A study on constraining Connectionist Temporal Classification for temporal audio alignment

Yann Teytaut, Baptiste Bouvier, Axel Roebel

A study on constraining Connectionist Temporal Classification for temporal audio alignment.

Yann Teytaut, Baptiste Bouvier, Axel Roebel

IRCAM, Sorbonne University, CNRS, UMR 9912 STMS, F-75004 Paris, France

yann.teytaut@ircam.fr, baptiste.bouvier@ircam.fr, axel.roebel@ircam.fr

Abstract

Connectionist Temporal Classification (CTC) has become a standard for deep learning-based temporal alignment allowing relevant probabilistic distributions to be learned. However, by nature, CTC is a transcription objective that can be minimized without guaranteeing any alignment properties. This work aims to study several constraints to help CTC generating alignments. With a fully convolutional architecture coupled with multi-head attention, we investigate the task of phonetic alignment for clean speech and singing signals. The focus is set on the impact of additional losses, namely spectral envelope reconstruction, temporal structure invariance and guided monotony. Results show that, once scaled to have identical temporal dependence, combining all of these constraints produces best performances.
Index Terms: Connectionist Temporal Classification, phonetic alignment, multi-objective training, loss scaling, attention.

1. Introduction

Connectionist Temporal Classification (CTC) is an objective function used to train deep neural networks. It has originally been developed for labeling and segmenting sequences [1] and has gained tremendous popularity in the past few years as seen in many sequence-to-sequence (seq2seq) problems, especially towards end-to-end (E2E) architectures [2, 3, 4, 5].

The main idea of CTC algorithm is to output probabilistic distributions from which sequences are estimated. Its founding principle is a one-to-many prediction framework based on the existence of a blank label, usually denoted ε, that allows several acceptable sequences from a given input [6].

The proposed framework is particularly adapted for audio data that are sequential in time. As a result, CTC has been applied with success to various audio-oriented tasks including speech recognition [7], note transcription [8], singing language identification [9], and detection of sung explicit content [10].

In this work, we aim to study in details the behaviour of the CTC loss for audio alignment tasks [11, 12], *i.e.*, automatic synchronization of audio representations (*e.g.*, voice recordings, music performances) with information often of symbolic nature (*e.g.*, music scores, text transcripts).

CTC has launched a new trend in this literature and has the great benefit of not requiring aligned data for training models [13, 14]. Yet, alignment remains intrinsically difficult to couple with CTC precisely due to this one-to-many mapping it exploits. Indeed, CTC measures by nature a transcription cost, therefore it can be minimized without guaranteeing alignment properties.

Some approaches have tried regularization on the CTC loss to better capture the role of its blank label [15], prevent peaky probability distribution [16], or improve its scalability with Cross-Entropy through sampling [17]. However, to the best of the authors' knowledge, very few works were dedicated to ensure the emergence of alignment from CTC probabilities [18].

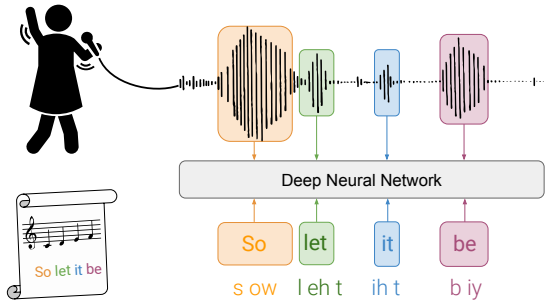

Figure 1: *An illustration of audio-to-text alignment with deep neural network for singing or speech. The network, given audio and lyrics, creates a mapping between them. Icons from [21].*

This paper proposes to evaluate how *constrained* CTC, in opposition to *basic* CTC, performs for the alignment between audio and text (see **Fig. 1**). We set our sights on phoneme-to-audio alignment [19, 20], which is particularly challenging due to the high temporal precision needed. We focus on the impact of additional losses, namely envelope reconstruction, temporal structure invariance and guided monotony, that we introduce.

The contributions of this work are:

- A fully convolutional network for CTC-based voice alignment, without phone transcripts as inputs and free from recurrent layers (contrary to recent works [22, 18]);
- A procedure to consistently combine duration-dependent loss functions into multi-objective losses;
- An overview on how to guarantee alignment from CTC posteriogram with definition of additionnal constraints.

Section 2 introduces the baseline neural architecture used for benchmarking in this study. The additive losses and their scaling are then presented in section 3. Next, section 4 exposes alignment results on clean speech and singing datasets. Finally, in section 5, we conclude on the influence of constraining CTC for temporal alignment with multi-objective training.

2. Neural architecture

This section presents the neural architecture used in this paper. In opposition to recent deep learning approaches for alignment [18, 22], the current proposal has two main advantages: (1) it does not need non-aligned text transcripts as model's inputs, but only at inference time (forced alignment) ; and (2) it is fully convolutional to ensure that information is processed while respecting temporality. The absence of recurrent layers also leads to more stable and easier trainings. Moreover, a multi-head self-attention mechanism [23] is introduced following previous works inciting to link CTC with attention [24, 25, 26]. The complete baseline architecture is depicted on **Fig. 2**.

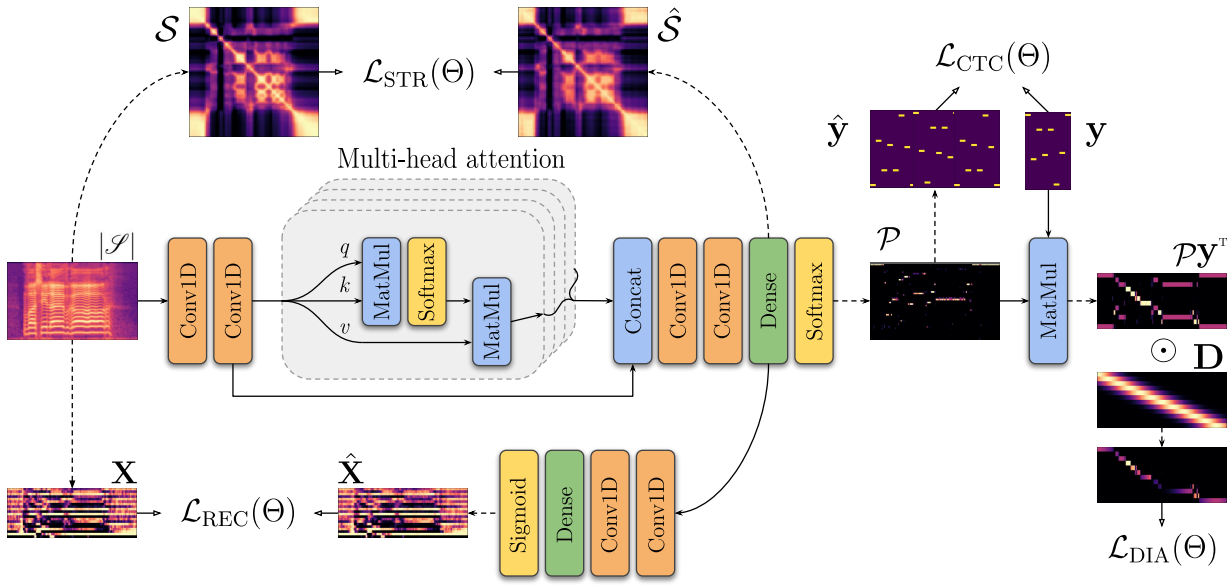

Figure 2: *Baseline network architecture. Spectrograms are processed with convolutional layers and (self-)multi-head attention towards the generation of robust CTC posteriograms allowing envelope (MFCCs) reconstruction, structural invariance, and sequence monotony.*

The network takes as inputs log-scaled mel-spectrograms $|\mathscr{S}| \in \mathbb{R}^{T\times B}$ from which normalized MFCCs $\mathbf{X} \in [0,1]^{T\times F}$ are derived. Through several 1D-convolution blocks and a self H-head spectral attention mechanism, the system converts the mel-spectrograms $|\mathscr{S}|$ into a posteriogram $\mathcal{P} \in [0,1]^{T\times(L+1)}$ over an alphabet $\mathcal{A}$ composed of L labels and the CTC blank ε. From $\mathcal{P}$, a sequence can be predicted $\hat{\mathbf{y}} \in (\mathcal{A}\cup\{\varepsilon\})^{T\times 1}$ along with an estimation of MFCCs features $\hat{\mathbf{X}} \in [0,1]^{T\times F}$.

3. A study on constraining CTC

This sections gives background on the loss functions that we manipulate and their combination. Since all losses are duration-dependent, we propose to ensure that they all scale similarly, linearly with the time length T. Their associated scaling will be established from worst-case scenario studies. The model's learnable parameters are denoted Θ.

3.1. Connectionist Temporal Classification

The CTC loss ensures that the sequence $\hat{\mathbf{y}}$ decoded from the posteriogram $\mathcal{P}$ is close to the groundtruth text $\mathbf{y} \in \mathcal{A}^{M\times 1}$, once that repeated labels are merged and blank labels removed with an operator $\mathcal{B}$. It is defined as a negative log-likelihood computed from all acceptable sequences $\hat{\mathbf{y}}$ and time frames t,

$$\mathcal{L}_{\text{CTC}}(\Theta) = -\log\left(\sum_{\substack{\ell\in\hat{\mathbf{y}}\\ \hat{\mathbf{y}}\in\mathcal{B}^{-1}(\mathbf{y})}}\prod_{t=0}^{T-1}\mathcal{P}[t,\ell]\right). \quad (1)$$

Worst-case scenario. The number of alignments in CTC computation, *i.e.*, the cardinal of $\mathcal{B}^{-1}(\mathbf{y})$, is $\binom{T+M}{T-M}$ [6, 27]. For a uniform posteriogram, we thus estimate the loss value to

$$\mathcal{L}_{\text{CTC}}(\Theta) \sim -\log\left[\binom{T+M}{T-M}\left(\frac{1}{L+1}\right)^{T}\right]. \quad (2)$$

To go even further, we assume[1] that there are much more time frames than sequence length, *i.e.*, $T \gg M$, leading to

$$\mathcal{L}_{\text{CTC}}(\Theta) \sim \log(L+1)T. \quad (3)$$

3.2. Envelope reconstruction

From the final dense CTC layer, we generate an estimate of the spectral envelope (MFCCs) $\hat{\mathbf{X}}$. It must be as close as possible to the original features $\mathbf{X}$ to reinforce temporal coherence in the CTC predictions as in [18]. We therefore minimize the L1 loss

$$\mathcal{L}_{\text{REC}}(\Theta) = \left|\left|\mathbf{X}-\hat{\mathbf{X}}\right|\right|_1 = \sum_{t=0}^{T-1}\sum_{f=0}^{F-1}\left|\mathbf{X}[t,f]-\hat{\mathbf{X}}[t,f]\right|. \quad (4)$$

Worst-case scenario. Given that $\mathbf{X},\hat{\mathbf{X}} \in [0,1]^{T\times F}$, the maximum difference one can observe is

$$\mathcal{L}_{\text{REC}}(\Theta) \sim FT. \quad (5)$$

Note that a key difference with [18] is that the *envelope* $\mathbf{X}$ is reconstructed instead of the spectrogram $|\mathscr{S}|$. Indeed, $|\mathscr{S}|$ estimation implies that F0 values must propagate through the network, but F0 detection is not relevant for the alignment task.

3.3. Temporal structure invariance

In the same set of mind, we want the CTC predictions to have the same temporal structure as the original spectrum. To do so, we compute the cosine self-similarity matrices (SSM) [28] of $|\mathscr{S}|$ and the final CTC dense layer, denoted $\mathcal{S}$ and $\hat{\mathcal{S}} \in [0,1]^{\frac{T}{2}\times\frac{T}{2}}$, respectively.

The SSM are $(\frac{T}{2}\times\frac{T}{2})$-shaped because we use an (4×4)-average pooling operation with stride (2×2) to smooth local structural singularities and reduce memory storage as well.

[1] Note that the CTC algorithm necessarily needs that $T > M$. It is also worth mentioning that T and M may vary for each audio-text pair.

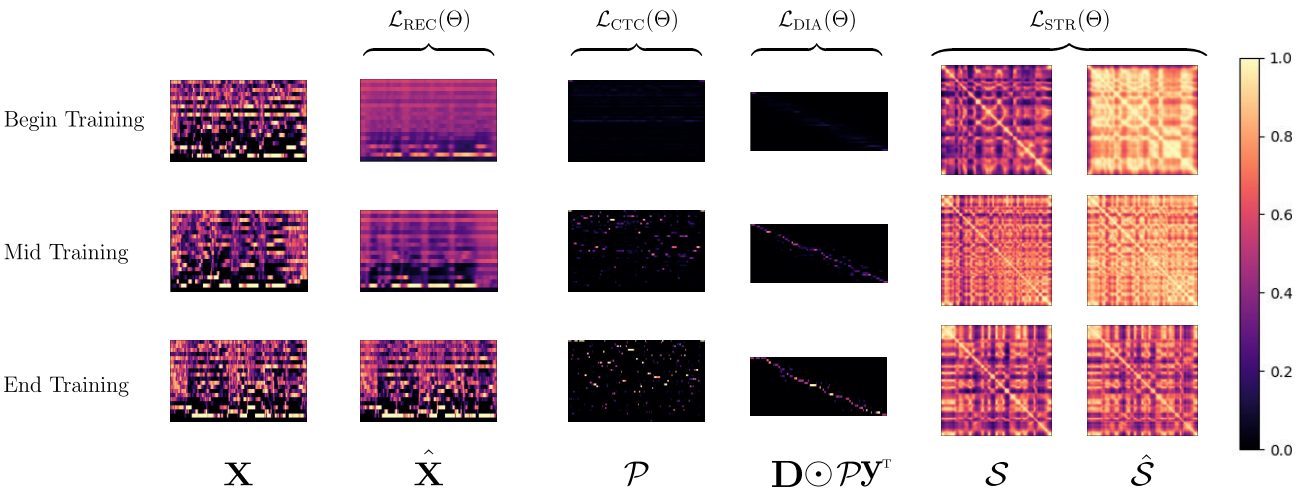

Figure 3: *Impact of the several losses constraining CTC throughout training. Columns, from left to right, depict original envelope (MFCCs), envelope reconstruction, CTC posteriogram, text-to-audio monotony and input/output structures as self-similarity matrices.*

We then minimize the L1 between $\mathcal{S}$ and $\hat{\mathcal{S}}$, that is

$$\mathcal{L}_{\text{STR}}(\Theta) = \left|\left|\mathcal{S} - \hat{\mathcal{S}}\right|\right|_1 = \sum_{t=0}^{\frac{T}{2}-1}\sum_{t=0}^{\frac{T}{2}-1}\left|\mathcal{S}[t,t] - \hat{\mathcal{S}}[t,t]\right|. \quad (6)$$

Worst-case scenario. Given that $\mathcal{S}, \hat{\mathcal{S}} \in [0,1]^{\frac{T}{2}\times\frac{T}{2}}$, the maximum difference one can observe is

$$\mathcal{L}_{\text{STR}}(\Theta) \sim \tfrac{1}{4}T^2. \quad (7)$$

3.4. Guided monotony

Speech/singing signals and phonetic transcripts are monotonic. Aligning audio with such sequences implies uncovering a pseudo-diagonal matrix showing that labels are pronounced with the flow of time. Let $\mathbf{D} \in [0,1]^{T\times M}$ be a Gaussian-decreasing matrix, with $\sigma = 0.1$, defined $\forall t, \forall m$ by the rule

$$\mathbf{D}[t,m] = \exp\left(-\left(\frac{t}{T} - \frac{m}{M}\right)^2 / 2\sigma^2\right). \quad (8)$$

One can notice that the element-wise product between $\mathbf{D}$ and the multiplication of posteriogram $\mathcal{P}$ (without blank) and one-hot target sequence $\mathbf{y} \in \{0,1\}^{M\times L}$ precisely yields the alignment matrix which is expected to be monotonic (*e.g.*, **Fig. 2** and **Fig. 3**). This can happen only if CTC systematically highlights the full duration of each label. Hence, we define and impose the guided monotony (inspired by *guided* attention [29]) constraint:

$$\mathcal{L}_{\text{DIA}}(\Theta) = \left|\left|\mathbf{D} \odot \text{softmax}\left(\mathcal{P}\mathbf{y}^{\text{T}}\right) - \mathbf{D}\right|\right|_1. \quad (9)$$

Worst-case scenario. This loss is maximized when CTC does not recognize any target labels, that is $\mathcal{L}_{\text{DIA}}(\Theta) = ||\mathbf{D}||_1$. The pseudo-diagonal structure of the matrix implies that its total L1 norm is equal to $\frac{3}{4}$ of the sum of Gaussian integrals on each of the M lines. It comes:

$$\mathcal{L}_{\text{DIA}}(\Theta) \sim \tfrac{3}{4}\sqrt{2\pi}\sigma MT \sim 2\sigma MT. \quad (10)$$

Note that this constraint is exclusively exploited during training, so that phoneme sequences are not inputs of the model at inference time, in opposition to both systems [18, 22].

3.5. Losses scaling

The present study aims to couple various losses, raising the question of how to combine them and their respective trade-off, which is a concern for all multi-task learning problems [25, 30].

It has been shown that chosen objective criteria do not result in similar variations when audio length changes. This is a major issue since different elements in the training set will not end up inducing comparable updates for gradients and weights.

Consequently, from previous worst-case scenario estimates, we propose to scale each loss so that they have identical, linear dependency on T, which is reasonable because time segments in sequences will always have the same impact independently of the phrase they are found in. The scaled losses are defined as

$$\mathcal{L}^n_{\text{CTC}}(\Theta) \leftarrow \tfrac{1}{\log(L+1)}\mathcal{L}_{\text{CTC}}(\Theta) \quad \mathcal{L}^n_{\text{REC}}(\Theta) \leftarrow \tfrac{1}{F}\mathcal{L}_{\text{REC}}(\Theta)$$

$$\mathcal{L}^n_{\text{STR}}(\Theta) \leftarrow \tfrac{4}{T}\mathcal{L}_{\text{STR}}(\Theta) \quad \mathcal{L}^n_{\text{DIA}}(\Theta) \leftarrow \tfrac{1}{2\sigma M}\mathcal{L}_{\text{DIA}}(\Theta)$$

The goal now is to quantify how alignment performances evolve when minimizing the global loss

$$\mathcal{L}(\Theta) = \mathcal{L}^n_{\text{CTC}}(\Theta) + \tfrac{1}{3}\sum_i \delta_i \mathcal{L}^n_i(\Theta) \quad (11)$$

with i an index over all above-mentioned constraints and δ_i the Kronecker delta. This results in 8 configurations to be tested. Their effect during training is shown on **Fig. 3**. The factor $\frac{1}{3}$ ensures that even all joint constraints do not dominate the CTC.

4. Evaluations

4.1. Datasets

To evaluate the impact of constraining CTC for phoneme-to-audio alignment, voice datasets are considered. For speech, TIMIT [31] offers 5h of clean solo English spoken by various speakers. We use 73.4%/13.3%/13.3% of data in training, validation, and test sets. For singing, DIMITRIOS proposes 3h of solo Greek byzantine singing [32]. We split it according to biphone (*i.e.*, two consecutive phonemes) distribution, making sure that unique biphones are in the train set, and balance less rare biphones between validation and test sets. The final split is 70.0%/10.0%/20.0%. The size of phonetic alphabet $\mathcal{A}$, with pause, is $L = 45$ for TIMIT and $L = 50$ for DIMITRIOS.

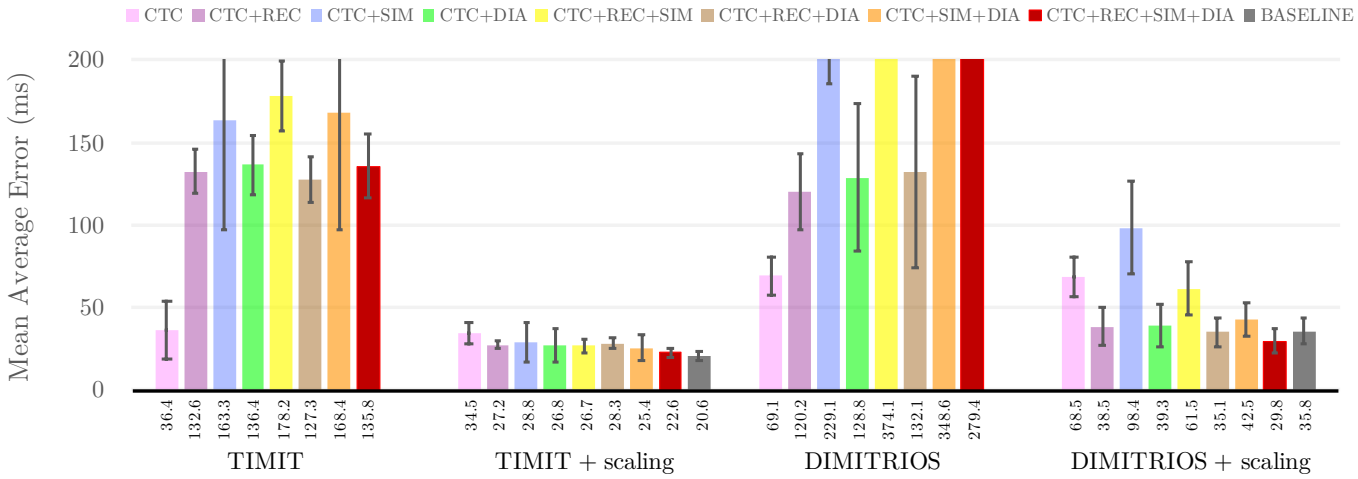

Figure 4: *Evaluation of alignment systems when mixing CTC, reconstruction (REC), structural (SIM) and monotonic (DIA) losses, with or without scaling. Error bars correspond to standard deviations. MAEs greater than 200ms are masked for the sake of readability.*

4.2. Pre-processing and training

Spectrograms are computed from mono, 16kHz signals with 1024-long Hamming window and hop length 256. The 1024 frequency bins are reduced to $B = 128$ with a Mel filterbank. Finally, only $F = 20$ coefficients are kept for F0-free MFCCs.

All convolution blocks are made of batch normalization, 512-filter `Conv1D` with a kernel of 3, and 0.2 dropout layers. When necessary (*i.e.*, **Fig. 2**), a final dense layer with relevant dimension and activation is applied. The number of attention heads has been fixed to $H = 4$ after a brief ablation study.

Trainings are done on a single GeForce GTX 1080 Ti. One epoch is composed of 128 steps, each processing 16-sample batch. Early stopping prevents overfitting. Codes are TF2.6-based and inspired from [33]. For all possible configurations, we minimize one or several objective(s), that are summed, with default ADAM optimizer and a learning rate set to 10^{-4}.

4.3. Results

4.3.1. Alignment predictions

In inference mode, the model is applied on a spectrogram and generates a posteriogram $\mathcal{P} \in [0,1]^{T\times(L+1)}$. Then, a forced alignment is computed on the (non-aligned) phonetic transcript $\mathbf{y} \in \mathcal{A}^{M\times 1}$ following the procedure in [6, 18], *i.e.*, cumulative score, blank distribution and beam search decoding.

4.3.2. Performance analyses

In **Fig. 4**, are depicted the alignment Mean Average Error (MAE), computed over each phone's edge (begin/end) time-stamp, on the test sets for all configurations and reference [18].

Regarding our contributions, one can see that (1) our neural baseline is suitable for clean voice alignment ; (2) there is a clear, significant impact of the proposed multi-training scaling procedure, changing the MAE's order of magnitude ; and (3) combining all introduced and scaled constraints with the CTC results in best performances, much comparable to the reference.

Yet, one must note that the similarity constraint by itself cannot guarantee alignment. In practice, posteriograms can be trickily shaped to have correct structures yet without predicting the full duration of labels. A positive impact is measured when coupled with envelope reconstruction and guided monotony.

The best configuration, CTC and all scaled constraints, leads to a MAE of 22.6ms on speech and 29.8ms on singing. This is in line with reference performances reported in **Fig. 4**. Although a bit weaker on speech (22.6ms *vs* 20.6ms), which will be addressed in future works, our network uses $10\times$ less parameters (4.5M *vs* 48M) and trains twice faster (2.1h *vs* 4.8h).

4.4. Perspectives

This work is meant to be continued in diverse research axes:

1) Informed guided monotony. In its current form, pseudo-diagonal matrix **D** used to force monotony carries a prior that all labels have similar duration, which is intrinsically not true. One could investigate a phoneme-informed or duration-focused approach by learning parameters defining matrix **D**.

2) Improve alignment precision. The results obtained on speech motivates for further improvement in alignment quality. For instance, pre-training the acoustic model with large speech corpora (*e.g.*, audio books) may be an interesting starting point.

3) Real world use cases. A major challenge is to apply our model on complete songs. Non-*a cappella* recordings are much harder to process due to the presence of music accompaniment. One can rely on singing voice extraction as a pre-processing step, though these separation algorithms introduce artifacts that might crop voice or some phonemes, which is hard to quantify.

5. Conclusion

This paper was aimed at constraining Connectionist Temporal Classification (CTC) to guarantee the emergence of temporal alignment properties from generated posteriograms. However currently limited to clean voice datasets, we have (1) designed a fully convolutional neural baseline with multi-head attention well-adapted to CTC-based alignment ; (2) presented a multi-task scaling method from worst-case scenario studies of each loss ; (3) shown that additional constraints such as envelope reconstruction, structural invariance and guided monotony are, once scaled, highly beneficial to phonetic alignment, inducing results in line with state-of-the-art precision.

6. Acknowledgement

This work has been funded by the French National Research Agency (ANR) project **ARS** (ANR-19-CE38-0001-01).

7. References

[1] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," *ACM/IEEE Supercomputing Conference (SC)*, vol. 148, pp. 369–376, 2014.

[2] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," *International Conference on Machine Learning (ICML)*, 2014.

[3] K. Xu, D. Li, N. Cassimatis, and X. Wang, "Lcanet: End-to-end lipreading with cascaded attention-ctc," *IEEE 13th International Conference on Automatic Face and Gesture Recognition*, pp. 548–555, 2018.

[4] J. Xue and J. Zhang, "A novel spec-cnn-ctc model for end-to-end speech recognition," *13th International Conference on Machine Learning and Computing*, pp. 141–145, 2021.

[5] H. Zhan, S. Lyu, Y. Lu, and U. Pal, "Densenet-ctc: An end-to-end rnn-free architecture for context-free string recognition," *Computer Vision and Image Understanding*, 2021.

[6] A. Hannun, "Sequence modeling with ctc," *Distill*, 2017, https://distill.pub/2017/ctc, accessed 18 March 2022.

[7] Y. Zhang, M. Pezeshki, B. P., S. Zhang, C. Laurent, Y. Bengio, and A. Courville, "Towards end-to-end speech recognition with deep convolutional neural networks," *Interspeech*, 2016.

[8] M. A. Roman, A. Pertusa, and J. Calvo-Zaragoza, "A holistic approach to polyphonic music transcription with neural networks," *Proceedings of the 20th International Conference on Music Information Retrieval (ISMIR)*, p. 731–737, 2019.

[9] L. Renault, A. Vaglio, and R. Hennequin, "Singing language identification using a deep phonotactic approach," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 271–275, 2021.

[10] A. Vaglio, R. Hennequin, M. Moussallam, G. Richard, and F. d'Alché Buc, "Audio-based detection of explicit content in music," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[11] A. Arzt and S. Lattner, "Audio-to-score alignment using transposition-invariant features," *Proceedings of the 19th International Conference on Music Information Retrieval (ISMIR)*, p. 592–599, 2018.

[12] H. Fujihara and M. Goto, "Lyrics-to-audio alignment and its application," *Dagstuhl Follow-Ups*, vol. 3, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012.

[13] D. Stoller, S. Durand, and S. Ewert, "End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[14] A. Vaglio, R. Hennequin, M. Moussallam, G. Richard, and F. d'Alché Buc, "Multilingual lyrics-to-audio alignment," *International Society for Music Information Retrieval Conference (ISMIR)*, 2020.

[15] T. Bluche, H. Ney, J. Louradour, and C. Kermorvant, "Framewise and ctc training of neural networks for handwriting recognition," *13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 81–85, 2015.

[16] H. Liu, S. Jin, and C. Zhang, "Connectionist temporal classification with maximum entropy regularization," *Advances in Neural Information Processing Systems*, vol. 31, pp. 831–841, 2018.

[17] E. Variani, E. McDermott, K. Lahouel, M. Bacchiani, and T. Bagby, "Sampled connectionist temporal classification," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4959–4963, 2018.

[18] Y. Teytaut and A. Roebel, "Phoneme-to-audio alignment with recurrent neural networks for speaking and singing voice," *Proceedings of Interspeech 2021*, pp. 61–65, 2021.

[19] J. Yuan, W. Lai, C. Cieri, and M. Liberman, "Using forced alignment for phonetics research," *Chinese Language Resources and Processing: Text, Speech and Language Technology. Springer*, 2018.

[20] D. Backstrom, M. C. Kelley, and T. B. V., "Forced-alignment of the sung acoustic signal using deep neural nets," *Canadian Acoustics*, vol. 47 No. 3, 2019.

[21] The Noun Project, "Singing song icon" from ProSymbols account (https://thenounproject.com/icon/singing-song-2106028/), "Sound Wave" from Alice Noir account (https://thenounproject.com/icon/sound-waves-2333598/) and "Treble clef icon" from Ben Avery account (https://thenounproject.com/icon/treble-clef-1620155/), web resource, https://thenounproject.com/, accessed 28 February 2022.

[22] K. Schulze-Forster, C. S. J. Doire, G. Richard, and R. Badeau, "Joint phoneme alignment and text-informed speech separation on highly corrupted speech," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[24] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[25] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4835–4839.

[26] H. Park, C. Kim, H. Son, S. Seo, and J.-H. Kim, "Hybrid ctc-attention network-based end-to-end speech recognition system for korean language," *Journal of Web Engineering*, pp. 265–284, 2022.

[27] L. Mao, "Number of alignments in connectionist temporal classification (ctc)," web resource, https://leimao.github.io/blog/CTC-Alignment-Combinations/, 2019, accessed 18 March 2022.

[28] J. Foote, "Visualizing music and audio using self-similarity," in *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, 1999, pp. 77–80.

[29] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," in *Proc. international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, p. 4784–4788.

[30] S. Liang, C. Deng, and Y. Zhang, "A simple approach to balance task loss in multi-task learning," in *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 2021, pp. 812–823.

[31] V. Zue, S. Seneff, and J. Glass, "Speech database development at mit: Timit and beyond," *Speech communitation 9(4)*, pp. 351–356, 1990.

[32] N. Grammalidis, K. Dimitropoulos, F. Tsalakanidou, A. Kitsikidis, P. Roussel, B. Denby, P. Chawah, L. Buchman, S. Dupont, S. Laraba *et al.*, "The i-treasures intangible cultural heritage dataset," in *Proceedings of the 3rd International Symposium on Movement and Computing*, 2016, pp. 1–8.

[33] Y. Soullard, C. Ruffino, and T. Paquet, "Ctcmodel: a keras model for connectionist temporal classification," research report, Université de Rouen Normandie, 2019, hal-02420358.