NOISE-ROBUST DYNAMIC TIME WARPING USING PLCA FEATURES

Brian King[*†‡] *Paris Smaragdis*[‡○] *Gautham J. Mysore*[‡]

[†] Department of Electrical Engineering, University of Washington
[‡]Advanced Technology Labs, Adobe Systems Inc.
[○]University of Illinois at Urbana-Champaign

ABSTRACT

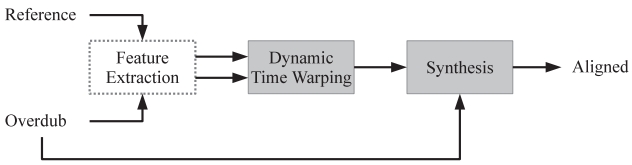
Conventional speech features, such as mel-frequency cepstral coefficients, tend to perform well in template matching systems, such as dynamic time warping, in low noise conditions. However, they tend to degrade in noisy environments. We propose a method of calculating features using the probabilistic latent component analysis (PLCA) framework. This framework models the speech and noise separately, leading to higher performance in noisy conditions than conventional methods. In this work, we compare our PLCA-based features with conventional features on the task of aligning a high-fidelity speech recording to a noisy speech recording, a scenario common in automatic dialogue replacement.

Index Terms— Probabilistic Latent Component Analysis, Dynamic Time Warping, Automatic Dialogue Replacement

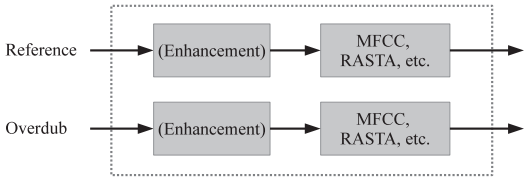
1. INTRODUCTION

Automatic time alignment of audio has many interesting applications, including synchronizing high-quality speech to a low-quality reference recording of the same utterance, aligning dialogue of different languages to aid in foreign overdubbing, and synchronizing recorded instrument tracks. In this paper, we refer to the unaligned recording as the overdub and the recording with the desired timing as the reference. Rather than simply shifting an audio clip by a global offset or sampling factor, automatic time alignment systems stretch and compress the signal dynamically within a clip. These systems typically consist of three steps [1] (see Figure 1(a)), which are calculating features for both the reference and the overdub signals, finding the optimal alignment mapping between the two signals using dynamic time warping (DTW) [2], and synthesizing a warped version of the overdub signal so that it temporally matches the reference signal [3]. In this paper, we concentrate on the first step. Specifically, we propose a method of feature calculation that exhibits better performance than conventional features when used with noisy recordings. This new method is based on probabilistic latent component analysis (PLCA) [4], a method that is commonly used in source separation [5] and can also be used for denoising. Instead of simply denoising the noisy reference signal and then calculating conventional features, we propose directly using certain estimated PLCA model parameters as features. We show that this new method provides significant improvement in time alignment tasks, particularly in the case in which we have noisy recordings.

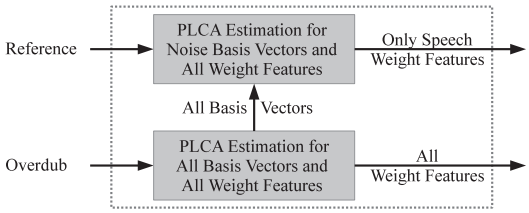
For our experiments, we focus on the application of synchronizing a high-quality speech overdub with a lower-quality reference recording. This scenario is commonly encountered in moviemaking, where the original video recordings use distant microphone setups and are often in a noisy environment, which result in poor audio quality. It is common procedure to record the actors in a proper sound studio environment afterwards, where they try to match their speech from the original video recordings. That higher quality audio is the one that gets used in the final movie mix. The process of re-recording actors in the studio is known as automatic dialogue replacement[1] (ADR). If an automatic alignment system is not used, then the actors must re-record their lines until the timing is perfect. Some manual alignment may be possible by a studio engineer, but can be even more time-consuming and difficult than recording perfectly-timed lines. The appeal of utilizing an automatic alignment system is that it has the potential of significantly speeding up



(a) Block Diagram of three-stage automatic aligment system



(b) Block diagram of conventional feature extraction method



(c) Block diagram of proposed PLCA feature extraction method

Fig. 1. Block diagrams of automatic alignment system, conventional feature extraction, and proposed PLCA feature extraction

*This work was performed while at Adobe Systems Inc.

[1]The "automatic" in ADR refers to the process of automatically looping a clip until the actor gets the overdub version perfect. However, the actual alignment is not automatic.

the ADR process, since the actor will need to record just one good performance in the studio, and the alignment system can automatically take care of the timing.

2. BACKGROUND

In the first step of the automatic alignment process, features are calculated on a frame-by-frame basis for both the reference and overdub recordings. In a typical automatic alignment system, the subsequent step is where the actual time alignment calculations are performed. The DTW algorithm does not use the reference and overdub features directly, but acts on the similarity matrix of the features. The similarity matrix contains the cosine distance of the reference and overdub features at each time window:

$$S(F_r(t_a), F_o(t_b)) = \cos(\theta) = \frac{F_r(t_a) \cdot F_o(t_b)}{\|F_r(t_a)\|\|F_o(t_b)\|} \quad (1)$$
$$\text{where } S \in [-1, 1]^{T_r, T_o}$$

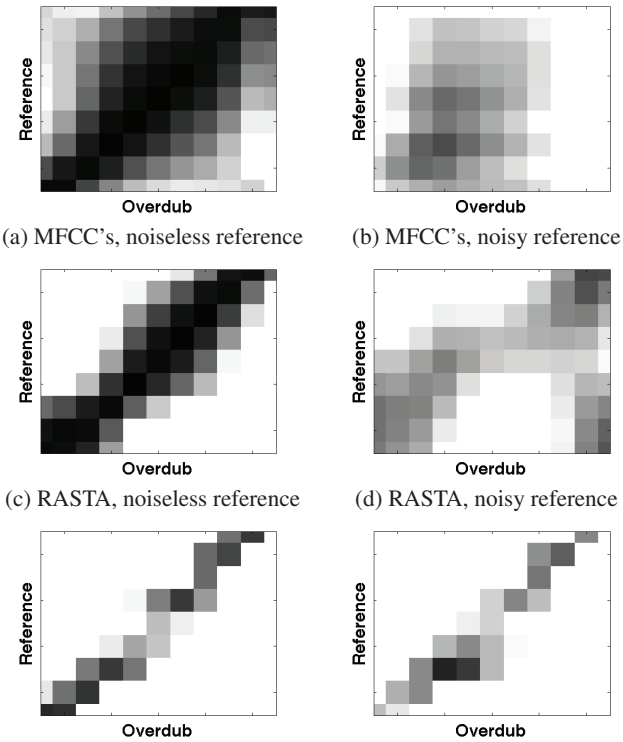
$F_r(t_a)$ is the feature vector of the reference signal at time frame t_a and $F_o(t_b)$ is the feature vector of the overdub signal at time frame t_b. T_r and T_o are the number of time frames of the reference and overdub signals. The dynamic time warping algorithm then uses this matrix to calculate a set of frame-to-frame correspondences that maximize the overall similarity between the two sequences [2]. Now that we have talked about the role of feature calculation and its importance in alignment, we will look at some conventional feature calculation methods and their performance in noise.

Conventional features for automatic time alignment tasks are the same features conventionally used for automatic speech recognition tasks. One such set of features is mel-frequency cepstral coefficients (MFCCs) [6, 7]. Similar to speech recognition systems, MFCCs can work very well in ideal recording conditions, but degrade in noise [8]. This is illustrated in Figure 2 which shows the same section of the similarity matrix for (a) a low-noise reference and overdub signal and (b) noisier reference (5 dB SNR) and overdub signal. The reference is on the y-axis, the overdub is on the x-axis, and the beginning of both the reference and overdub is on the bottom left corner. Black indicates a high similarity value and white indicates a low similarity value. Ideally, the path would be clearly indicated by having a much higher value than neighboring frames. In Figure 2(a), a high-similarity path is clearly seen traveling diagonally from the bottom left corner to the top right corner. A clear, accurate path like this will result in an accurate alignment. However, when the reference becomes noisier (see Figure 2(b)), the MFCC's begin to distort and the previously clear patch through the similarity matrix is obscured. Another popular set of features designed to be more robust to noise is the relative spectral transform-perceptual linear prediction [9] (RASTA-PLP). In Figure 2(c), the RASTA-PLP's performance with a non-noisy reference is accurate and on par with MFCC's, but also performs unsatisfactorily in the noisy 5 dB SNR case (see Figure 2(d)). In the next sections, we will introduce our PLCA-based method and compare its performance in noise with the conventional methods.

3. PROBABILISTIC LATENT COMPONENT ANALYSIS

3.1. Model

Before discussing the specifics of our feature calculation algorithm, we introduce probabilistic latent component analysis (PLCA) [4]. This model can be used to model spectrograms and is a member of



(a) MFCC's, noiseless reference (b) MFCC's, noisy reference

(c) RASTA, noiseless reference (d) RASTA, noisy reference

(e) PLCA, noiseless reference (f) PLCA, noisy reference

Fig. 2. Similarity matrix of reference (y-axis) and overdub (x-axis) for MFCC's, RASTA-PLP, and PLCA features

a family of non-negative matrix factorization models [10]. PLCA models a spectrogram as a probability distribution as follows:

$$X_N(f,t) \approx P(f,t) = \sum_{z \in \{1,2,\ldots,K\}} P(z)P(f|z)P(t|z) \quad (2)$$

$$\text{where} \qquad X_N(f,t) = \frac{|X(f,t)|}{\sum_{f,t}|X(f,t)|} \quad (3)$$

$X(f,t)$ is the short-time Fourier transform (STFT) of a signal, and $X_N(f,t)$ is the normalized STFT. $P(f|z)$ corresponds to the spectral building blocks, or spectral basis vectors. $P(t|z)$ corresponds to the temporal evolution of these spectral basis vectors. $P(z)$ corresponds to the relative contribution of each spectral basis vector. All distributions are discrete. Given a spectrogram, the model parameters are estimated using the EM algorithm.

3.2. PLCA-Based Feature Calculation

Before presenting our proposed feature calculation method, we first recall some key characteristics of the data that we will be able to take advantage of with our approach. In an automatic dialogue replacement (ADR) scenario, the goal is to time align a high-fidelity overdub recording to a potentially noisy, lower quality reference recording from the same speaker. Our proposed method takes advantage of the fact that both clips are from the same speaker. In essence, our method analyzes the overdub clip to find the speech characteristics, which we then use to analyze the noisy clip more accurately than if we had no prior knowledge of the speech. In contrast, typical feature

calculation methods do not take advantage of this knowledge, extracting the features of the two clips independently (see Figure 1(b)).

The first step of our method is to calculate the magnitude subband representation of each signal. We use a mel-spaced subband representation because we found that it increased alignment performance. When using a linearly-spaced narrowband spectrogram, if the pitch of an utterance from the reference clip differs slightly from the pitch in the studio clip, then they will not be represented well by the same basis vectors. A mel-spaced subband representation helps mitigate this so that small differences in pitch do not result in significant changes in PLCA model parameters. Although we chose mel-spaced subbands, logarithmically-spaced subbands could have been used to achieve similar results.

Next, we perform PLCA on the spectrogram of the overdub studio version. We construct F_o, the weight features of the overdub signal at each time frame t, as

$$F_o(z,t) = P_o(t,z) = P_o(t|z)P_o(z), \text{for } z \in Z_o \quad (4)$$

where Z_o is the set of N_{speech} speech basis vectors learned in the overdub signal. Intuitively, F_o can be thought of as a set of weights that indicate how the speech basis vectors can be linearly combined to approximate the observed signal. We reuse the basis vectors learned from the overdub signal to be the speech basis vectors for the reference recording. If we knew that the reference signal also had little to no noise, then we would simply perform PLCA on the reference to find a new $P_r(z)$ and $P_r(t|z)$, keeping the $P_o(f|z)$ basis vectors learned in the previous step fixed. We would then calculate the reference recording features $P_r(z,t)$, as before. Since we are using the same set of basis vectors for both the overdub and reference clips, parts of audio similar in the reference will have values of $P_r(t|z)$ similar to the values of $P_o(t|z)$. In other words, the features learned from the two signals correspond to the same sounds because they share the same speech basis. As long as the vocal characteristics between the two dialogue recordings are similar, the features of the corresponding parts of dialogue will match closely and result in accurate alignment.

The preceding method will work well when the reference has little noise, but will degrade with noise. Since the noise is not learned and separated from the speech, the weight features will try to account for the noise. With the PLCA framework, we are able to get around this problem. When the reference recording is noisy, we still use the speech basis vectors learned from the clean overdub signal, but we also estimate an additional N_{noise} basis vectors while running PLCA on the reference to account for the noise. Since PLCA models a spectrogram as a linear combination of basis vectors, introducing noise affects the speech weight features much less than conventional methods because the noise in the reference signal can be learned and modeled explicitly by the algorithm. If we were to then synthesize an enhanced speech signal with the speech basis and weight features, we would be performing semi-supervised source separation [5], but this is unnecessary for our task. The features for the reference signal are then calculated by

$$F_r(z,t) = P_r(t,z) = P_r(t|z)P_r(z), \text{for } z \in Z_o \quad (5)$$

where Z_o is the same set of speech basis vectors learned in the overdub signal. By modeling the noise separately from the speech, we are able to calculate features that are more robust to noise. This can be seen in Figure 2(e-f), which show same section of the similarity matrix values for PLCA-based features. Notice how the similarity matrix changes significantly less when noise is introduced than when the MFCC and RASTA-PLP are used, having a clearly-defined optimal path in both cases.

In summary, the steps of the algorithm ((see Figure 1(c)) are as follows:

1. Calculate the subband representations of both the reference and overdub signals.
2. Perform PLCA on the low-noise, overdub signal to find N_{speech} speech basis vectors.
3. Perform PLCA on the noisy reference signal, using the speech basis vectors learned in the previous step. Include an additional N_{noise} basis vectors with random initial values. Keep the speech basis vectors constant, but allow the noise basis vectors to be updated to adapt to the noise.
4. The features for the overdub and reference recordings are their speech weight features, F_o and F_r, as defined in (4) and (5).

4. EXPERIMENTAL DESIGN AND RESULTS

We will now discuss our pilot study, where we compare our proposed PLCA method with dynamic time warping using several conventional features in an automatic time alignment task of a low-noise overdub recording to a noisy reference recording. Two recordings of the same four sentences were made by three male speakers and one female speaker. The speakers were prompted to speak naturally in both recordings, but no feedback was given about the timing of the recordings. Thus, the timings, pitch contours, and pronunciation were noticeably different between the two recordings, but not extremely different. To create noisy reference recordings, factory noise from the NOIZEUS corpus [11] was added at -10, -5, -, +5, and +10 dB SNR. The SNR was calculated by the average power of active speech vs. the average power of the noise. All clips were sampled at 16 kHz. The conventional features we compared to the proposed method were MFCC and RASTA-PLP [9] features. We also compared MFCC and RASTA-PLP feature sets calculated from a denoised version of the noisy signal. For the denoising, we used Ephraim and Malah's classical enhancement algorithm [12] as implemented by Loizou [13]. The same dynamic time warping algorithm was used for all sets of features. The alignments calculated by the DTW algorithm were compared to "ground truth" alignment. The ground truth was found by first performing alignment on the reference and overdub clips with no noise added to the reference. Conventional MFCC's were used to calculate the ground truth. Time-aligned synthesized signals were then compared with the reference signals and manually verified that the performance of the ground truth alignment was satisfactory, both by listening to the two signals together as well as comparing their spectrograms.

For all feature sets, the same window length (32 ms) and hop size (50%) were used. The MFCC features were calculated using the first 8 discrete cosine transform (DCT) coefficients from the 29 mel-spaced subband representation [7]. The RASTA-PLP model used an 8th-order PLP model and was calculated with the RASTA code from [14]. Both the MFCC and RASTA-PLP features worked well with noise-free references. Two PLCA models were used in the comparison. The first used 40 basis vectors for speech and 40 for noise. The second PLCA model just used 40 basis vectors for speech, so the noise was not modeled separately. This was done to see if our hypothesis that modeling the noise separately from the speech would increase alignment performance. For the PLCA-based methods, we combined the STFT subbands into 102 mel-spaced subbands.

In comparing DTW frame mapping of the aligned versions to the ground truth, a frame was labeled as correct if it was within 2

frames of the ground truth mapping. Since we used 32 ms windows with a 50% overlap in our experiments, this corresponds to being accurate within 32 ms. This value was chosen because it translates to being accurate within one video frame for the common 30 frames per second rate. We have plotted the alignment performance of the features in figure 3. PLCA with separate speech and noise basis vectors performed the best, followed by PLCA with speech basis vectors only, Ephraim-Malah (E-M) with RASTA-PLP, RASTA-PLP, Ephraim-Malah with MFCC's, and finally MFCC's. With this, we have seen that processing the noisy signals with speech enhancement techniques such as Ephraim-Malah can improve performance over unprocessed signals, and that the RASTA-PLP, which was designed to be more robust to noise, performs better in this task than MFCC's. In order to see whether PLCA may be performing so well just because it had the highest feature dimensionality, we tried keeping all 29 DCT coefficients for the MFCC's. This improved their alignment performance, but still performed worse than all of the other RASTA and PLCA sets of features. However, no combination of conventional features and signal processing techniques that we tried was able to match the performance of PLCA features at any of the tested SNR's, especially in lower SNR's. This is consistent with our prediction that all methods would perform fairly well in minimal noise situations, but that the conventional methods would decrease in performance significantly more than the proposed method in the presence of noise.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a new PLCA-based feature calculation method used in template matching and DTW applications such as automatic alignment for automatic dialogue replacement. Our alignment experiments showed a marked improvement over conventional features in noisy environments, and put forth an example of how data-driven machine-learning approaches can improve feature selection. Some manual work was done in finding optimal parameters for the PLCA-based approach, but in future work, it would be helpful to conduct a thorough investigation on choosing optimal parameters. All of the experiments used the same set of parameters, but it is likely that optimal parameters will vary depending on speech characteristics, noise type, and SNR. In conclusion, we have demonstrated that PLCA-based features for automatic time alignment are useful and that preliminary results show high potential to aid in alignment, especially when one of the clips is noisy.

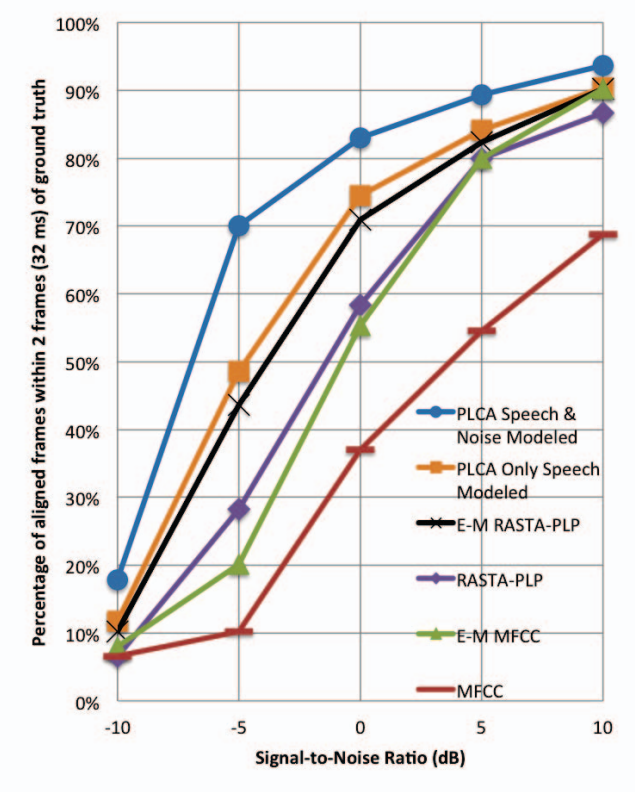


Fig. 3. Pecentage of aligned frames within 2 frames (32 ms $\approx$ 1 video frame) of correct alignment

6. REFERENCES

[1] J.P. Hosom, *Automatic time alignment of phonemes using acoustic-phonetic information*, Ph.D. thesis, Oregon Graduate Institute of Science and Technology, Oct. 2000.

[2] L. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1st edition, Apr. 1993.

[3] B. Ninness and S.J. Henriksen, "Time-scale modification of speech signals," *IEEE Transactions on Signal Processing*, vol. 56, no. 4, pp. 1479–1488, Apr. 2008.

[4] P. Smaragdis, B. Raj, and M. Shashanka, "A probabilistic latent variable model for acoustic modeling," in *NIPS Workshop on Advances in Modeling for Acoustic Processing*, 2006.

[5] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," *Independent Component Analysis and Signal Separation*, pp. 414–421, 2007.

[6] S. Greenberg and W. Ainsworth, *Listening To Speech: An Auditory Perspective.*, Lawrence Erlbaum Associates, Mar. 2006.

[7] D.P.W. Ellis, "Dynamic time warp (DTW) in matlab," 2003, online web resource.

[8] R. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1–15, July 1997.

[9] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP speech analysis," *ICSI Technology Report TR-91-069*, 1991.

[10] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999, PMID: 10548103.

[11] Yi Hu and P.C. Loizou, "Subjective comparison of speech enhancement algorithms," in *ICASSP*. May 2006, vol. 1, IEEE.

[12] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[13] P.C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, 1 edition, June 2007.

[14] D.P.W. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005, online web resource.