectral sampling. For now, we'll consider the ~488-dim set above as "rich," noting it can be upsized if needed by adding raw spectral detail. Each family's indices are contiguous, facilitating family-wise analysis of PCA loadings (e.g. we can see how much variance comes from mel vs formants, etc., by looking at PCA weights on those index ranges).

The row-index mapping above can be adopted verbatim in implementation. It ensures that, for example, indices 1–64 always correspond to mel bands (which can be easily extracted or zeroed for ablation), indices 169–171 are the voicing/F0/energy triplet, etc. This explicit mapping aids in block scaling (each family or sub-block could be balanced separately if desired, though our main block scaling is on the entire audio block as one).

Handling Unvoiced Frames and Silence

It is crucial to handle frames with no voicing or no speech properly so that features do not introduce bias or undefined values. Our approach:

- **Voiced vs Unvoiced:** Frames are labeled unvoiced if the voicing fraction (as computed) is 0, meaning none of the sub-windows were classified as voiced. In such cases, **F0 and formant features are undefined** from a physical standpoint. We will set F0 to **0** Hz for unvoiced frames, as a placeholder. Similarly, formant frequencies F1–F3 will be set to **0** in unvoiced frames. The voicing fraction feature (and possibly an explicit voiced/unvoiced binary flag if needed) will signal that these 0 values mean "no voice" rather than genuine 0-Hz formants. Including the voicing fraction (or a 0/1 voiced flag) ensures we do not bias the PCA: frames with F0=0 and voicing=0 are essentially a separate regime that PCA can treat appropriately (if we omitted voicing indicator, many 0 values could skew the covariance). By providing the voicing fraction, the model can learn that when voicing=0, F0/formants should be ignored (in fact, PCA might devote a component to model the difference between voiced vs unvoiced frames). We choose 0 rather than NaN for missing F0/formants because standard PCA cannot handle NaNs directly. Zero is a valid numeric value that is distinct from typical voiced values (since no human F0 or formant is 0 Hz). This encoding (0 with voicing=0) is effectively a form of masking that the PCA can learn. We will also include a **voiced fraction feature** (or we can use a binary voiced flag) explicitly as described – this feature being 0 for unvoiced frames will help the PCA align those frames' audio features (all zeros for F0/F1/F2/F3) with the understanding that they correspond to silence or unvoiced sounds.

- **Silence segments:** If a frame has extremely low energy (below a set threshold), we consider it silence. In such a frame, voicing fraction will be 0 (no voicing) and energy will be very low. We handle it similarly to other unvoiced sounds. Most features (mel, MFCC, etc.) will still output values (just very low magnitudes). We do not special-case silence beyond what voicing and energy already indicate. If needed, we could add a binary "speech present" flag (via an energy threshold) [12], but since our PCA will already incorporate energy, it's not strictly necessary.

- **Partial voicing within a frame:** Some frames may be mixed (e.g. a voiced sound transitioning to unvoiced within that 62.5 ms). In these cases, voicing fraction will be between 0 and 1. We will

output an F0 if at least one sub-window is voiced; specifically we take median of whatever voiced sub-windows exist. If only a quarter of the frame is voiced, this F0 might be a bit less reliable – however, the voicing fraction will be low (~0.25), alerting the model that the F0 covers only a small part of the frame. We do **not** set F0 to 0 in such cases (because there *was* voicing, albeit not throughout). Instead, we rely on voicing fraction to modulate its effect. In PCA, this means those frames will have an intermediate voicing feature and a possibly outlier F0 – if it's inconsistent, PCA might not emphasize it strongly. (If we found including such partial F0s is problematic, an alternative would be to require e.g. >50% voiced to report an F0; but our default is to include any detected F0 to maximize information).

- **Formants in unvoiced frames:** Typically, formant tracking in unvoiced segments yields nonsense (because the vocal tract is excited by noise or silence, LPC might still return resonances but they are not reliable). By setting F1–F3 to 0 for unvoiced frames, we effectively drop that information. We also consider including a **"fraction of frame with valid formant"** feature, but that correlates strongly with voicing (since formants are only defined in voiced sounds for our purpose). Thus, voicing fraction is sufficient proxy for "formant validity." We also take a conservative approach: we only compute formants on sub-windows that are marked as voiced (we can couple formant tracking with a voiced frame test – e.g. run LPC only on portions where some periodic energy is present). This further ensures we don't get random formant values in unvoiced parts. If an unvoiced consonant has a vocal tract shape that technically has resonances (it does, but the source is noise), we are **not** including those as formant features because they wouldn't be reliable or consistent (and the spectral shape features will already capture any resonant peaks in fricatives).

- **Voiced-fraction as feature:** Yes, we explicitly include the voiced fraction per frame (or a derived measure like "% of subwindows voiced" or a 0/1 flag). This avoids bias because the model won't assume that, say, an F0 of 100 Hz with voicing=0 is a low pitch – it will know voicing=0 means that 100 Hz value likely won't occur (in practice we set F0=0 in voiceless frames as said). The inclusion of voicing fraction also helps in PCA because one principal component can largely separate voiced vs unvoiced conditions (e.g. capturing the presence of periodic energy), rather than forcing the PCA to approximate voicing via other features like energy or centroid alone. In sum, **unvoiced frames will have:** voicing fraction = 0, F0 = 0, F1=F2=F3=0, and typically lower energy and higher spectral flatness. The PCA can learn to characterize this "silence/unvoiced" cluster distinctly. We will verify that this handling introduces no bias – essentially, by treating 0 as a default filler and providing the voicing indicator, we ensure these zeros do not pull the mean or covariance in a misleading way (they will mostly lie on a separate subspace flagged by voicing=0).

In implementation, after computing features for all frames, we will run a quick check: any frame marked unvoiced (voicing=0) should indeed have F0 and formants as 0; any frame voiced (voicing ~1) should have nonzero F0 and plausible formants. If any inconsistencies appear (e.g. voicing=0 but F1 nonzero due to a glitch), we will rectify by zeroing those values or adjusting threshold. This rule-based imputation ensures a clean separation between voiced and unvoiced frame feature profiles.

Quality Control and Sanity Checks

Before integrating audio features into the PCA, we will perform several quality-control checks per sentence (and globally) to catch any abnormal feature values or extraction failures. Key checks and thresholds include:

- **NaN/Inf check:** No feature value should be NaN or infinite. This can happen if, for example, log of zero occurs. We mitigate this by adding a very small floor in log calculations (e.g. floor power at 10^{-12} before log) so we shouldn't get $-\infty$. But we will still scan each feature matrix for NaNs/Infs. **Action:** If any are found, investigate their source. Likely causes: a bug in formant finding (e.g. no roots found), or a division by zero in spectral flatness (if spectrum is zero). We will replace NaN with 0 or a benign value and mark that frame, or adjust the algorithm (e.g. if flatness undefined because all zero energy, just set flatness to 0 which is logically "no signal"). In practice, since we always have some signal in speech frames, NaNs should be rare.

- **Feature range sanity:** We expect each feature to fall in a plausible range. We will compute summary stats (mean, ±3 std) for each dimension across the dataset. Red flags: e.g. an MFCC that is extremely large in magnitude (could indicate an unstable numerical issue), or negative values where only positives make sense. For instance:

- MFCC coefficients typically range roughly between -20 and +20 (depends on log energy, etc.). If we find an MFCC with range say -100 to 100, something's off (perhaps not liftered or a scaling issue). We'll confirm MFCC ranges look reasonable (this also checks that we did not accidentally include an unnormalized energy that dominates).
- Log-mel energies: since we log10, values might be around, say, 0 for 1 (reference) down to -inf for 0 energy. In practice, with floor, we might see minimum ~-12 (if floor at 1e-12). The maximum log-mel might be 0 (if 1.0 is reference) or higher if input wasn't normalized. We'll ensure the audio amplitude scaling is consistent (we might normalize input wave amplitudes similarly across sentences so that absolute log-energy is comparable; if not, at least block scaling will adjust).
- Pitch (F0): should lie in [50, 300] Hz by design (for voiced frames). We'll verify no F0 values outside this (if there are, likely octave errors that slipped through). If a few frames have F0 doubling (e.g. ~400 Hz), we might tighten the voicing threshold or median filter more. Likewise, formants: F1 should typically be in 200–800 Hz (for adult voices), F2 ~800–2500 Hz, F3 ~1500–3500 Hz. We'll scan the formant outputs: if we see obviously swapped values (e.g. a frame with F1=1500, F2=800, indicating an error), or values out of human range (F1 50 Hz or F2 5000 Hz for a human speaker), those indicate tracking errors. We expect our median-over-subwindows to mitigate momentary errors, but if any persistent out-of-range values occur, we might incorporate a rule to clip them or drop that frame's formant data. For example, we might cap F1 at 1000 Hz, F2 at 3500 Hz for safety – values beyond are likely noise. Frames with such anomalies can be flagged; if they are rare, we could optionally zero out the formant features for those frames (treat like unvoiced for formant purposes). However, given that PCA is robust to a few outliers in a large dataset, we likely don't need to remove frames entirely – just ensure they don't produce extreme values that skew scaling.
- Spectral centroid: should lie between ~500 Hz (for very mellow sounds) and ~~5000–6000 Hz (for very bright sounds like /s/). If we see centroid consistently at 0 or extremely high ~8000, something's wrong. We also check **spectral flatness**: 0 to 1 (or in dB, -∞ to 0 dB). Flatness = 1 (0 dB) means white noise; =0 means a perfect tonal signal. Values outside [0,1] indicate a calculation issue (e.g. negative power).
- Voicing fraction: by definition 0–1. If we ever see a value >1 or <0, that's a bug.

- Energy: Should be <= 0 in log10 if we normalize reference to 0 dB. If not normalized, some reference is needed. We can simply check that energy values cluster around a reasonable range (e.g. if audio is 16-bit PCM normalized to ±1, a typical frame energy might be on order 1e-3 to 1e-1, so log10 ~ -3 to -1). If an entire sentence's energy is drastically lower or higher, it might indicate varying recording levels – PCA's block scaling will handle it, but we might consider normalizing per sentence or speaker if needed. At least, we note any sentences with unusually low overall energy (which could affect SNR and feature reliability).

- **Voiced/unvoiced frame proportion:** We expect a realistic distribution of voiced frames per sentence (depending on content, maybe ~40–60% voiced for English). If for a particular sentence or actor the voiced fraction is extremely low or high (e.g. one actor shows 90% voiced frames across sentences, or 10%), that could indicate the voicing detection threshold might be mis-tuned for that voice (e.g. a high-pitched voice might not cross our default threshold or a very low voice might confuse zero-crossing logic). We will examine the average voicing rate. If needed, adjust the voicing detector (e.g. lower threshold for female voices). However, since we use an autocorrelation method with probability output, it should generalize well; we can dynamically adjust threshold to match an expected voiced frame rate (~50%) if we detect clear skew. For example, if an entire sentence is marked unvoiced (voicing fraction all zeros) but we know it contains voiced sounds, that's a fail – we'd lower the voicing threshold and recompute for that sentence. Conversely, if voicing fraction is 1 for every frame (implying continuous voicing, which is unlikely because English sentences have voiceless consonants), we might have threshold too low (labeling everything as voiced due to background noise). We can incorporate a basic **voice activity detection** check: frames with extremely low energy should never be marked voiced; if our voicing detector violates that, we refine it (e.g. require energy > some floor to consider voicing).

- **Formant continuity and validity rate:** We will compute the percentage of voiced frames for which we obtained valid formants. Ideally, that should be high (close to 100% in voiced regions, since we use robust median). If we find, say, that only 70% of voiced frames have "sensible" formant values and 30% were too erratic (out of range) and effectively got zeroed, that indicates the LPC parameters might need adjustment (e.g. increase order or window length, or exclude problematic frames differently). We may tweak LPC order per speaker if needed (e.g. if a speaker has a higher pitch or different vocal tract, sometimes a slightly higher model order resolves tracking). We might set a threshold: if <90% of voiced frames yielded good formants, try increasing LPC order by 2 or adjust pre-emphasis. If formant tracking remains unreliable (due to fast articulation or SNR issues), we might consider dropping formant features for that speaker to avoid injecting noise.

- **Mel bandwidth energy coverage:** We will ensure that all mel bands carry some energy across the dataset. If we find that, for example, the highest mel band (covering ~7–8 kHz) has near-zero energy for all frames, it might be due to aggressive low-pass filtering or simply the dataset lacking energy there. That's not a problem per se, but it means those dimensions are almost constant (or just noise). The PCA will naturally assign them low variance. However, if they truly have zero variance, they provide no information and could be removed. We might drop the top few mel bands if they are always zero (or ensure any frequency filtering is adjusted so that mel band isn't entirely zero). Similarly, if the lowest band is always extremely strong (e.g. if there's DC offset or hum captured), that could dominate – but our pre-emphasis and high-pass (~50 Hz) should remove DC components. We also check that the mel energy distribution per frame looks reasonable (e.g. not all concentrated in one band due to an artifact).

- **Thresholds and actions summary:** As a concrete rule, we'll flag any feature dimension whose values exceed 5 σ from the mean or fall outside physically plausible bounds. If flagged, we inspect those frames. For minor single-frame glitches, we may simply leave them (PCA can accommodate a few outliers, and our block scaling will reduce their influence). For systematic issues (e.g. all frames of a sentence have MFCC[5] = 50 which is out of expected range), we identify the cause (perhaps an error in computing that coefficient) and recompute if necessary. We prefer not to do heavy handed clipping on a large scale (to avoid biasing distributions), but small adjustments (like capping formant values at a max) essentially remove unphysical extremes and are acceptable.

- **PCA block variance check:** After block scaling (each modality block scaled by $1/\sigma_1$ as described), we can double-check that no single feature dominates. For instance, if our audio block's top singular value is dominated by the energy feature, then after scaling, energy will become quite small relative to others (since we divide by σ_1 of audio block). But within the audio block, one feature could still have much larger variance than others, which might skew the PCA within the audio subspace. To detect this, we look at the variance of each feature (after our processing, before PCA). If one feature has, say, >50% of total variance of the audio set, that's problematic. Typically, log-energy might have larger variance than individual MFCCs. However, since we do not z-score each feature (by design), some disparity is expected. It will be handled by PCA (first principal component of audio will align with that dominant feature). That's okay as long as that dimension truly carries meaningful info. If we see, for example, that "energy" is 10× more variable than any spectral coefficient (which could happen if speaking volume varies a lot), we might consider normalizing energy to a comparable scale (maybe implicitly done by block scaling, but block scaling sees the combination). In practice, block scaling uses the top singular value of the entire audio feature set, which will be heavily influenced by the largest-variance direction (likely energy + some spectral mix). That will downweight the whole audio block appropriately relative to MR and video. So we should be fine. Nonetheless, we list it as a check: if needed, we could reduce the range of energy by applying a compressive transform (we already use log, which helps).

These QC checks will be run on a per-sentence or per-speaker basis to identify any anomalies early. If thresholds are exceeded, our actions range from adjusting extraction parameters, removing or replacing outlier values, or at least annotating those frames so we interpret PCA results with caution. Overall, these steps ensure the feature set is reliable and interpretable, which is critical for downstream analysis of variance explained and correlations.

Ablation Experiments (Features and Parameters)

We propose 5 small toggles (ablation conditions) to assess the sensitivity of H1 and H2 outcomes to the audio feature design. Each ablation involves altering one aspect of the audio features while keeping the rest of the pipeline the same, then evaluating the impact on audio reconstruction VAF% (H1) or MR/Video reconstruction error (H2). We will consider changes "material" if they produce a substantial change in metrics beyond expected noise (e.g. more than a few percentage points of VAF, or a notable fraction of the current performance level).

1. **Mel filterbank resolution – 40 vs 64 bands:** Reducing to 40 mel bands (from 64) will test if high-frequency resolution is contributing to reconstruction. 64 mel provides finer detail in higher formants and fricative noise [6], whereas 40 mel is coarser (standard in many systems). **Expectation:** If articulator-to-audio mapping benefits from extra spectral detail, using 64 should slightly improve H1 audio reconstruction (especially for high-frequency components), perhaps

raising VAF by a small amount (e.g. on the order of 0.5–1 percentage point absolute). For H2 (MR/Video reconstruction), the difference might be minimal or slightly positive, because extra audio detail could help the PCA find modes that align with subtle articulator motions. A **material** change here would be >+1% VAF or so. If the difference is <0.5% VAF, we deem it negligible (the null p-value is ~0, so a 0.5% change might be at the edge of detectability given baseline 5–6% VAF). We also check H2: e.g. if using 64 mel vs 40 mel changes the loading-space correlation R between true and reconstructed video by more than ~0.05 or changes SSE by >5%, that's material. We predict maybe a slight benefit for richer mel (if any). If we see virtually no change, it means the additional bands were redundant given other features.

2. **Number of MFCCs – 13 vs 20 coefficients:** Here we compare using 13 cepstra (including C0) versus 20. 13 is traditional and may miss some spectral detail (particularly above ~4 kHz), while 20 captures more nuance (e.g. F4, F5 regions). **Expectation:** If articulatory data (especially MR) can predict higher-frequency details, we might see a small uptick in H1 VAF with 20 MFCC. However, those higher cepstral components might also be noisier/unpredictable, potentially adding dimensions that PCA treats as residual. A material improvement would be on the order of +0.5–1% absolute VAF. We expect perhaps a modest increase in VAF or no significant change – if MR/video can't predict those fine details linearly, the extra MFCCs will mostly add reconstruction error variance (which PCA might just assign to low-variance components). For H2, including more MFCCs (which emphasize finer acoustic variation) could slightly influence the face-MR reconstruction metrics: possibly the PCA will allocate a bit more weight to audio variation that doesn't align with face/MR, which *could* even hurt MR/video reconstruction slightly. If adding MFCCs degrades H2 metrics (say increases video reconstruction SSE notably or reduces correlation by >0.02), that suggests those added features were essentially noise from the perspective of MR/video. In contrast, if H2 metrics remain the same while H1 improves, that's a win. We'll treat >+1% VAF or >5% change in SSE as material.

3. **Within-frame deltas – with vs without Δ features:** We will test dropping the within-frame Δ coefficients for MFCC (and similarly not using Δ mel). In this condition, each frame's feature vector only contains static coefficients (MFCC or mel averaged over frame). This ablation reveals whether capturing intra-frame dynamics improves the linear mapping. **Expectation:** Without Δ, the PCA might lose some sensitivity to articulator velocity (e.g. a rapid tongue movement within 62 ms might not register in static features). We anticipate **some** impact: possibly a reduction in H1 audio reconstruction (since dynamic cues can help predict transients). It might be modest – e.g. a drop of 0.5–1% VAF – because 62.5 ms is not very long, and many frames may be quasi-static. For H2, we expect temporal context (even within-frame) helps slightly: Scholes et al. found that adding contiguous frames improved reconstruction [19] [20]. Our within-frame Δ is a mini version of adding context. So removing Δ might slightly reduce the coupling between modalities, possibly visible as a small drop in loading-space correlation or an increase in SSE for MR reconstructed from video and vice versa. A material change would be if, for example, VAF drops from ~6% to 4% (i.e. ~2 pp, which is quite noticeable, ~33% relative drop), or if video reconstruction error SSE increases significantly (>5–10%). We suspect the effect will be noticeable but not huge, perhaps on the borderline of material (maybe ~0.5–1% absolute VAF drop). If it's below 0.5%, then within-frame dynamics might not be crucial at this frame rate – an interesting finding itself.

4. **Raw spectral snapshot – included vs excluded:** We gauge the value of the raw FFT features by removing them. In the rich set, this means comparing the full trimodal PCA with and without the 128-point raw spectrum appended (keeping mel/MFCC, etc. the same). **Expectation:** The raw spectrum might be partially redundant with mel/MFCC, but it could carry fine details (like exact spectral peaks) that the PCA could use. If those details aren't predictable from MR/video (likely

not in a linear sense), then including raw spectrum might actually add mostly noise from the perspective of MR+Video prediction. We suspect excluding it *might* slightly **increase** H1 VAF (paradoxically) because the PCA will focus on more predictable mel/formant features instead of trying to also account for unpredictable spectral variance. Conversely, including it increases the audio feature variance (most of which MR/Video cannot explain), which could reduce the proportion of variance (VAF) explained by MR+Video in the audio space. So we wouldn't be surprised if Tri-real VAF% is a bit lower when raw spectrum is included, simply because there's more total variance in audio features. We will check Tri-real vs Tri-real-minus-raw VAF. A material change might be >±1% VAF. For H2, including raw spectrum likely has minimal positive effect on MR or video reconstruction – it could even slightly degrade the defined metrics if the PCA uses some degrees of freedom to capture audio-only variance. If, for instance, we see MR→Video correlation drop by more than 0.02 when raw audio is included (relative to using only mel/MFCC/formants), that would indicate the audio was "distracting" the PCA. If metrics stay essentially the same, then the raw spectrum was either largely ignored by PCA (low weight components) or it provided some orthogonal info that didn't interfere. We consider any change beyond normal variability (say >5% change in SSE or >0.05 change in correlation) as material here.

5. **Formant features – included vs excluded:** Finally, we test the impact of adding explicit formant and ratio features. We compare the PCA results with our full feature set vs a feature set with formant-related features removed. **Expectation:** If formant features are truly providing unique articulatory information not fully captured by mel/MFCC, then removing them might hurt performance. We anticipate that MR in particular benefits from formant features: since MR can directly predict formant movements, including them should boost audio reconstruction from MR. If we drop formants, MR will rely on broader spectral features to convey that info. A **material** effect would be a drop in H1 audio reconstruction VAF by perhaps >1–2 percentage points when formants are removed (especially in an MR-only reconstruction scenario). We might measure H1 for MR+Video vs audio; but to isolate the effect of formants, we can also look at how well MR-alone PCA modes predict the audio: with formants included, MR-alone should explain a decent chunk of variance in the audio-formant dimensions [2], boosting overall audio VAF. Without formants, MR-alone might do worse in capturing certain vowel distinctions. So we expect a reduction in audio VAF (and possibly lower Pearson r in the audio space) without formants. For H2 (visual target), including audio (with formants) might have made subtle improvements to MR→Video mapping; if formants were mostly aiding audio reconstruction, their removal might not significantly change video reconstruction fidelity. But it could: audio features like formants might reinforce the correspondence between face and MR (since face correlates to formants too to some degree). So possibly Tri-real vs Tri-ShufA difference might shrink if formants are removed (meaning audio no longer adds as much useful info). We consider it material if, for example, the **improvement** that real audio gives over shuffled audio in reconstructing MR/Video (H2) drops notably. Currently, we'd compare Tri-real vs Tri-ShufA performance; if excluding formants reduces that gap significantly, it shows formants were a key contributor. Concretely, if with full features Tri-real has, say, 0.10 higher loading-space R than Tri-ShufA for MR reconstruction, but without formants this drops to 0.05, that's a material halving of the audio's benefit. We'd quantify any such change.

In summary, we deem changes on the order of 10–20% relative difference in the metric as **material**. Given a baseline of ~5–6% VAF, a 1% absolute change is ~20% relative – clearly material [21]. For correlation (if baseline ~0.8 for face–MR loading correlation, a 0.02 change is 2.5% relative – mild, whereas 0.05+ (6%+ relative) is more noticeable). SSE changes are harder to gauge in abstract, but we use ~5% of baseline SSE as a rough threshold. These toggles are small and controlled; we expect mostly small effects, but anything exceeding the thresholds above would indicate that feature choice has a meaningful impact on integration performance, guiding us to refine the feature set accordingly.

Expected Patterns for H3: Modal Contributions per Feature Family

H3 will examine how each feature family's variance is explained by Video-only, MR-only, or combined MR+Video. We predict the following patterns, based on articulatory mechanisms and prior research on audio-visual speech:

- **Low-frequency spectral envelope (vowels/formants – e.g. MFCC low coeffs, formant features):** These should be best predicted by **MR (internal articulators)**. The vocal tract shape (tongue, pharynx, velum) captured in MRI directly determines formant frequencies [2]. For example, F1 and F2 of vowels have high linear correlation with tongue and jaw positions [2], which MR sees, but only partial correlation with facial appearance (the lips convey some info, especially for F2 via rounding, but not the whole story). So we expect feature families like formant frequencies and broad MFCC/mel variations to align strongly with MR dynamics. **Video-only** will be weaker here but not zero: visible lip height and width correlate with jaw and tongue to some extent (e.g. an open mouth often indicates a low vowel ⇒ high F1, a protruded lip often accompanies a back/rounded vowel ⇒ low F2). Indeed, Scholes et al. showed that facial cues alone can predict vocal tract configuration to a degree [20] [22]. We expect Video to recover maybe the gross formant patterns (perhaps capturing on average 50% of variance that MR can). **MR+Video combined** should outperform either alone for these features. The combination merges internal and external information – analogous to how adding lip info helps disambiguate tongue positions that produce similar acoustics. For instance, lip rounding vs unrounding can produce different F2 even if tongue is similar; MR gives tongue, video gives lip – together they nail F2. Thus, H3 should show a significant boost in variance explained for envelope features with MR+Video versus either modality alone. This reflects the complementarity: MR covers tongue and pharynx, video covers lips and jaw; both are needed for full vowel/acoustic coverage [23]. In articulatory terms, **F1** (jaw/tongue height) might be fairly well predicted by video alone (jaw movement is visible) [23], whereas **F2** (tongue front/back) is mostly invisible externally but partly signaled by lip shape – so MR carries most weight, with video adding a bit. Formant ratio F2/F1 likely heavily MR-driven but video can contribute for rounded vs spread distinctions. MFCC coefficients 1–3 (which roughly correspond to overall formant spacing and tilt) will likely follow a similar trend: MR > Video, with combined best. This aligns with the common generator hypothesis that face and vocal tract share underlying motor commands [23] – the internal articulations drive the acoustics, and the face moves in coordination, but not as directly.

- **High-frequency energy / fricative spectral features (e.g. spectral centroid for consonants, high mel bands):** These are related to place of articulation for fricatives and stop bursts. For fricatives, internal articulator shape (particularly tongue tip/blade position and grooving, which MR sees) plus lip shape (which video sees) both influence the spectrum [4]. Example: /s/ vs /ʃ/ – MR can see the tongue posture difference (alveolar vs postalveolar), and video sees lip rounding for /ʃ/. /ʃ/ has a lower-frequency spectrum (centroid) than /s/ [4]. **MR+Video combined** should reconstruct fricative spectral shape best, because MR can indicate the articulator placement and degree of constriction, while video adds cues like lip rounding/protrusion. **MR-only**: likely effective for many consonants' spectral features – for instance, the presence of a sibilant constriction or a velar constriction will be evident in MR and correlates with high-frequency noise production. But MR might miss some fine detail like whether teeth are visible (affecting high frequency), which video might indirectly capture (through lip opening). **Video-only:** can distinguish labial vs non-labial sounds easily (visible), and can somewhat distinguish coronal vs dorsal (e.g. for /θ/ vs /k/, video sees tongue tip on teeth sometimes, but for many it's tough). Video is good at recognizing bilabial sounds (which have a characteristic acoustic of silence then

broad burst – though silent intervals aren't spectral features per se). It may not predict the precise spectrum of a fricative as well as MR can, but it gives cues like "jaw very closed, lips spread" could indicate /s/. We suspect MR will outperform video for non-labial consonant spectra (since internal tongue shape dominates those), whereas for bilabials (like /f/ vs /θ/ vs /s/ differences), video might be on par or better (because /f/ involves visible lip-teeth contact, /θ/ tongue tip visible between teeth, etc.). In aggregate, high-frequency spectral features likely need both modalities: MR for tongue shape, video for whether lips are rounded or open. Thus, the **combined MR+Video** should exceed either alone for spectral centroid/spread of fricatives [4]. If H3 analysis is done per feature, we'd see e.g. spectral centroid variance: maybe MR explains a good portion (since place of articulation is mostly internal) but video adds the rest (especially distinguishing similar internal place with different lip shapes).

- **Voicing-related features (F0, harmonicity, spectral flatness differences):** These are primarily **source** characteristics. We expect neither video nor MR to excel at predicting F0 – it's controlled by laryngeal muscles not captured in our modalities (MR might show larynx height or shape indirectly but not vibration rate; video doesn't see vocal folds). So F0 feature variance will largely be **unexplained by either modality** (the evaluation-only null will show near-zero correlation, which is why H1 VAF is low overall). MR+Video combined also won't do much better, since combining two weak predictors doesn't magically solve it. However, **voicing state (binary)** can be somewhat inferred: video can detect the presence of vocal fold vibration only via extremely subtle cues (perhaps skin vibration or just by knowing the phonetic context – e.g. if they see a /b/ vs /p/ by timing). MR might show glottal opening for voiceless sounds (if the MRI slice covers the glottis – not sure if it does clearly, but possibly some part of it). The **energy feature** might correlate with voicing (voiced sounds often have more low-frequency energy, different energy distribution). But overall, F0 amplitude variations (intonation) and exact pitch will be basically independent of articulator motion under a linear model. We expect H3 to show very low explained variance for the F0 dimension from either modality (likely at chance level ~0% by video, maybe a few percent by MR if any). The combined might be slightly higher if, say, both modalities together give a hint (for instance, certain articulatory settings correlate with prosodic emphasis which correlates with slight F0 raising – but that's second-order). We note that some correlation could come from **intrinsic F0** effects: high vowels often have slightly higher F0 than low vowels on average [2], so if MR+Video know the vowel height, they could very weakly predict F0 differences. Also, if the speaker tends to raise their chin or tighten neck muscles with higher pitch (some do), video might indirectly correlate with F0. But these are minor. So expectation: **F0 features will be largely unrecovered (low R, low VAF) by either modality**, confirming that a linear PCA can't capture voice pitch – which is fine as it's not articulatory in origin. The **voicing (binary)**, however, might be significantly predicted by articulators: E.g., MR can indicate a spread glottis or a lack of periodic vocal tract resonance during unvoiced consonants; video can indicate when a consonant is voiceless by context (like seeing a /p/ explosion). We might see that frames classified as voiced vs unvoiced can be separated by the PCA using features like spectral flatness (which our model does include and which video/MR can predict somewhat – e.g. unvoiced frames have higher flatness which might correlate with certain face shapes). So voicing *state* could be moderately well predicted, but the actual F0 value (in Hz) not.

- **Energy (intensity):** The loudness of speech can correlate with articulatory posture: louder speech often involves larger mouth openings and exaggerated movements (to project sound). Video captures mouth opening degree directly, so it should correlate reasonably with frame energy (open mouth vowels are typically louder) [23]. MR captures velum opening (nasal coupling can reduce energy) and overall tract opening (which affects output gain). Also, if someone speaks louder, they might have a wider pharynx or mouth – MR sees that. So we expect both modalities to have some predictive power for energy. Possibly video might be slightly better for

energy (since lips/jaw position directly control radiated sound area). MR also sees tongue backing (which if the tongue is very back, it might reduce oral aperture, affecting energy). Combined MR+Video should do best, though energy might already be largely explained by either one (this might be somewhat redundant info). Perhaps face signals intensity by facial strain or bigger movements. We anticipate at least moderate correlation (not as high as for formants, but significant). If we quantify: maybe video-alone could explain e.g. 50% of energy variance, MR-alone 40%, combined 60–70%. If one had to guess, yes, MR+Video will give the best estimate of loudness (since they jointly reflect articulatory setting that influences acoustic efficiency).

- **Spectral shape features (tilt, flatness):** These are somewhat mixed – spectral tilt (related to voice quality, e.g. breathy vs pressed voice, and vowel type) may not be strongly predicted by either modality (because it has to do with glottal source too). But some articulatory configurations (like a wide open jaw yields a slight decrease in high-frequency tilt due to increased low-frequency energy) might correlate. Spectral flatness distinguishes voiced (harmonic, flatness low) vs unvoiced (noise-like, flatness high). That essentially ties to voicing again. So flatness being high is predictable by articulators as voiceless consonant presence (which MR/Video know moderately). Tilt might correlate with whether the speaker is producing a consonant (tilt flatter for noise) or a vowel (tilt steeper for voiced). Thus MR+Video could moderately predict these. We expect pattern similar to voicing: MR+Video combined separates noise vs voice better than either alone. Video might catch unvoiced (via visible articulator tension or known consonant shapes), MR definitely catches presence of a constriction that would cause turbulent noise (hence flat spectrum). So for flatness: MR likely > video (because internal constriction is a strong indicator of noise excitation). Combined best.

- **Overall: synergy (MR+Video > either) and unique contributions:** We anticipate **MR-alone** will excel at features that depend on hidden articulators (tongue, velum) – notably vowel formants, many consonant place distinctions. **Video-alone** will excel at features tied to visible articulators – labial consonant bursts, lip rounding effects on spectrum, jaw-related loudness and F1. For example, a bilabial stop's acoustic signature (silence + burst) might be predicted by video seeing the lips close and open, whereas MR sees relatively "nothing changing in the vocal tract interior except maybe tongue staying neutral" – video gives the key cue. So for features like burst energy or presence (not explicitly in our list except via high-frequency energy changes), video might contribute more. **MR+Video** will *at least* match the better of the two for each feature, and usually exceed because any aspect one modality misses the other can supply. This is in line with the idea that audiovisual integration occurs because face and vocal tract together provide a more complete representation of the speech production state [23]. Previous studies found that combining modalities improves reconstruction fidelity of one modality from the other [24] [19] – here we extend that to audio: MR+Video should best reconstruct audio features (H1), better than either alone (we expect e.g. MR+Video yields ~6% VAF, whereas MR-alone or Video-alone would be lower, maybe 3–4% VAF each, roughly). H3 will quantify those differences per family. Specifically, we expect: for **mel/MFCC** variance, MR+Video > MR > Video; for **formants**, MR+Video > MR >> Video (since video is relatively weak alone on formants but improves combined by adding rounding info); for **F0**, MR+Video ~ MR ~ Video ~ (very low) – none can explain well; for **voicing/flatness**, MR+Video > MR ~ Video (both can cue voicing, MR maybe slightly better if it captures glottal opening; combined helps resolve ambiguous cases); for **spectral centroid/high-freq**, MR+Video > MR ~ Video (with MR likely a bit better for non-labial fricatives, video helping especially for labial ones and rounding contexts).

These predictions are supported by the notion that the face and vocal tract provide complementary views of the "common generator" of speech [23]. The PCA trimodal analysis should reflect this: features tied to that common articulatory state (formant structure, broad spectrum) will be well captured by the

joint variation of face+MR, whereas features tied to independent source variations (pitch, voice quality) will remain largely unexplained. We will validate these expectations in H3 by examining, for each feature family, how much variance is captured by PCs derived from different modality combinations (e.g. do PCs from face-alone capture mel variations, etc.).

Ultimately, we expect MR+Video to outperform either alone on **almost all feature families** (since even if one modality is dominant, the other rarely hurts and often adds), aligning with the findings that using both facial and vocal tract info yields the best prediction of acoustic-relevant configurations [20] [23]. The degree of improvement, however, will vary by feature: minimal for F0 (none can do it), moderate for energy (both can do it, so combined just a bit better), and large for something like F2 or a fricative's spectrum (where each modality has a piece of the puzzle). These results will help interpret which aspects of speech audio are encoded redundantly vs complementarily in face and MR modalities.

Sources: These expectations are grounded in known articulatory-acoustic relationships [2], audio-visual speech findings (face cues suffice to recover some vocal tract info [20] but not all), and the common generator theory [23] suggesting that combining modalities yields the complete picture. Each feature family's behavior in H3 should reflect these established principles of speech production and perception.

---

[1] [10] [11] [12] [13] Speaker Identification Using Pitch and MFCC - MATLAB & Simulink
https://www.mathworks.com/help/audio/ug/speaker-identification-using-pitch-and-mfcc.html

[2] Relationship between tongue positions and formant frequencies in female speakers - PubMed
https://pubmed.ncbi.nlm.nih.gov/26827037/

[3] Spectral Descriptors - MATLAB & Simulink
https://www.mathworks.com/help/audio/ug/spectral-descriptors.html

[4] Differential contributions of the two cerebral hemispheres to temporal and spectral speech feedback control | Nature Communications
https://www.nature.com/articles/s41467-020-16743-2?error=cookies_not_supported&code=0978563d-2140-44cf-8bf0-b62f9eae6f8a

[5] Formants - Acoustic Phonetics
https://home.cc.umanitoba.ca/~krussll/phonetics/acoustic/formants.html

[6] [21] [2104.11598] Reconstructing Speech from Real-Time Articulatory MRI Using Neural Vocoders
https://ar5iv.org/pdf/2104.11598

[7] [8] [9] [15] [16] Formant Estimation with LPC Coefficients - MATLAB & Simulink
https://www.mathworks.com/help/signal/ug/formant-estimation-with-lpc-coefficients.html

[14] How to decide filter order in Linear Prediction Coefficients (LPC ...
https://stackoverflow.com/questions/61519826/how-to-decide-filter-order-in-linear-prediction-coefficients-lpc-while-calcu

[17] audioFeatureExtractor - Streamline audio feature extraction - MATLAB
https://www.mathworks.com/help/audio/ref/audiofeatureextractor.html

[18] spectralCentroid - Spectral centroid for audio signals and auditory ...
https://www.mathworks.com/help/audio/ref/spectralcentroid.html

[19] [20] [22] [24] scholes-et-al-the-interrelationship-between-the-face-and-vocal-tract-configuration-during-audiovisual-speech (1) 1.pdf
file://file-NsDLXj1wzcVAcuY6fe7xj9

23 [PDF] Analysis of the variability of the coupling between facial motion and ...
https://www.cefala.org/issp2006/cdrom/articles/moreira.pdf