ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

Phone Duration Based Confidence Scoring

[1] Punnoose A K

Abstract: This paper discuss about an approach of making a confidence scoring for phone duration in speech recognition. A confidence scoring mechanism is derived out of correspondence between an Hidden Markov Model(HMM) based forced aligner and a Multi Layer Perceptron(MLP) based frame classifier. Phone duration for noise is also factored into the approach which makes it more reliable.

Key Words — **Noise Robustness, Neural Networks, Interactive Voice Response Systems, Phone Duration**

I. INTRODUCTION

In automatic speech recognition and speech synthesis, phone duration models are mostly treated as an add-on. In most speech recognition systems, temporal aspects of speech are implicitly modeled by HMMs, rather than explicitly modelled. The speech recognition system is built in such a manner to make it independent of the speaking rate in general. Except in the case of very fast speech or slow speech, the issue is not needed to be addressed separately.

In pronunciation evaluation tasks, rate of speech is a critical component. Rate of speech is even the major factor in deciding one's fluency. In automatic pronunciation evaluation, explicit models for the duration of different phones, are needed. In most of the cases a Gaussian Mixture Model is used to model phone duration. One of the assumptions is phone duration modeling is that the duration for a consonant remains the same across the speakers, while it's the vowels which get shortened or elongated.

In this paper, an approach to provide a confidence score for phone duration, is discussed. An HMM based Forced Aligner and an MLP based frame classifier is used as the tools, for deriving some required statistics out of the speech data. A simple approach is discussed with its positives and drawbacks, to assign a confidence score based on the phone duration. The statistics derived from the noise data ensures the robustness of the approach.

II. PROBLEM DEFINITION

Given MLP output, a set of frames classified as a single phone, how to derive a confidence measure for that chunk of phone, using a forced aligner.

III. PRIOR WORK

Phone duration models in the very basic form assigns a score to a given duration of phone. Though most speech recognition systems ignore the phone duration, there has been many approaches to incorporate the duration information into the ASR process. In [1], Ron Dong et al. proposed syllable duration modeling with the gamma distribution for Mandarin continuous digits recognition system. In [2], authors discuss about Classification and Regression Tree for modeling the phone duration.A number of features and their usefulness for segmental duration prediction is assessed. They use Root Mean Square Prediction Error as an evaluation mechanism.

In [3], authors use a neural network for learning the phone durations. The input features are derived from the phone identities of a context window of phones, along with the durations of preceding phones within that window. In [4], author use a phone duration model based on a learned Deep Neural Network based acoustic model. The duration model calculates the probability density function of phone duration from phones contextual features using a neural network which is then applied for word lattice rescoring. In [5], authors deal with word duration rather than phone duration. Word durations are represented by log normal densities with a method of predicting new infrequent words by using well represented sub-word units.

Most of the phone duration models are used for improving the accuracy of a HMM based speech recognition system, by trying to incorporate the duration models into the HMM decoding process. In this paper a related but different issue is addressed. Given a frame classifier and a forced aligner used to build the frame classifier, how can we use the forced alignment information, along with the frame classifier to provide a

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

confidence score for phone chunks. The confidence score heavily depends on the accuracy of the frame classifier.

IV. APPROACH OUTLINE

The approach is outlined as following
1) Get the count of instances of different chunk sizes for every phone, from the forced aligner.
2) Use the same data for getting the getting the count of instances of different chunk sizes for every phone, from the frame classifier.
3) Take the chunk size where the count difference between 2 measures are minimum.
4) Threshold the difference between the two measures and assign a scoring mechanism, where the chunk size with minimum difference gets highest score
5) Use the statistics from the pure noise data, to make the confidence scoring mechanism more robust.

V. BASIC EXPERIMENTAL PLATFORM

A. Rationale for Voxforge as Training Data

For Experiments Voxforge data is used, which is available free for public use. The reason for selecting Voxforge data is multi-fold. First is that it's telephonic narrow band data. Second and foremost reason is that it's recorded in an uncontrolled environment by different people with different accent, with different mother tongue, etc. This will give the necessary variability in the data, which is very much crucial for making a speaker independent telephonic information access system. This is very much against the popular notion of using a very well known database like TIMIT, as the focus here is on real world telephonic IVRS, where the user response is simply silence or background speech, or just murmuring, or traffic noise, or noise of any other kind. A rough approximation of analyzing a real world speech based information access system will show that roughly only 20% of the user utterance is of any significant speech content. This heavily bias us to use a speech database which is uncontrolled and with wide variability.

B. Basic Neural Network Training Details

A neural network is trained to predict phones from speech features. Perceptual Linear Prediction Coefficients(plp) are used as feature. Standard 42 phone set of English is used as the labels. Mini batch gradient descent is used as the training mechanism. Cross Entropy Error is used as the measure for backpropagation training. 3 hidden layers are used and weights of MLP are initialized randomly between -1 and +1.

C. Noise Data Details

Pure background noise from CHiME4 Dataset is used as noise data. Background noise in various environment like street, bus, etc are used. Unlike older CHiME datasets, CHiME4 is not segregated based on SNR.

VI. APPROACH & ANALYSIS

Define chunk count variables μ, λ and η for forced alignment based chunk counts, frame classifier based chunk counts for speech and frame classifier based chunk counts for noise respectively. A chunk is a continuous sequence of the same phone, corresponding to a frame. Chunk size corresponds to the phone duration. Chunk width for a phone, based on the forced alignment, is found from the forced aligned output. Count of each chunk width is calculated. Same is done for the frame classifier decoded output for speech and noise. The chunk count variable takes the form, $\mu_{px}, \lambda_{px,} \eta_{px}$ where p represents the phone and x is the chunk width

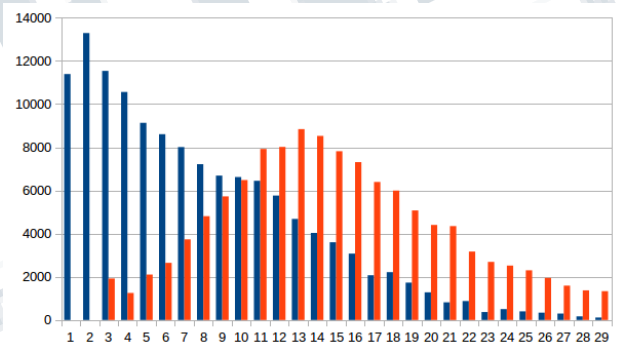

Fig. 1. Phone /aa/ : Expected vs Predicted Chunk Counts

Fig 1 plots the counts λ and μ for the phone /aa/. μ is plotted in red and λ is blue. It's apparent from the plot that there is a shift in the peak between λ and μ. While forced alignment based count is more spread around its peak both to the left and to the right almost symmetrically, the frame classifier based counts are skewed more towards the smaller chunk size. This difference in the chunk size is due to the fact that, in forced alignment, the parameter to choose is the starting and ending position of the phone, not the phone itself. And moreover it's based on a likelihood scoring by HMMs. On the other hand a frame classifier is discriminatively trained. For a frame classifier to classify a set of contiguous frames as a single phone, each of the frame has to be independently classified as the same phone. In the case of a likelihood scoring of a

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

chunk, if for a frame in the chunk gets a low score, it could be still included in the chunk, if some other adjacent frame cover the required likelihood for the low scoring frame. The bottom line here is that in forced alignment, the scoring is based on a chunk, which in turn makes it more fuzzy in terms of boundaries.

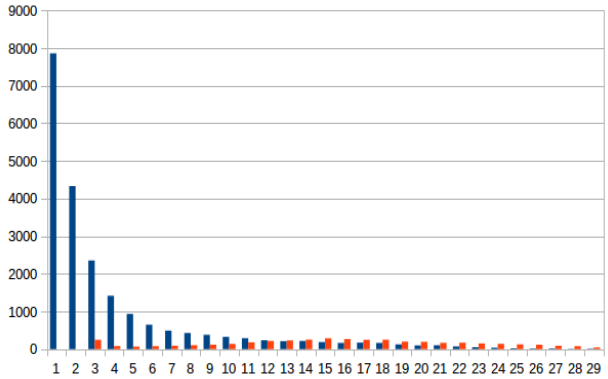

Fig. 2. Phone /aw/ : Expected vs Predicted Chunk Counts

Fig 2 plots both the counts for the phone /aa/. It's more or less similar to the Fig 1, except the fact that counts are less, due to the data. The skewness of the frame classifier to the smaller chunk width is evident. Note that in all of the figures, the count is taken only for chunk size upto 30. There are phones with chunk sizes above 30 frames detected, but not taken into consideration in this analysis.

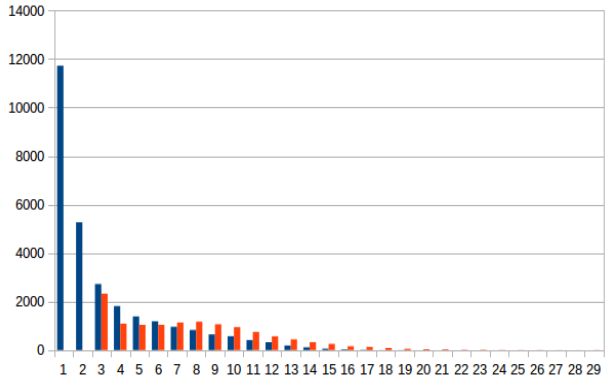

Fig. 3. Phone /p/ : Expected vs Predicted Chunk Counts

Fig 3 is a plot for phone /p/. It is evident the overall phone width for the forced alignment itself has been shifted to left, ie, around 6-8 frames, as opposed to that of the previous 2 phones /aa/ and /aw/. This is true to the understanding that stops are of less duration compared to a vowel.

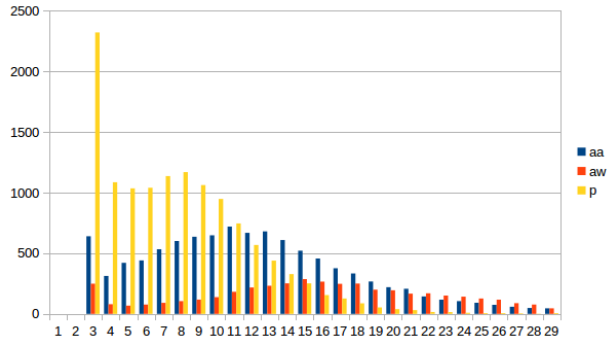

Fig. 4. Chunk Counts

In fig 4, the forced alignment based chunk width count is plotted, for the three phones /aa/, /aw/ and /p/. Disregarding the count, the point to be noted here is that all three phones are centered differently, thus suggesting that forced aligner is aligning in a plausible correct manner.

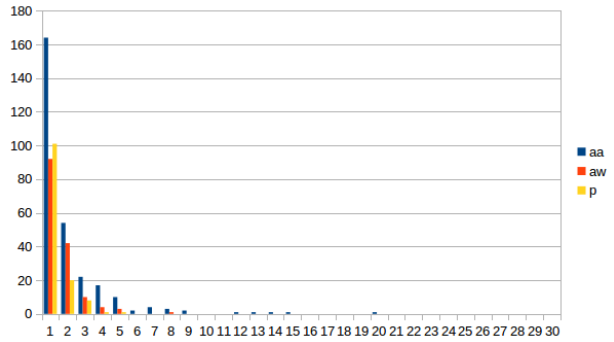

Fig. 5. Chunk Counts

Fig 5 shows the chunk size count of noise for 3 phones. /aa/,/aw/ and /p/. The noise data used is background noise of CHiME dataset. The counts are less because of the less amount of noise data available. It is clear from the plot that the count mass is more focused on the small chunk size.

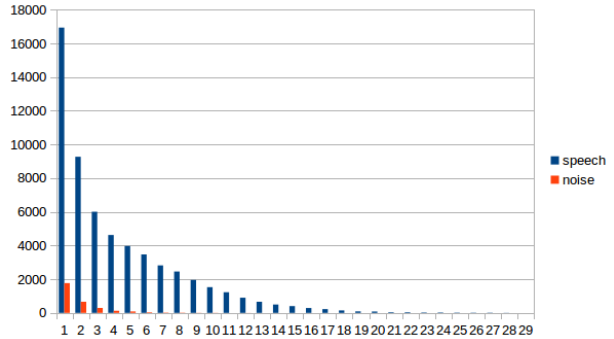

Fig. 6. Chunk Counts for phone /l/

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

Fig 6 shows the count of chunk size of the lateral phone /l/ for speech and noise, for the frame classifier. Despite the sparsity in noise data, there is not even one chunk of size > 12. What this hints is that, noise data tends to concentrate more on small chunk size irrespective of the phone. The overall goal is to find a method which assigns a scoring mechanism for a phone, based on the decoded chunk width and the forced aligned chunk width. A mechanism to assign a confidence score to chunk size is described. The issue here is in figuring out the how well the forced alignment based chunks matches with that of frame classifier based chunks. Rather than calculating a chunk sized based confusion matrix, an approach, based on a set of assumptions and constraints is discussed.

1) Assumptions:

1) As for certain phones, $\sum_{i=1}^{T} i\mu_{pi} > \sum_{i=1}^{T} i\lambda_{pi}$, a chunk of size m, be misrecognized only as with chunk size $n < m$.
2) A chunk of size m will be misrecognized as a chunk with size $m-1$, with probability $\frac{1}{2}$ and as a chunk of size $m-2$ with probability $\frac{1}{4}$. Generalizing, a chunk of size m , will be misrecognized as a chunk of size $m-i$ with probability $\frac{1}{2^i}$.
3) Following the above assumption, if there are M chunks of size m, then the expected chunk misrecognitions are $\frac{M}{2}$ of size $m-1$, and $\frac{M}{2^i}$ of size $m-i$.

The first assumption hold true for a number of phones, especially for stops. This is the case where forced aligner assigns more duration than the actual duration to the phones. For such phones, there is a chance for such phones to be recognized to a smaller chunk size, than that is assigned. This is the basis for the second assumption. With the above assumptions a confidence scoring function is defined, with the maximum confidence at

$$k = \arg\min_{x} (\mu_{px} - \lambda_{px}) \text{ s.t } \mu_{px} > \lambda_{px} \text{ and } \eta_{px} = 0$$

To reinforce the above found chunk length k, which the forced aligner and the frame classifier both agree upon, to be the ideal chunk length, another constraint is put forth as follows,

$$\frac{\lambda_{pk}}{2} > \sum_{i=k}^{i=T} \frac{1}{2^i}(\mu_{pi} - \lambda_{pi})$$

What the above conditions ensures is that total number of chunks of size > k, which got misrecognized as a chunk of size k, should not be more than half of the count of the chunk of size k.

The rationale for choosing the chunk size, for which there is minimal difference between count from 2 different approaches is as follows. In the speech detection tasks, it is the frame classifier rather than the forced aligner, which is employed. There is no much point in analyzing the ideal duration of a phone from forced aligner. What matters is to what extend the frame classifier agrees with the forced aligner, in terms of the phone duration. The reason frame classifier is not able to detect longer frames consistently is because the stationarity aspect of a long phone, as in the case of vowels when uttered, is compromised. It could be as simple as a slight inflection towards the end part of the vowel, which will be picked by the classifier as a seperate phone. This is evident from the presence of large count of small chunks of 1 and 2 frames in all of the plots discussed above.

Any simple scoring mechanism which takes into account the count of longer chunks, as well as the difference between chunk counts from forced aligner and frame classifier, can be employed. One such scoring mechanism is given

$$c_j = \left(1 - \frac{\mu_{pj} - \lambda_{pj}}{\sum_{i=k}^{T}(\mu_{pi} - \lambda_{pi})}\right) \frac{\lambda_{pi}}{\sum_{i=k}^{T} \lambda_{pi}} \, , \forall j = k, k+1, ..T$$

where k is the chunk size with maximum confidence, and T is the maximum chunk size. Note that there is no score assigned for chunk size from 0 to k−1. This is because the small chunks are not reliable enough to calculate any confidence score. Moreover, noise tends to get recognized as small chunks.

A. A Failure Case

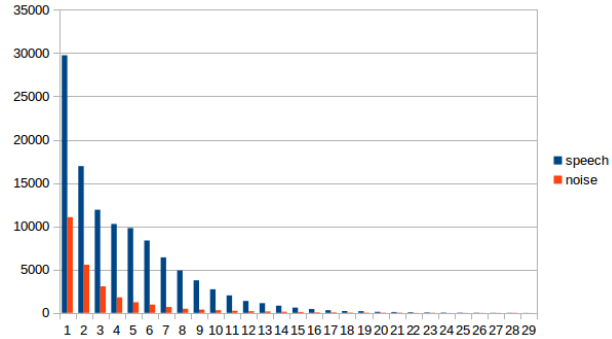

Fig. 7. Chunk Counts for phone /n/

Fig 7 plots noise and speech for the nasal phone /n/. It is clear from the plot that the noise and speech chunk overlaps across the whole chunk sizes. It's difficult to

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

discriminate between noise and speech, let alone do the confidence scoring for the speech chunks. This makes it very difficult to robustly analyze the recognized /n/ frames.

VII. CONCLUSION AND FUTURE WORK

In this paper, a method of scoring confidence from the duration of phone chunks recognized by a frame classifier, is discussed. A forced aligner is used as a tool to find measures suggesting the correspondence between the frame classifier and the forced aligner. Noise data is also used to cut out the noise related statistics coming out of the frame classifier. This method is applicable in calibrating a pronunciation assessment system, using a frame classifier. It also throws light into some plausible errors in the forced alignment.

Count of noise and speech for various phones are given, which helps in the development of a set of assumptions, which holds in real-world speech recognition systems. The reason for the assumptions plays a vital role in figuring out the type of confidence scoring mechanism to be used. A simple confidence scoring mechanism for phones, which takes into account the chunk counts for various chunk size is also described. A case where this approach fails is also discussed.

REFERENCES

[1] R. Dong and J. Zhu, "On use of duration modeling for continuous digits speech recognition"

[2] N. Sridhar Krishna, Partha Pratim Talukdar, Kalika Bali, A.G. Ramakrishnan, Duration Modeling for Hindi Tex t-to-Speech Synthesis System",

[3] Hossein Hadian, Daniel Povey, Hossein Sameti, Sanjeev Khudanpur, Phone duration modeling for LVCSR using neural networks"

[4] Tanel Alumae, "Neural Network Phone Duration Model for Speech Recognition",

[5] Dino Seppi, Daniele Falavigna, Georg Stemmer, Roberto Gretter, "Word Duration Modeling for Word Graph Rescoring in LVCSR ",