IDIAP RESEARCH REPORT

A Discriminative Decoder for the Recognition of Phoneme Sequences

David Grangier [1,2] Samy Bengio [1,2]

IDIAP–RR 05-67

November 2005

[1] IDIAP Research Institute, Martigny, Switzerland, {grangier,bengio}@idiap.ch
[2] Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

IDIAP Research Institute www.idiap.ch
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 05-67

A Discriminative Decoder for the Recognition of Phoneme Sequences

David Grangier Samy Bengio

November 2005

Abstract. In this report, we propose a discriminative decoder for the recognition of phoneme sequences, i.e. the identification of the uttered phoneme sequence from a speech recording. This task is solved as a 3 step process: a phoneme classifier first classifies each accoustic frame, then temporal consistency features (TCF) are extracted from the phoneme classifier outputs, and finally a sequence decoder identifies the phoneme sequence according to the TCF.

1 Notation

The training data consists of a set of sequences of acoustic vectors:

$$\mathcal{X} = (X^1, \ldots, X^{|\mathcal{X}|})$$

along with the corresponding phoneme labels:

$$\mathcal{Y} = (Y^1, \ldots, Y^{|\mathcal{X}|}).$$

The labels are considered to be *aligned* in the sense that an acoustic sequence X^i and Y^i have the same length $|X^i|$ and the j^{th} vector x^i_j of X^i is assigned to the phoneme class y^i_j. In the following, the number of phonemes is denoted N_{phone}, i.e. $\forall i, j, y^i_j \in \{1, \ldots, N_{phone}\}$.

2 Phoneme Classification

The phoneme classifier is a parametric function f_θ which takes as input a frame x and outputs a N_{phone}-dimensional vector denoted

$$f_\theta(x) = [f_\theta(x, 1), \ldots, f_\theta(x, N_{phone})].$$

Ideally, f_θ should classify each frame x such that the *true* phoneme y is assigned the highest output, i.e.

$$\forall p \neq y, f_\theta(x, y) - f_\theta(x, p) > 0. \tag{1}$$

The parameters θ of f are hence selected to minimize the number of non respected constraints in (1):

$$C^{0/1}_{phone} = \sum_{i,j,p} C^{0,1}_{phone}(x^i_j, y^i_j, p)$$

where

$$C^{0,1}_{phone}(x, y, p) = I\{p \neq y\} I\{f_\theta(x, y) - f_\theta(x, p) < 0\}$$

and $I\{\cdot\}$ is the indicator function.

This problem can be solved by applying stochastic gradient descent to the following upper bound of $C^{0/1}_{phone}$:

$$C_{phone} = \sum_{i,j,p} C_{phone}(x^i_j, y^i_j, p) \tag{2}$$

where

$$C_{phone}(x, y, p) = I\{p \neq y\} |1 - f_\theta(x, y) + f_\theta(x, p)|_+$$

and $z \to |z|_+$ is z if $z > 0$ and 0 otherwise. C_{phone} is actually an upper bound of $C^{0/1}_{phone}$ as,

$$\forall z \in \mathbb{R}, I\{z < 0\} \leq |1 - z|_+.$$

Moreover, if there exists θ^* such that $C_{phone} = 0$, then θ^* is a global minimum of both C_{phone} and $C^{0/1}_{phone}$ (with $C^{0/1}_{phone} = 0$). This means that, if the gradient descent procedure identifies θ^*, all the constraints in (1) will be verified.

The cost C_{phone} is actually similar to the optimized criterion in multi-class Support Vector Machine (SVM) [2]. However, other models than SVM may also be trained with this type of costs, for instance, f_θ could be a Multi-Layer Perceptron (MLP), see e.g. [1, 3] . In fact, the parameters of any differentiable function f_θ can be identified through (stochastic) gradient descent over C_{phone}.

3 Temporal Consistency Features

The phoneme classifier does not model temporal dependencies between frames. In order to recover some temporal context, we define simple binary features that measure the temporal consistency of the phoneme classifier output, such features being then used as input for the decoder (see Section 4). The intuition behind Temporal Consistency Features (TCF) relies on two simple observations: first, phonemes generally last longer than a single frame [4]. Second, even when the phoneme classifier fails to correctly classify one frame (i.e. assign the highest output to the correct class), the correct phoneme is generally among the best outputs. Hence, TCF features measure whether a phoneme *consistently* appear within the n-best classifier outputs for a certain duration d, as explained below.

Given a phoneme p, an integer $n < N_{phone}$, a duration d and a ratio α, the binary feature $\phi_{p,n,d,\alpha}$ is defined as follows:

$$\forall x^i_j, \phi_{p,n,d,\alpha}(x^i_j) = \begin{cases} 1 \text{ if } (i) \text{ is verified} \\ 0 \text{ otherwise} \end{cases}$$

where condition (i) corresponds to:

> "In X^i, there exists a subsequence S of duration d containing x^i_j such that the phoneme p is among the n-best phonemes according to f_θ for at least α percent of the frames of S."

The TCF features are then defined relying on ϕ for a large set of triplets,

$$\mathcal{T} = \left\{(n_1, d_1, \alpha_1), \ldots, (n_{|T|}, d_{|\mathcal{T}|}, \alpha_{|\mathcal{T}|})\right\},$$

this set being a-priori chosen using prior knowledge about the phoneme duration and the generalization performance of f_θ. Precisely, given $\mathcal{T}$, the TCF vector $\Phi(x^i_j)$ is defined for each frame x^i_j as a $N_{phone} \cdot |\mathcal{T}|$-dimensional vector whose components are

$$\left\{\phi_{p,n,d,\alpha}(x^i_j), \forall (p, n, d, \alpha) \in \{1, \ldots, N_{phone}\} \times \mathcal{T}\right\}.$$

Such a vector is hence high dimensional (i.e. $\Phi(x^i_j) \in \{0,1\}^{N_{phone} \cdot |\mathcal{T}|}$) but sparse (i.e. most $\phi_{p,n,d,\alpha}(x^i_j)$ are null). These TCF features hence aim at detecting complex temporal patterns through the use of many simple binary features. Such approaches have already shown to be effective in other domains, e.g. Haar-like features allow to model complex spacial patterns through the use of many simple binary pattern detectors [7].

4 A Discriminative Decoder

The sequence classifier G takes as input a sequence of TCF vectors

$$Z^i = (z^i_1, \ldots, z^i_{|X^i|}),$$

where $\forall i, j, z^i_j = \Phi(x^i_j)$, and a sequence of phonemes P,

$$P = (p_1, \ldots, p_{|X^i|}),$$

where $\forall j, p_j \in \{1, \ldots, N_{phone}\}$, and outputs a real value:

$$G_{\theta'}(Z^i, P) = \sum_{j=1}^{|X^i|} g_{\theta'}(z^i_j, p_j). \tag{3}$$

Ideally, $G_{\theta'}$ should be such that, for any sequence Z^i, the output $G_{\theta'}(Z^i, Y^i)$ for the *true* phoneme sequence Y^i is higher than for any other phoneme sequence,

$$\forall i, \forall P \neq Y^i, G_{\theta'}(Z^i, Y^i) > G_{\theta'}(Z^i, P), \tag{4}$$

Hence, the parameters θ' are selected to minimize the amount of non-respected constraints in (4):

$$C^{0/1}_{decoder} = \sum_i C^{0/1}_{decoder}(Z^i, Y^i) \tag{5}$$

where

$$C^{0/1}_{decoder}(Z^i, Y^i) = \frac{1}{|W^i|} \sum_{P \in W^i} I\{G_{\theta'}(Z^i, Y^i) - G_{\theta'}(Z^i, P) < 0\}, \tag{6}$$

W^i corresponding to the set of *wrong* phoneme sequences, i.e.

$$W^i = \{1, \ldots, N_{phone}\}^{|X^i|} \setminus \{Y^i\}.$$

The normalization with respet to W^i further hypothesizes that each training sequence X^i should led to the same penalty (i.e. $C^{0/1}_{decoder}(Z^i, Y^i) = 1$) when none of its corresponding constraints are respected.

As for the phoneme classifier, we propose to select θ' through the minimization of an upper bound of $C^{0/1}_{decoder}$ using stochastic gradient descent. For that purpose, we introduce the function $H_{\theta'}$ which is defined as follows,

$$H_{\theta'}(Z^i, P) = \sum_{j=1}^{|X^i|} h_{\theta'}(z^i_j, p_j).$$

where,

$$h_{\theta'}(z^i_j, p_j) = g_{\theta'}(z^i_j, p_j) - I\{p_j = y^i_j\}.$$

Then, we select θ' through the minimization of

$$C_{decoder} = \sum_i \frac{1}{|W^i|} \sum_{P \in W^i} |H_{\theta'}(Z^i, P) - H_{\theta'}(Z^i, Y^i)|_+ \tag{7}$$

This cost $C_{decoder}$ is actually an upper bound of $C^{0/1}_{decoder}$ as shown in the Appendix A and the minimization of $C^{0/1}_{decoder}$ can be performed through the minimization of $C_{decoder}$. However, this optimization problem may rise some tractability issues as the sum over W^i could hardly be performed for long sequences (i.e. this set grows exponentially[1] with sequence length, $|W^i| = (N_{phone})^{|X^i|} - 1$). Hence we introduce the following upper bound of $C_{decoder}$,

$$C^{max}_{decoder} = \sum_i |\max_{P \in W^i} H_{\theta'}(Z^i, P) - H_{\theta'}(Z^i, Y^i)|_+$$

which also bounds $C^{0/1}_{decoder}$ by transitivity. $C^{max}_{decoder}$ is hence much easier to compute as the identification of the max over W^i only involves a best-path-decoding whose cost grows linearly with respect to sequence length. Moreover, like for the phoneme classifier cost, if there exists θ^* s.t. $C^{max}_{decoder} = 0$, then θ^* is a global minimum of both $C^{max}_{decoder}$ and $C^{0/1}_{decoder}$ (with $C^{0/1}_{decoder} = 0$). This means that if this optimum is identified during training, all the constraints in (4) will be verified.

5 Incorporating Phoneme Transitions

Like in a Hidden Markov Model (HMM) [4], it may be desirable to also penalize or promote some phoneme transitions in the decoder. Moreover, it may also be interesting to model phoneme duration.

[1]The cost of the computation of the gradient $\frac{\partial C_{decoder}}{\partial \theta'}(Z^i, Y^i)$ may however be lower than $O(|W^i|)$ when efficient dynamic programming techniques are used.

These two characteristics can easily be incorporated into the model described above. For that purpose, we modify the function $G_{\theta'}$ as follows:

$$\begin{aligned} G_{\theta'}(Z^i, P) &= \sum_{j=1}^{|X^i|} g_{\theta'}^{full}(z_j^i, p_j) \\ &= \sum_{j=1}^{|X^i|} g_{\theta'}^{e}(z_j^i, p_j) + I\{p_{j-1} \neq p_j\} \cdot \left(g_{\theta'}^{t}(p_{j-1}, p_j) + g_{\theta'}^{d}(p_{j-1}, d_{j-1})\right) \end{aligned} \tag{8}$$

where d_{j-1} corresponds to the duration of phoneme p_{j-1} in the sequence p (i.e. the length of the longuest homogene sequence preceeding p_j). In this definition (8), the function $g_{\theta'}^e$ plays the same role as g in (3) while the functions $g_{\theta'}^t$, $g_{\theta'}^d$ are respectively used to model phoneme transition and phoneme duration. This modified model can then be trained with the cost $C_{decoder}^{max}$, this only requires to redefine h as

$$h_{\theta'}^{full}(z_j^i, p_j) \quad = \quad g_{\theta'}^{full}(z_j^i, p_j) - I\{p_j = y_j^i\}.$$

Like for the phoneme classifier, the parameterization of the functions $g_{\theta'}^e$, $g_{\theta'}^t$, $g_{\theta'}^d$ are not given. Similarly to f_θ, these functions should simply be differentiable in their parameters in order to apply gradient descent. A possible choice could be, for instance, to select MLPs for g^e and g^d and a table for g^t.

6 Extension to Weakly Aligned Sequences

The presented approach requires training acoustic sequences to be labeled with aligned phoneme sequence, i.e. the phoneme sequence provided should provide phoneme boundary with a resolution of one frame. However, such an alignment can hardly be obtained automatically [5] and is also very difficult to label manually. Hence, we propose an extension of the above approach for *weakly-aligned* sequences. In the following, we first define *weak-alignment* and we then describe how the proposed approach can be trained from such data.

The concept of *weak-alignment* considers that phoneme boundaries do not precisely occur between two frames but during several frames, as shown in Figure 1. Hence, a weakly aligned phoneme sequence Y is a succession of stable sub-sequences S_i and transitional sub-sequences T_i, i.e.

$$Y = S_0(T_1 S_1) \dots (T_k S_k)$$

where $k \geq 0$. Stable and transitional frames are denoted with the following formalism,

$$\forall j, y_j = (y_{j,1}, y_{j,2}),$$

which either corresponds to the phoneme $y_{j,1}^i$ if $y_{j,1} = y_{j,2}$ or to the transition $y_{j,1} \rightarrow y_{j,2}$ otherwise. According to such definition, the weakly align sequence Y hence encompasses all aligned sequences $Y^{aligned}$ such that, the phoneme labels are identical for the stable sub-sequences, i.e.

$$\forall i, \forall j \in S_i, y_j^{aligned} = y_{j,1} \tag{9}$$

and, the phoneme labels either corresponds to $y_{j,1}$ or $y_{j,2}$ for transitional sub-sequences with the additional constraint that each of these subsequences contains only a single transition which is $y_{j,1} \rightarrow y_{j,2}$, i.e.

$$\forall i, \quad \exists t \in T_i \text{ s.t. } \forall j \in T_i, \begin{cases} j < t \Rightarrow y_j^{aligned} = y_{j,1} \\ j \geq t \Rightarrow y_j^{aligned} = y_{j,2} \end{cases} \tag{10}$$

We now present how the phoneme classifier and the sequence decoder can be adapted to be trained from such data.

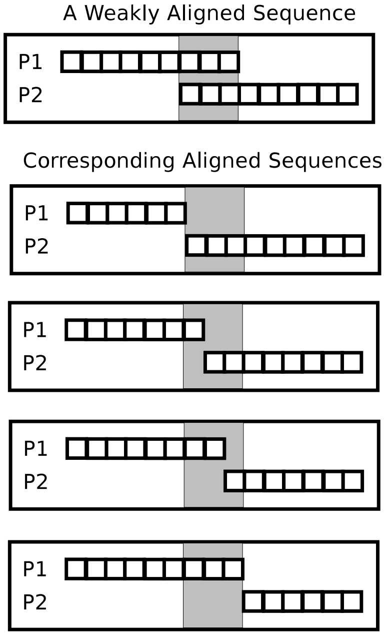

Figure 1: An example of weak alignment: this weak alignment specifies that the transition between phoneme $P1$ and phoneme $P2$ is located after frame 6 and before frame 10, and hence 4 frame-level alignments are possible according to such information.

6.1 Phoneme Classifier Training

Weak alignment specifies that each frame x^i_j should be classified either as $y^i_{j,1}$ or as $y^i_{j,2}$. Therefore, the phoneme classifier can be trained such that its outputs for the two possible *true* phonemes would ideally be higher than those of any other phoneme. Consequently, a cost to minimize can be

$$C^{0/1}_{phone} = \sum_{i,j,p} C^{0,1}_{phone}(x^i_j, y^i_j, p)$$

where

$$\begin{array}{l} C^{0,1}_{phone}(x, y, p) = \\ \quad I\{p \notin y\}\left(I\{f_\theta(x, y_1) - f_\theta(x, p) < 0\} + I\{f_\theta(x, y_2) - f_\theta(x, p) < 0\}\right) \end{array}.$$

We then similarly introduce C_{phone} which bounds $C^{0,1}_{phone}$, i.e.

$$C_{phone} = \sum_{i,j,p} C_{phone}(x^i_j, y^i_j, p)$$

where

$$\begin{array}{l} C_{phone}(x, y, p) = \\ \quad I\{p \notin y\}\left(|1 - f_\theta(x, y_1) + f_\theta(x, p)|_+ + |1 - f_\theta(x, y_2) + f_\theta(x, p)|_+\right) \end{array}.$$

Stochastic Gradient Descent can then be applied to select the parameter vector θ which minimizes C_{phone}.

6.2 Sequence Decoder Training

As explained above, each weakly aligned sequence Y^i corresponds to a set of frame-level assigned sequences (i.e. the set of sequences whose phoneme boundaries are located within the ranges specified by Y^i). If this set is referred to as R^i, the phoneme decoder $G_{\theta'}$ should identify a phoneme sequence $P^R \in R^i$ from the TCF sequence Z^i. In other words, the function $P \to G_{\theta'}(Z^i, P)$ should be maximal for a sequence $P^R \in R^i$. This means that, our goal is to find the parameters θ' which minimizes

$$C^{0/1}_{decoder} = \sum_i C^{0/1}_{decoder}(Z^i, Y^i) \tag{11}$$

where

$$C^{0/1}_{decoder}(Z^i, Y^i) = \frac{1}{|W^i|} \sum_{P^W \in W^i} I\{\max_{P^R \in R^i} G_{\theta'}(Z^i, P^R) - G_{\theta'}(Z^i, P^W) < 0\}, \tag{12}$$

and W^i corresponds to the set of *wrong* phoneme sequences, i.e.

$$W^i = \{1, \ldots, N_{phone}\}^{|X^i|} \setminus R^i.$$

For that purpose, we redefine $C_{decoder}$ with the same approach as above,

$$C_{decoder} = \sum_i |H_{\theta'}(Z^i, P^W) - \max_{P^R \in R^i} H_{\theta'}(Z^i, P^R)|_+$$

where H is defined as above,

$$H_{\theta'}(Z^i, P) = \sum_{j=1}^{|X^i|} h_{\theta'}(z^i_j, p_j).$$

and h is redefined to take into account weak alignement labels. In fact, we define $h_{\theta'}(z^i_j, p_j)$ differently depending whether p_j is located in a stable or a transitional subsequence of Y^i. If j belongs to a stable subsequence S,

$$h_{\theta'}(z^i_j, p_j) = g^{full}_{\theta'}(z^i_j, p_j) - I\{p_j = y^i_{j,1}\}.$$

If j belongs to a transitional subsequence T,

$$\begin{aligned} h_{\theta'}(z^i_j, p_j) \quad = \quad & g^{full}_{\theta'}(z^i_j, p_j) \\ & -I\{p_j = y^i_{j,1} \text{ and } n_{trans}(y^i_{j,1}, P, b_T, j) = 0\} \\ & -I\{p_j = y^i_{j,2} \text{ and } n_{trans}(y^j_{j,2}, P, b_T, j) \leq 1\} \end{aligned}$$

where b_T corresponds to the starting point of the transitional phase T and $n_{trans}(y^i_{j,1}, P, b_T, j)$ measures the number of transitions to phoneme $y^i_{j,1}$ between b_T and j in the sequence P, i.e.

$$n_{trans}(y^i_{j,1}, P, b_T, j) = \sum_{k=b_T+1}^{j} I\{p_{k-1} \neq y^i_{j,1} \text{ and } p_k = y^i_{j,1}\}.$$

As showing that $C_{decoder} \geq C^{0/1}_{decoder}$ is less obvious than for the phoneme classifier cost, a proof is presented in the Appendix A. Finally, the cost $C_{decoder}$ is itself bounded by,

$$C^{max}_{decoder} = \sum_i |\max_{P^W \in W^i} H_{\theta'}(Z^i, P^W) - \max_{P^R \in R^i} H_{\theta'}(Z^i, P^R)|_+.$$

7 Conclusions

In this report, we presented a discriminative approach for the recognition of phoneme sequences. This task has been divided into two steps: frame classification and sequence decoding. For frame classification, we introduced a classifier which is trained with a cost inspired by multi-class SVM criterion. For sequence decoding, a discriminative approach has also been adopted, the objective being to train a model G such that, for any accoustic sequence X and any phoneme sequence P, $G(X, P)$ is maximal (i.e. $G(X, P) = \max_S G(X, S)$) when P actually corresponds to the *true* phoneme sequence of X. Moreover, we have introduced the Temporal Consistency Features that should allow the decoder to be more robust with respect to frame classification errors. The underlying idea is to feed the decoder with features which do not only rely on the descision of the phoneme classifier for a single frame but which also depend on the neighboring frame descisions.

This proposed approach is hence an alternative to the state-of-the-art Hidden Markov Model. Compared to the HMM, this model has two major differences, discriminative training criterion and non-probabilistic parameterization, which is advantageous for several theoritical reasons [6]. However, even if it is theoretically attractive, the proposed approach can only be validated through empirical comparisons with respect to HMM. Hence, our future works will focus on such evaluations over benchmark datasets.

Acknoledgments The authors would like to thank Joseph Keshet for his suggestions and comments. This work has been performed with the support of the Swiss NSF through the NCCR–IM2 project. It was also supported by the PASCAL European Network of Excellence, funded by the Swiss OFES.

References

[1] Ronan Collobert and Samy Bengio. Links between perceptrons, mlps and svms. In *International Conference on Machine Learning, ICML*, 2004.

[2] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research, JMLR*, 2002.

[3] David Grangier and Samy Bengio. Inferring document similarity from hyperlinks. In *Conference on Information and Knowledge Management, CIKM*, 2005.

[4] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall, 2001.

[5] Joseph Keshet, Shai Shalev-Shwartz, Yoram Singer, and Dan Chazan. Phoneme alignment based on discriminative learning. In *European Conference on Speech Communication and Technology, INTERSPEECH*, 2005.

[6] Yann LeCun and Fu-Jie Huang. Loss functions for discriminative training of energy-based models. In *International Workshop on Artificial Intelligence and Statistics, AISTATS*, 2005.

[7] Paul Viola and Michael Jones. Robust real-time object detection. *International Journal of Computer Vision, IJCV*, 2002.

A Appendix: Bounding the Decoder 0/1 Loss

In this section, we want to show that $C_{decoder} \geq C^{0/1}_{decoder}$. There are slight differences in this proof regarding whether training phoneme labels are provided as aligned sequences or weakly-aligned sequences, hence we present these two cases successively.

A.1 Aligned Phoneme Sequence Case

We should show that

$$C_{decoder} \geq C^{0/1}_{decoder}.$$

According to the definition of both costs, it would actually be sufficient to show that:

$$\begin{aligned}&\forall i, \forall P \in W^i,\\ &\qquad |H_{\theta'}(Z^i, P) - H_{\theta'}(Z^i, Y^i)|_+ \geq I\{G_{\theta'}(Z^i, Y^i) - G_{\theta'}(Z^i, P) < 0\}.\end{aligned}$$

For that purpose, we rewrite the right-hand side of the equation and then bound it: given i and $P \in W^i$,

$$\begin{aligned}&|H_{\theta'}(Z^i, P) - H_{\theta'}(Z^i, Y^i)|_+\\ &\quad = \left|\sum_{j=1}^{|X^i|} h_{\theta'}(z^i_j, p_j) - h_{\theta'}(z^i_j, y^i_j)\right|_+\\ &\quad = \left|\sum_{j=1}^{|X^i|} g_{\theta'}(z^i_j, p_j) - g_{\theta'}(z^i_j, y^i_j) - \sum_{j=1}^{|X^i|} \left(I\{p_j = y^i_j\} - 1\right)\right|_+\\ &\quad = \left|\sum_{j=1}^{|X^i|} \left(1 - I\{p_j = y^i_j\}\right) - G(Z^i, Y^i) + G_{\theta'}(Z^i, P)\right|_+\end{aligned}$$

As the sequence P belongs to W^i, $P \neq Y^i$. This means that P differs from Y^i on at least one of its phoneme label, and hence,

$$\sum_{j=1}^{|X^i|}(1 - I\{p_j = y^i_j\} - 1) \geq 1.$$

This inequality leads to the following bound

$$|H_{\theta'}(Z^i, P) - H_{\theta'}(Z^i, Y^i)|_+ \geq \left|1 - G(Z^i, Y^i) + G_{\theta'}(Z^i, P)\right|_+$$

since $x \rightarrow |x|_+$ is an increasing function over $\mathbb{R}$. Then, we can conclude that

$$|H_{\theta'}(Z^i, P) - H_{\theta'}(Z^i, Y^i)|_+ \geq I\{G_{\theta'}(Z^i, Y^i) - G_{\theta'}(Z^i, P) < 0\}$$

since $\forall x \in \mathbb{R}, |1 - x|_+ \geq I\{x < 0\}$.

A.2 Weakly Aligned Phoneme Sequence Case

Similary to the above proof, we want to show that

$$\begin{aligned}&\forall i, \forall P^W \in W^i,\\ &\qquad |H_{\theta'}(Z^i, P^W) - \max_{P^R \in R^i} H_{\theta'}(Z^i, P^R)|_+ \geq I\{\max_{P^R \in R^i} G_{\theta'}(Z^i, P^R) - G_{\theta'}(Z^i, P^W) < 0\}.\end{aligned} \tag{13}$$

which is a sufficient condition to $C_{decoder} \geq C^{0/1}_{decoder}$ according to the definitions of both cost. This proof is divided in two steps: first, we show that

$$\begin{aligned}&\forall (P^W, P^R) \in W^i \times R^i,\\ &\qquad |H_{\theta'}(Z^i, P^W) - H_{\theta'}(Z^i, P^R)|_+ \geq I\{G_{\theta'}(Z^i, P^R) - G_{\theta'}(Z^i, P^W) < 0\}\end{aligned} \tag{14}$$

and then,

$$\arg\max_{P^R \in R^i} H_{\theta'}(Z^i, P^R) = \arg\max_{P^R \in R^i} G_{\theta'}(Z^i, P^R). \tag{15}$$

These two propositions would then obviously imply (13).

Proof of (14). We first remark that, given a phoneme sequence P,

$$\forall j, h_\theta(z^i_j, p_j) = g_\theta(z^i_j, p_j) - 1 \Leftrightarrow P \in R^i.$$

This proposition is also equivalent to

$$\exists j \text{ s.t. } h_\theta(z^i_j, p_j) = g_\theta(z^i_j, p_j) \Leftrightarrow P \in W^i.$$

since R^i and W^i are complementary sets. This proposition then leads to, given $(P^W, P^R) \in W^i \times R^i$,

$$\begin{cases} H_{\theta'}(Z^i, P^R) = G_{\theta'}(Z^i, P^R) - |Z^i| \\ H_{\theta'}(Z^i, P^W) \geq G_{\theta'}(Z^i, P^W) - |Z^i| + 1. \end{cases}$$

This implies that

$$H_{\theta'}(Z^i, P^W) - H_{\theta'}(Z^i, P^R) \geq 1 + G_{\theta'}(Z^i, P^W) - G_{\theta'}(Z^i, P^R).$$

As $x \rightarrow |x|_+$ is an increasing function, we have

$$|H_{\theta'}(Z^i, P^W) - H_{\theta'}(Z^i, P^R)|_+ \geq |1 + G_{\theta'}(Z^i, P^W) - G_{\theta'}(Z^i, P^R)|_+$$

which leads to

$$|H_{\theta'}(Z^i, P^W) - H_{\theta'}(Z^i, P^R)|_+ \geq I\{G_{\theta'}(Z^i, P^R) - G_{\theta'}(Z^i, P^W) < 0\}$$

since $\forall x, |1 - x|_+ \geq I\{x < 0\}$. □

Proof of (15). As mentioned above, we have

$$\forall P^R \in R^i, H_{\theta'}(Z^i, P^R) = G_{\theta'}(Z^i, P^R) - |Z^i|,$$

which implies that the maximum of $P \rightarrow H_{\theta'}(Z^i, P)$ and $P \rightarrow G_{\theta'}(Z^i, P)$ over R^i is reached at the same point. □