ChatGPT

1. One-Dimensional Audio Features for Low-FPS Speech Integration

For short speech utterances (2–3 seconds) synchronized at 16 frames per second, the most suitable audio features are those that **compactly represent the acoustic content per frame** while correlating with articulatory events. Common choices include **Mel-Frequency Cepstral Coefficients (MFCCs)** and related cepstral features, which succinctly encode the short-term speech spectrum in a low-dimensional form [1]. MFCCs (typically 12–13 coefficients plus energy) are widely used in speech processing to capture the spectral envelope, making them a logical candidate for compressible audio representation in multimodal settings [1]. Another option is to use **formant frequencies** (e.g., F1, F2), the resonant frequencies of the vocal tract. Formants directly relate to vocal tract configurations and have been used in articulatory studies as low-dimensional acoustic descriptors. Real-time MRI speech research often tracks formants automatically alongside imaging data [2], underscoring their relevance. Similarly, the **fundamental frequency (F0)** or **pitch** is a 1D feature capturing voicing and prosody; it is routinely tracked in multimodal speech analysis [2]. Pitch indicates voiced vs. unvoiced frames and intonation, providing information absent in silent facial/magnetic resonance images. **Intensity (energy/envelope)** is another simple 1D feature – the short-time root-mean-square energy of the audio – which correlates with mouth opening and loudness [3]. In summary, features like MFCCs (or their first principal component), formant frequencies, pitch, and energy are all viable 1D descriptors. Each can be computed per 62.5 ms frame to align with the 16 Hz imaging, yielding a compact time-series to integrate with video/MRI. These features have been utilized in speech science and multimodal learning due to their compactness and interpretability. For example, a recent audio-visual dataset release included both denoised audio and derived low-dimensional representations, noting that such features are "more suitable for various forms of research" than raw waveforms [4] [5]. Ultimately, **the "best" feature may be a combination** – e.g. a small vector per frame comprising energy, pitch, and formant-related parameters – as this captures complementary aspects of the speech signal in a form amenable to PCA integration.

2. Good vs. Bad Features: Reconstructability (H1) and Added Value (H2)

A "good" audio feature for this multimodal PCA framework is one that **provides complementary information** to the MRI and video, rather than duplicating what those modalities already convey. In terms of H1 (audio *reconstructable* from video/MR) and H2 (audio provides *added value* to reconstructing video/MR), an ideal feature strikes a balance: it should not be completely predicted by the articulatory data (otherwise it's redundant), yet it should correlate with cross-modal events enough to be relevant. For instance, **voicing-related features** like pitch are largely invisible to facial and MR imagery, so they offer new information (high H2) and are not easily reconstructed from silent visuals (low H1). Including pitch has the potential to improve multimodal reconstruction by signaling when vocal fold vibration occurs (e.g. distinguishing a voiced vs. voiceless consonant that might look similar on video). In contrast, a feature like overall **loudness (intensity)** is partly correlated with mouth opening – as the mouth opens wider, sound intensity tends to increase [3]. Such a feature would be **partly reconstructable from video** (moderate H1); if it's too predictable from lip/jaw motion, it adds little new information (low H2). Empirical studies of audio-visual speech confirm these intuitions: Yehia et al.

(1998) found that a large fraction of facial motion during speech could be predicted from acoustic variables, and vice versa, but not all – indicating some acoustic features (especially those related to vocal tract resonances) share common information with facial movements, while others do not [6] [7]. In one study, linear regression could recover internal tongue movements (EMA sensor data) quite well from both facial motions and acoustic parameters [6]. This implies features tightly linked to articulation (e.g. formant frequencies shaped by tongue/jaw position) might be **"too good" (too correlated)** – they risk being largely reconstructable from the MRI/video alone, thus offering limited added value. On the other hand, features capturing more independent dimensions of speech (e.g. voice source characteristics or high-frequency spectral details) are "good" in that they inject novel information (improving H2) and are not fully determined by the other modalities (avoiding trivial H1). Additionally, a good feature should be **stable and reliable**, not dominated by noise. For example, a random spectral detail that fluctuates erratically frame to frame would be a "bad" feature: it contributes variance that is essentially noise, offering neither reconstructable structure nor useful new signal. By contrast, a **consistent acoustic parameter** that tracks the speech content (like a formant trajectory or harmonic amplitude) is meaningful. In sum, **a good feature is one that the other modalities cannot entirely "guess" (preventing redundancy), but that still aligns with cross-modal speech dynamics sufficiently to enhance the joint model**. It should improve the PCA's ability to represent the multimodal data (high H2 contribution) without merely re-stating what's already in the images (low H1 redundancy).

3. Using H1 and H2 to Guide Feature Selection

The hypotheses H1 and H2 provide a framework for choosing audio features that optimally complement the video and MRI. **H1 (audio is reconstructable from video/MR)** suggests evaluating how much of a candidate audio feature's variance can be explained by the other modalities. If a feature yields a high H1 (almost entirely predictable from the articulatory data), it indicates redundancy – such a feature is likely not worth including, since the PCA could already capture that information from MRI/video alone. For example, if one considers an audio feature like the **amplitude envelope**, which correlates with visible mouth movements, a substantial portion of its variation might be recovered by a linear mapping from lip aperture (as shown by prior studies where mouth area and audio intensity co-vary [3]). In that case, H1 is strong and the feature might be deemed "bad" (it doesn't add a new dimension to the joint subspace). **H2 (audio improves reconstruction of video/MR)** points to the feature's unique contribution: does adding this feature help the PCA represent the *other* modalities better? A feature with high H2 is one that carries information that was missing in the video/MR alone. For instance, including a **voicing feature** could help the PCA distinguish frames where the vocal tract configuration is the same (say, a voiceless vs. voiced consonant in MRI looks similar), but the audio feature differentiates them – thereby improving the reconstruction or discrimination of those MRI/video frames when mapped back from the PCA space. In practice, one would **prefer features that have low mutual redundancy and high complementarity**. This can be tested by measuring cross-modal correlations or performing ablation experiments: if removing the audio feature significantly degrades the combined PCA's ability to reconstruct or represent the video/MR, then the feature had high added value (H2) [6]. Conversely, if removing it makes little difference, likely it was mostly redundant (high H1). Another way H1/H2 guide selection is via **multi-modal variance analysis**: one can examine how much variance in the joint PCA components comes from the audio feature vs. video/MR. A "good" feature will manifest as a component that mixes modalities (indicating it contributes to explaining covariance between audio and image data), whereas a "bad" one might either form a trivial separate component (explaining only itself) or not appear at all until very late components. In summary, H1 urges us to avoid features that lie in the same subspace as the visual articulatory data, and H2 urges us to find features that extend that subspace usefully. Practically, this means favoring audio descriptors that capture aspects like **voice timing, spectral cues of articulation, or prosody** that the MRI/video alone miss. We should also

ensure the feature is **reliably measured** – poor measurement can make even a theoretically useful feature behave like random noise, which neither H1 nor H2 would favor.

4. Optimal Number of Audio Features per Frame

Choosing how many audio feature dimensions to use per frame involves a trade-off between capturing sufficient acoustic information and maintaining balance with the high-dimensional visual data. In a concatenated PCA, if too **few audio features** are used (e.g. a single scalar), the audio modality may carry negligible weight relative to tens of thousands of image pixels. On the other hand, too **many audio features** could over-emphasize audio variance if each feature contributes unique variance. A practical approach noted in the literature is to use a *small number of dominant components* of the audio. For instance, research on PCA-based audio coding has shown that just a handful of principal components can capture most of the information in short-time audio spectra – often the first **4–6 components** are sufficient to reconstruct speech with high fidelity [8]. Indeed, one study reported that preserving the first 6 PCA components of audio frames covered essentially all important spectral information (the 5th and 6th contributed only minor improvements) [8]. This suggests that on the order of **5 or fewer audio features per frame** might be near-optimal for speech in a short window. Many audio-visual systems historically use about 13 MFCCs (plus perhaps pitch) for the audio frontend, but since MFCCs are highly correlated, their effective dimensionality is smaller. In a multimodal PCA context, one might start with **a small feature vector (e.g. 2–5 features)** that includes the most significant audio parameters (say, energy + first MFCC or first formant + pitch, etc.). This size is large enough to encompass multiple acoustic dimensions (spectral shape, F0, etc.) yet small relative to image pixels. Keeping audio features minimal also helps ensure the **PCA doesn't overfit the audio modality** at the expense of the others. It is also important to consider **scaling**: if each image frame has, say, 96,000 pixel features and we include only 1 audio feature, the audio's variance might be dwarfed unless scaled up. Thus, sometimes the **effective number of features** can be increased by scaling or replicating features. Rather than literally duplicating features, one can include a few independent audio descriptors. As a concrete example, one could choose **3 features per frame** (e.g. [F0, first formant, intensity]) which has been common in some audiovisual speech analyses. This number is small enough to avoid upsetting the PCA balance, but enough to represent key acoustic dimensions. Ultimately, the optimal number may be determined empirically by testing the **explained variance ratio** in the joint PCA. If adding a second or third audio feature significantly increases the total variance explained (or improves cross-modal reconstruction), it's worth including; beyond a certain count, returns will diminish. The goal is to reach a point where **each modality's variance is comparably represented** in the initial principal components. In summary, a *few* well-chosen features per frame (on the order of 1–6) is typically optimal, with literature evidence pointing to **4–6 principal features capturing most speech information** [9] [8]. This keeps the audio modality's dimensionality in the same ballpark as the degrees of freedom of articulatory motion (which are also relatively low, despite the high pixel count).

5. Features Robust to Modality Imbalance and Normalization

When one modality (images) has orders of magnitude more raw features than another (audio), normalization is critical. In an unnormalized joint PCA, the modality with more features can dominate the principal components simply due to sheer number. The most **robust audio features under such imbalance** are those that concentrate significant variance into each feature and can be scaled to be commensurate with image variation. Features derived from **spectral energy distributions** – for example, the first principal component of the audio spectrogram – tend to have high variance and thus

hold their own. One study found that the **first PCA component of short-term audio spectra often carries over 95% of the signal's energy/information** [9]. Using such a feature means the audio modality's information isn't spread thinly across many dimensions; instead, it's packed into one or few dimensions with relatively large variance. This makes the feature less likely to be swamped by thousands of low-variance pixel features. On the flip side, features that are inherently low-variance or binary (e.g. a voicing flag 0/1) could be problematic – their dynamic range is small, and without careful scaling they may be treated as near-constant by the PCA when alongside huge pixel variation. **Normalization strategies** can mitigate these issues. A common practice is to **z-score or whiten each feature** (subtract mean, divide by std) so that each dimension contributes ~1 unit variance a priori [10]. However, even if each single audio feature is normalized to unit variance, 96,000 image features each with unit variance would collectively still dwarf the audio in total variance. To counteract this, one can give the audio features a larger weight or include multiple related features. For example, using a small set of **filter bank amplitudes** (say 10 mel-band energies) instead of one MFCC yields ~10 dimensions for audio; if each is normalized, the audio block's total variance is ~10, which is still far less than 96,000 from images, but an order-of-magnitude boost over a single feature. Among candidate features, those that are **less sensitive to arbitrary scaling** are preferable. **Formant frequencies** measured in Hz can have a range of several hundred Hz – this range can be normalized to a similar scale as pixel intensity variations (which might be normalized to a few standard deviations). **Log-energy or log-power** features also tend to have reasonably large dynamic ranges (since silence vs loud speech can differ by tens of dB). They can be made commensurate with image intensity variances by scaling (e.g., multiply or divide by a constant) without losing meaning. In contrast, a feature like **zero-crossing rate** (proportion of sign changes in the waveform per frame) yields a number between 0 and 1 that might show subtle changes; it could easily be washed out in a combined PCA unless amplified. Therefore, features with an intrinsically broad scale or those that can be **amplified without distortion** are more resistant to imbalance. Another aspect is **consistency across frames** – if a feature has wildly variable variance from frame to frame (i.e., sometimes big spikes, sometimes flat), it complicates global normalization. Features such as MFCCs or LPCs are computed with stabilized algorithms and tend to produce steady ranges for voiced speech frames, making them easier to normalize reliably across the dataset [1]. Finally, it's worth noting that some multimodal PCA approaches explicitly apply weighting to modalities (so-called *weighted PCA* or block normalization) to ensure each modality contributes proportionally. If such a scheme is used, it can make the choice of feature less precarious. But absent explicit weighting, one should select features that **carry maximal information per dimension** and then consider duplicating or scaling them to boost the audio modality's overall variance. In summary, features concentrating audio variance (like principal spectral components, formant frequencies, or energy on a log scale) are most robust to normalization issues, especially when combined with appropriate scaling. By contrast, very low-amplitude or numerous fine-grained features would require aggressive normalization to avoid being overwhelmed by the high-dimensional image data.

6. PCA Architecture and Constraints on Feature Design

The project extends an existing PCA framework that already combines facial video and rtMRI, likely by **concatenating modality feature vectors for each timeframe and performing a joint PCA** on the concatenated vectors. This "early fusion" PCA means that for each frame (time-synchronized across modalities), a single feature vector is formed (e.g., [MRI pixels; video pixels; audio features]), and principal components are computed on the covariance of these large fused vectors. Such an architecture was used, for example, by *Stewart et al.* (2020) to jointly analyze midsagittal MRI and video frames of speech using PCA [11]. In their case, the PCA on concatenated images showed that one modality could be approximately reconstructed from the other, indicating the PCA captured coupled

motion [11]. This **joint PCA** imposes several constraints on feature design: **(a) Synchronized Sampling:** All modalities' data must be frame-aligned in time. The audio features must correspond exactly to the same 16 fps timeline as the MRI and video frames. Any misalignment (even a few frames) can degrade the PCA's ability to find common structure. Thus, audio features likely need to be extracted on sliding windows aligned to each video/MRI frame timestamp. **(b) Linear Representation:** PCA is a linear method, so it will best capture linear correlations among features across modalities. This encourages choosing audio features that relate linearly (or at least approximately linearly) to articulatory kinematics. For instance, a formant frequency might change roughly linearly with certain tongue/jaw displacements, making it suitable. A highly nonlinear feature (say, a phoneme class label) would not integrate as smoothly. If the architecture simply concatenates raw pixels with audio, it assumes all features are continuous numeric variables. **(c) Scaling and Units:** Because PCA is influenced by variances, all features must be put on comparable scales (often via pre-normalization). Each modality may have different units (pixel intensity, perhaps already unitless; audio features in Hz or dB). As noted, typically one would z-score each modality or each feature type before concatenation [12]. The architecture might therefore require the audio features to be normalized (zero-mean, unit-variance) so that the PCA isn't biased to one modality's units. **(d) Dimensionality Constraints:** In concatenated PCA, the total feature dimension equals the sum from each modality. If MRI+video contribute, say, ~100k dimensions and audio adds, e.g., 5, the PCA will produce at most 100k+5 components. Practically, one might perform preliminary PCA or compression on each modality (for instance, perhaps the existing framework already does PCA on video and MRI separately, then merges those scores) – the question suggests "joint PCA across concatenated modalities," so likely no prior dimensionality reduction except possibly some ROI extraction. Regardless, the audio features chosen should be *few* enough to not explode the total dimensionality, but *informative* enough to influence the first several principal components. **(e) Common Variance vs. Private Variance:** In a concatenated PCA, if an audio feature varies completely independently of the images, PCA will treat that variation as orthogonal components. This doesn't violate PCA per se, but it means some principal components will be *audio-only* while others are *image-only*. The framework likely aims for *coupled components* that span modalities (for example, a PC that represents a certain articulatory gesture which has both an MRI pattern and an audio signature). Therefore, features should be designed to encourage shared variance – if audio has a slight lead or lag relative to the video, for instance, one might adjust for that so that peaks align. If audio features have too much modality-specific noise, PCA may yield separate components for them, reducing multimodal integration. In practice, researchers have addressed this by including temporal context or pairing frames. Stewart et al. included consecutive paired frames in the PCA sample to implicitly incorporate temporal continuity [13] [11]. While that is a specific tweak, it highlights that the PCA architecture expects *structured, correlated inputs*. Hence, audio features may benefit from a bit of temporal smoothing (so that they align in phase with the slower frame rate visuals). **(f) Feature Contiguity:** If the PCA algorithm is implemented in a straightforward manner, it doesn't "know" which features belong to which modality – it just sees one long vector. However, sometimes after PCA, one examines the loaded components and maps them back to modality segments. Designing audio features that are clearly separable (in index) from image features (e.g., appended at the end of the vector) is trivial, but one must ensure that this ordering and any scaling are correctly handled by the software. In summary, the extension of a joint PCA means the audio features must be **co-temporal and numerically compatible** with the image data. The linearity constraint in particular suggests focusing on continuous acoustic features that correlate with articulatory movements or events. The feature design must avoid introducing large discontinuities (which could distort PCA) or overly categorical variables. Instead, smoothly-varying, quantitative features (frequency, amplitude, etc.) are favored. Lastly, because the PCA treats all inputs uniformly, any pre-processing like centering/scaling must be applied similarly to audio as done for the image pixels (the existing framework likely already centers each pixel intensity across time). Ensuring the audio features undergo analogous preprocessing is necessary for a cohesive analysis [12].

7. Verifying Audio Features Are Meaningful (Not Random Noise)

It is crucial to confirm that the integrated audio features truly correspond to the speech dynamics and aren't just capturing random fluctuations. Researchers employ several validation techniques: **shuffle tests, surrogate data, and cross-modal predictability checks**. A powerful approach is the **frame shuffle null test**. This involves randomizing the temporal alignment between audio feature frames and the corresponding video/MRI frames (or shuffling the audio feature sequence) and then re-running the PCA or reconstruction analysis. If the original feature is meaningful, the *true* time alignment should produce significantly better cross-modal reconstruction or correlation than any random misalignment. For example, *Stewart et al.* performed such a test in their audiovisual study: when they randomly permuted the pairing between facial and vocal-tract frames, the ability to reconstruct one from the other dropped dramatically, and was far worse than for the correctly ordered data [13]. All 1000 random shuffles yielded reconstructions with higher error, demonstrating that the actual synced features carry genuine predictive power beyond chance [13]. We can apply the same logic to audio: shuffle the audio feature time-order (breaking its coupling with the articulators) and see if the PCA-derived correlation between audio and image modalities collapses to baseline. **Lag testing** is another method: systematically introduce a time lag between the audio feature stream and the video/MRI, and measure any correlation or PCA mixing. A correctly aligned feature should show peak correlation or maximal shared variance at *zero* lag. If significant correlation appears only at a non-zero offset, that suggests the feature might be inadvertently misaligned or responding after the fact. Additionally, one can compare against **null feature controls**. For instance, generate a "feature" that is pure noise (or taken from a different speech segment entirely) and feed it into the PCA; it should not meaningfully couple with the articulatory PCs. If our chosen feature were similarly uncorrelated, it would behave like the null – e.g., get sequestered into its own PCA component with no impact on video/MR reconstruction. So we expect a real feature to outperform such null features on metrics like cross-modal covariance. Another validation is to use **frame-level permutation tests**: for each principal component that involves audio, evaluate whether the correspondence between audio and image loadings is statistically higher than would be expected by chance. This can be done by Monte Carlo shuffling or phase randomization of the audio time series to destroy temporal structure, then checking PCA outcomes. If our feature is meaningful, the true coupling (e.g., correlation of audio PC loading with an articulator trajectory) will lie outside the distribution of couplings from phase-scrambled audio. We also have **predictive reconstruction tests**: attempt to reconstruct the audio from the video/MRI using a regression or learned model (independent of PCA) and vice versa. Prior studies using multimodal data often do this – e.g., reconstructing tongue motion from face video, or acoustics from silent video [6]. If the audio feature is meaningful, one should be able to train a model (even a simple linear regression) on a portion of the data to predict it from the MRI/video features with some success (better than chance, but presumably with error since it's not fully redundant), and the same for predicting some articulatory measure from the audio. For instance, Jiang *et al.* (2000) showed significant correlations when estimating EMA articulator positions from acoustic LSP features and vice versa, far above chance levels [6]. If our feature were random, such cross-predictions would yield no better than chance results. Finally, visual inspection across time can help: overlay the audio feature curve with measured articulatory kinematic traces (e.g., lip aperture from video, or an ROI intensity from MRI) to see if they rise and fall together in a plausible way. One can even do a **framewise correspondence test**: sort frames by the audio feature value and check if the articulator configurations sort accordingly (e.g., do high-feature frames correspond to open-jaw images, etc., if feature is, say, energy). This kind of sanity check can catch features that might be measuring something irrelevant. In summary, by using **randomized alignment tests (shuffles), surrogate features, and cross-modal reconstruction evaluations**, we can be confident that our extracted audio features are not merely random noise. A

properly chosen feature will demonstrate significant, repeatable coupling with the visual/MR data that vanishes or greatly diminishes under permutation, as shown in prior multimodal analyses [13].

8. Visualizable Audio Features for Side-by-Side Display

Ideally, the audio representation can be visualized in a way that complements the MRI and video, facilitating intuitive interpretation. Several audio features lend themselves to **visual depiction alongside images**. A common choice is the **spectrogram**, which is essentially an image of the audio's time-frequency content. In fact, spectrograms have been used in tandem with rtMRI videos in the literature – for example, Toutios et al. display spectrograms of concurrently recorded audio beneath the MRI frame sequences to show acoustic timing and quality [14]. A spectrogram can be plotted as a grayscale or color heatmap (frequency on one axis, time aligned to the frames on the other), giving a visual "audio image" that can be directly synchronized with the video/MRI frames. This is useful to see phoneme boundaries or voicing (e.g. voiceless fricatives show high-frequency noise, vowels show distinct formant bands in the spectrogram). If only a 1D feature is used, one can plot it as a **curve over time** under or over the video frames. For instance, a **waveform** (amplitude vs. time) can be drawn, but since the waveform oscillates at high frequency, usually a smoothed envelope is more interpretable. The **intensity envelope** or loudness contour can be plotted as a single line that peaks where speech is loud (e.g., vowels) and dips in silence; this aligns well with visible mouth movements. Similarly, a **fundamental frequency (F0) track** can be plotted as a curve representing pitch over time. Researchers often overlay pitch contours on spectrograms or plot them below video – e.g., the pitch track could be superimposed on the spectrogram image as a dotted line showing how intonation rises and falls. Because pitch is a single value per frame, it's easy to visualize as a time series. **Formant trajectories** (F1, F2 over time) can also be plotted; typically, one would plot F1 and F2 as separate colored lines or two separate subplots aligned in time. These indicate vowel quality changes and can be directly compared with tongue/jaw motion in MRI. In fact, tools for rtMRI data allow automatic formant tracking and display, indicating their importance in visualization [2]. For example, one could annotate on an MRI frame sequence where F1 and F2 are at that moment, or simply show their timeline under the video. Another visual representation is a **mel-band filterbank energy plot**, which is similar to a coarse spectrogram (e.g., 10–20 bands) – this can be shown as a "bar graph" per frame or as a low-res spectrogram image. If the question is specifically about showing the feature "as an image," a spectrogram or any time-frequency representation naturally qualifies. Even a single-feature time series can be turned into an image by plotting it (time on x, feature value on y) – for instance, plotting the trajectory of the audio feature against time and placing it below the video like a signal strip. In multimodal studies, it's not uncommon to see **audio plotted in parallel with articulatory traces**. For example, *Silva et al.* (2019) aligned plots of MRI-derived articulator movement trajectories with the speech waveform and spectrogram to illustrate timing [15]. Also, many datasets (like USC's) provide ROI intensity traces and suggest visualizing them alongside spectrograms. If one uses an abstract feature like MFCC 1 or a principal component of audio, one could still visualize it by **resynthesizing an equivalent spectrogram** or by plotting its value as a function of time (though interpretation might be less direct). In summary, **spectrograms** and **time-series plots (for pitch, formants, energy)** are the most straightforward visual analogs of the audio features. They can be easily placed below or next to the MRI/video frames. Spectrograms in particular have been used in published rtMRI studies as a visual aid [14], since they mirror the time resolution of the images. Formant and pitch tracks, being highly interpretable, are also commonly visualized; for instance, an analysis GUI for rtMRI shows formant trajectories and even overlays them for analysis [2]. These visual representations help confirm that the audio feature is synchronized (you can literally see formant movements aligning with tongue movements, etc.) and allow qualitative assessment of how the audio relates to the articulatory events.

9. Audio Preprocessing: Beyond Denoising (Downsampling, Smoothing, Alignment)

Before extracting features, the audio must be preprocessed to ensure it aligns with the low frame rate and is free of artifacts that could confuse the PCA. **Noise removal** (e.g., MRI scanner noise reduction) is the first step, which we assume is already handled. Beyond that, a critical step is **downsampling and segmentation** of the audio to match the 16 fps frames. Typically, one would record audio at a high sampling rate (e.g. 16 kHz or 44.1 kHz). This needs to be mapped to frame-wise features. If using short-time analysis (like computing MFCCs or energy), one would choose a window length (perhaps ~40–60 ms) per frame and extract features for each window, effectively yielding a 16 Hz feature sequence. In some cases, audio is explicitly downsampled in time. For instance, in an MRI-video-audio dataset, the high-rate audio was **downsampled to 16 kHz and then dubbed onto the MRI video timeline** [16]. Even after such downsampling, one must ensure that each video frame has a defined audio snapshot (often by taking the center of the frame's time interval or averaging within the frame interval). **Temporal alignment** is absolutely crucial: any offset between audio and video/MRI must be corrected. As noted by Maekawa (2022) for a Japanese rtMRI database, even with careful procedures there can be "some uncertainty in the synchronization" between audio and frames due to differing sampling rates [12]. It's often necessary to calibrate alignment using a known cue – for example, the sound of the MRI gradient start (or a clap or beep) can serve as a sync marker [12]. After initial syncing by such a marker, one might further refine by aligning phonetic events with visible articulations (for instance, ensuring that the acoustic release burst of a /p/ coincides with the frame where lips part). Many setups use **forced alignment** with transcripts to get precise timing of phoneme boundaries, which can then be matched to video frames. Once aligned, one may need to consider **smoothing** or filtering the feature trajectory. Because we are effectively sampling audio features at 16 Hz, we should low-pass filter the feature stream to avoid aliasing high-frequency modulations. In articulatory studies, it's common to low-pass filter motion data at ~20–30 Hz to remove spurious rapid fluctuations [17] – by analogy, an audio feature like energy or formant could be filtered below ~8 Hz (since 16 Hz sampling gives ~8 Hz Nyquist) to ensure no rapid oscillations cause frame-to-frame jitter. For example, if using pitch, one might smooth the pitch contour to remove micro-fluctuations (like microprosody or halved/doubled pitch errors). If using MFCCs, typically delta-cepstrum (time derivatives) are optionally computed; in our case, we might omit deltas but could apply a light moving average on the static coefficients over a few frames to ensure continuity. **Downsampling** itself can be done by computing features on non-overlapping frame windows, which inherently does a form of smoothing (integration over the frame). However, if the frame stride is large, sometimes features benefit from overlap – e.g. computing MFCCs every 10 ms and then averaging or picking every Nth frame for 16 fps. Another preprocessing step is **band-limiting**: if focusing on certain frequency bands (say, formant ranges), one might apply a band-pass filter to the audio before feature extraction. For instance, if one wanted to track formants, a formant tracker might pre-emphasize the speech (a typical pre-processing in MFCC extraction is a high-frequency pre-emphasis filter to balance the spectrum). Pre-emphasis (like 6 dB/octave boost to highs) can be considered a preprocessing, as it can improve the prominence of formant peaks in the spectrum for MFCC calculation [1]. Additionally, if there are known linear trends or biases in the audio features (for example, if using a microphone with roll-off, etc.), one might normalize those out. **Normalization** of the audio feature over the utterance or dataset (beyond just noise removal) is also important: e.g., subtract the mean and divide by std of the feature across the dataset (so that it's zero-mean when entering PCA). This puts it on the same footing as the video/MRI which are likely intensity-normalized. In practice, one might perform **utterance-level normalization** for certain features. For example, pitch might be speaker-normalized by converting to semitone deviation from a speaker's average, which removes speaker-specific offsets. If multiple recordings are involved, one might equalize the overall gain so that one recording's audio isn't consistently louder (which would otherwise reflect in its

features). If the audio has **silence padding**, trimming leading/trailing silences to exactly match the video duration is also a preprocessing step. Silence frames can still be included (as they correspond to closed-mouth or rest positions), but one should ensure no misalignment (e.g., an extra silent audio frame at the end with no corresponding video frame). To summarize, beyond denoising, **the audio should be resampled or analyzed to yield features at 16 Hz, carefully synchronized to the imaging frames** [12]. A gentle **low-pass filter or smoothing** of the feature trajectory is advisable to prevent aliasing and erratic jumps (since articulatory changes are relatively slow, most relevant audio feature content is below ~8 Hz in modulation frequency). **Normalization** (z-scoring, etc.) is done to match the scale of other modalities. By executing these steps – align, downsample, smooth, normalize – we prepare the audio features so that they cleanly integrate into the PCA with video and MRI. These practices are reflected in multimodal speech studies; for example, in one corpus construction, audio was downsampled and time-synchronized to MRI to within a small error [16] [12], and articulatory signals are often filtered to match sampling rates [17]. Following such procedures ensures the audio features are not only noise-free but also appropriately conditioned for joint analysis.

Sources: Recent studies and reviews in speech science and multimodal learning were used to support these answers, including works on audio-visual speech correlation [6] [3], multimodal PCA of speech articulators [11] [13], audio feature extraction techniques [1] [8], and best practices from real-time MRI speech corpora [14] [12]. All cited sources are from the last ~20 years except where foundational work is referenced for context. Each point is grounded in findings or methodologies reported in these technical sources.

---

[1] Trends in audio signal feature extraction methods
https://calebrascon.info/PDA/Topic4/addresources/features.pdf

[2] [10] [14] Advances in real-time magnetic resonance imaging of the vocal tract for speech science and technology research - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC5100697/

[3] Development of speech rhythm in first language: The role of syllable ...
https://pubs.aip.org/asa/jasa/article/143/6/EL463/917955/Development-of-speech-rhythm-in-first-language-The

[4] [5] A long-form single-speaker real-time MRI speech dataset and benchmark
https://arxiv.org/html/2509.14479v1

[6] seas.ucla.edu
http://www.seas.ucla.edu/spapl/paper/jiang_icslp00.pdf

[7] [11] [13] The interrelationship between the face and vocal tract configuration during audiovisual speech - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC7768679/

[8] [9] A Multi-Frame PCA-Based Stereo Audio Coding Method
https://www.mdpi.com/2076-3417/8/6/967

[12] [16]
https://www.jstage.jst.go.jp/article/ast/46/1/46_e24.22/_pdf

[15] A fast and flexible MRI system for the study of dynamic vocal tract ...
https://pmc.ncbi.nlm.nih.gov/articles/PMC4947574/

[17] A generalized smoothness criterion for acoustic-to-articulatory ...
https://www.researchgate.net/publication/47531215_A_generalized_smoothness_criterion_for_acoustic-to-articulatory_inversion