Neural Network Phone Duration Model for Speech Recognition

Tanel Alumäe

Institute of Cybernetics
Tallinn University of Technology, Estonia
tanel.alumae@phon.ioc.ee

Abstract

In this paper, we describe a novel phone duration model that is used to improve the accuracy of a large vocabulary speech recognition system based on state-of-the-art speaker-adapted DNN acoustic models. The duration model calculates the probability density function of phone duration from phone's contextual features using a neural network which is then applied for word lattice rescoring. Experimental results are given for Estonian, English and Finnish transcription tasks. An absolute word error rate reduction of 0.8-1.4% is observed across all evaluation sets.

Index Terms: duration model, neural network, speech recognition

1. Introduction

Durational patterns of phonetic events carry important information about the linguistic content of an utterance. However, most speech recognition systems rely only on hidden Markov models (HMMs) in modeling the variability of phone durations. The weaknesses and limitations of HMMs in utilizing the knowledge provided by durational cues are well known. HMMs model phone durations using state transition probabilities which result in geometric probability density functions (PDFs) for phone durations, which inappropriately represent the temporal behavior of speech, as phone durations are known to follow a gamma or log-normal distribution. Several methods have been proposed to resolve this limitation. Improved duration modeling can be integrated directly into the HMM framework, by replacing the HMM state transition probabilities with explicit duration PDFs [1] or by modifying the HMM topology [2]. However, these approaches significantly increase the computational complexity of decoding and offer no or only limited decrease in word error rates. Therefore, an alternative approach is to model word or phone durations using an independent model and use it as a separate knowledge source during N-best rescoring [3, 4] or lattice rescoring [5, 6, 7]. Such approaches usually result in a small (usually about 5% relative) word error rate reduction in large vocabulary continuous speech recognition (LVCSR) tasks.

Recently, we proposed a new method for modeling phone durations for speech recognition [8]. The model was based on a decision tree that finds clusters of phones in various contexts that have similar durations. Wide contexts with rich linguistic and phonetic features were used. To better model varying and non-stationary speaking rates, the contextual features also included the observed duration values of previous phones. For each resulting phone cluster, a log-normal duration PDF was estimated. The resulting decision tree and the log-normal PDFs were used for N-best rescoring. Experiments on two Estonian recognition tasks showed a small but significant improvement in speech recognition accuracy. In this paper, we improve the previous model by replacing its decision tree based binned probability model with a neural network that computes the parameters (mean and standard deviation) of the log-normal duration PDF from the phone's contextual features. We also replace N-best rescoring with more-efficient and more accurate lattice rescoring. Speaker adapted deep neural network (DNN) based acoustic models are used as the baseline acoustic models, resulting in baseline word error rates that are relatively 40% lower than in our previous work where speaker independent Gaussian mixture models (GMMs) were used. Experiments are performed on two Estonian, one English and one Finnish recognition task.

Previously, neural networks have been used for determining the duration of phonetic segments in speech synthesis. For example, [9] describes a system where inputs to the neural network correspond to phoneme label, articulatory features, syllable and word prominence, and proximity to syntactic boundaries. The network is trained to generate the logarithm of phone duration. Our model is similar and partly inspired by such model, however, instead of determining the phone duration, it determines the PDF of the duration.

The following section describes the duration modeling approach used in this paper. Section 3 presents evaluation setup and experimental results. Section 4 concludes the paper.

2. Neural network based duration model

With explicit duration models, the speech recognition problem is usually reformulated as a task of finding the best word sequence $W^\star$ and the corresponding word durations $D^\star$, given the acoustic signal. Assuming that, given the word sequence W, acoustic features A can be viewed as conditionally independent from word durations D, we can write:

$$\begin{aligned} W^\star, D^\star &= \underset{W,D}{\operatorname{argmax}} P(W, D|A) \\ &= \underset{W,D}{\operatorname{argmax}} P(A, D|W)P(W) \\ &= \underset{W,D}{\operatorname{argmax}} P(A|W)P(D|W)P(W) \end{aligned}$$

The third line in the last equation relies on the (invalid) assumption that A is independent from D, once conditioned on W. Therefore, the task of the explicit duration model is to estimate the likelihood $P(D|W)$.

We decompose the duration structure D into m phone durations $d_i^{(p)}$:

$$P(D|W) = P(d_1^{(p)}, ...d_m^{(p)}|W)$$

The likelihoods of phone durations can be decomposed using

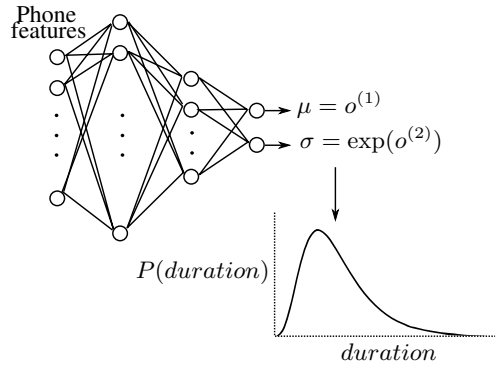

Figure 1: Architecture of a deep conditional density estimation network. Input features describe the current phoneme, phonetic context of the current phone, and the durations of the previous phones. Output is used for calculating the log-normal duration density function of the current phone.

the chain rule and approximated using the Markov assumption:

$$P(d_1^{(p)}, ...d_m^{(p)}|W) = \prod_{i=1}^{m} P(d_i^{(p)}|d_1^{(p)}, .., d_{i-1}^{(p)}, W)$$
$$\approx \prod_{i=1}^{m} P(d_i^{(p)}|d_{i-n-1}^{(p)}, .., d_{i-1}^{(p)}, W)$$

That is, we approximate the likelihoods of phone durations by conditioning on the durations of the previous n phones and the hypothesized words.

We use binary features to encode the dependence of the i-th phone duration $d_i^{(p)}$ on the word sequence W. The features describe the type and properties of the phone, position of the phone in relation to the word and utterance, and features of the neighboring phones. To condition the duration probability on the previous durations, we use the duration values of previous phones simply as discrete-valued features (representing the number of feature frames or milliseconds). A detailed look at the features used in practice is given in section 3.

Previously proposed duration models often take special care to handle two phenomena that have a big impact on phone duration: speech tempo and pre-pausal lengthening. Our method handles both of those effects intrinsically: (local) speaking rate can be derived from the duration values of the previous K phones that are encoded in our model as features. Pre-pausal context can be also represented using features, as the feature vector of a pre-pausal phone includes entries such as $pos_{i+1} =$ `SIL`, and $pos_{i+2} =$ `</s>` if the pause is followed by an utterance end.

Our task is now to construct a model for estimating phone duration likelihoods, based on a set of binary and discrete features $x_i^{(p)}$ for that phone. In other words, we want to build a model for estimating $P(d_i^{(p)}|x_i^{(p)})$ for unseen $(x_i^{(p)}, d_i^{(p)})$. We propose to implement the model using a neural network. A neural network that estimates the parameters of a specified probability distribution is referred to as a conditional density estimation network (CDEN) [10], [11, 4.7]. In the CDEN framework, neural network outputs are associated with parameters of a user-specified PDF. A CDEN for a log-normal PDF would thus have two outputs: one for the conditional mean μ and second for the conditional standard deviation σ. A CDEN with two hidden layers is shown in Figure 1. Because the outputs of the CDEN correspond to the parameters of the log-normal distribution, exponent function is applied to the output corresponding to the standard deviation to make it strictly larger that zero.

The model can be trained using back-propagation, with negative log-likelihood of the data as the cost function. When using log-normal distribution as the target distribution, this translates to summing over the logarithm of the log-normal PDF $f(d; \mu, \sigma)$ applied to training data:

$$cost = -\log(\mathcal{L}) = -\sum_{t=1}^{N} \log f(d_t^{(p)}; \mu_t, \sigma_t)$$
$$= -\sum_{t=1}^{N} \log \left(\frac{1}{d_t^{(p)} \sigma_t \sqrt{2\pi}} \exp(-\frac{(\log d_t^{(p)} - \mu_t)^2}{2\sigma_t^2}) \right)$$
$$= \sum_{t=1}^{N} \frac{(\log d_t^{(p)} - o_t^{(1)})^2}{2\exp(o_t^{(2)})^2} + \log \left(d_t \exp(o_t^{(2)}) \sqrt{2\pi} \right)$$

where N is the number of phones in the training data, $d_t^{(p)}$ is the phone's duration, and $o_t^{(1)}$ and $o_t^{(2)}$ are the outputs of the neural network, given the phone's features.

The neural network is trained using the features and duration values obtained from a large speech corpus. Given a training corpus with orthographic transcriptions, duration values are obtained after forced alignment between the phone transcriptions (generated from the orthographic transcriptions and the pronunciation dictionary) and the speech signal. We used the same acoustic model as was used during decoding for performing the forced alignment.

CDEN models are known to be prone to overfitting, if the number of training observations is small and the model is complex enough. With excess model capacity, optimization of the neural network may result in a conditional PDF that collapses onto the training sample, thus leading to an infinite value of the likelihood. However, in our case, the number of training observations is equal to the number of phones in the aligned speech data, which is quite large even for a moderately sized speech corpus, and we have found overfitting not to be a problem (max-norm regularization was used however).

3. Experiments

3.1. Training data

We tested our method on Estonian, English and Finnish speech.

The details of our Estonian system are described in [12]. For training the duration models (DMs) and acoustic models (AMs), we used various wideband Estonian speech corpora, about 140 hours in total: dictated speech (9h), broadcast news (36 h), broadcast conversations (38 h), conference speeches and university lectures (40 h) and studio-recorded spontaneous monologues and dialogues (16 h).

For tuning and measuring system performance, two different domains were used: broadcast news (BN) and broadcast conversations (BC), with separate development and evaluation sets for both of the domains. The development and evaluation sets for BN consist of 30 short radio news programs with mainly dictated speech, both around two hours in total. The development and evaluation sets for the BC domain contain spontaneous TV and radio interviews, both around one hour in total.

For English experiments, we used the VoxForge[1] open source speech corpus. The corpus contains dictated speech recorded and submitted by volunteers. The VoxForge English corpus downloaded on September 27, 2013 was used. Speakers with American, British, Australian and New Zealand pronunciation dialects were selected for further experiments. Out of those, two random 20 speaker sets were extracted to be used as a development and evaluation set. The resulting training set contained 1207 unique speakers, with about 70 hours of speech signal in total. The development and test sets both contained about 40 minutes of audio.

The Finnish experiment was based on the Finnish Speecon database [13], from which 31 hours of clean dictated wideband speech from 310 speakers was used for training. Development and evaluation sets both contain around 1.9 hours of speech from 40 distinct speakers.

Estonian phonemic inventory includes 43 phonemes, including short and long variants of all phonemes except /j/. English phonemic inventory has 39 items. Finnish inventory contains 43 phonemes, including short and long variants for most of the phonemes. Position dependent (word beginning, word ending, intra-word and single phone) models were used for all languages. To avoid the mismatch of position dependent phones between training and decoding conditions in the Estonian and Finnish experiments, words in the training data were segmented into shorter units, using the same method as for language model training data.

3.2. Duration model architecture and features

We experimented with a few different neural network architectures, by varying the number of hidden layers, the number of units in the hidden layers, hidden layer activation functions and regularization methods. The architecture that worked best and was used in the reported experiments has two hidden layers. The first layer uses a rectified linear activation function, with the number of units equal to 1.5 times the number of features. The second layer uses the *maxout* activation function [14], with the number of units equal to 0.75 times the number of input features. We regularized the model by imposing a constraint on the norm of each hidden layer weight vectors, applied after every training batch.

Input vector to the neural network contains mostly binary entries, except for the entries corresponding to the syllable number of the current and neighbouring phones and the duration values of the previous K phones. The duration features were normalized using a sigmoid-like function

$$d' = \frac{2}{1 + e^{-0.01d}} - 1$$

where d is the original duration value in milliseconds. The normalization function maps the unbounded original duration values non-linearly to $(0, 1)$. The function grows almost linearly in the $d = 0..200$ range and saturates at larger input values. This allows to normalize duration values independently of the dataset. We found that such normalization greatly accelerates the convergence of the neural network, especially when the rectified linear activation function is used in the first hidden layer.

We used the the following features in our duration model (all features, except for those marked otherwise, are binary):

- Current and $\pm K$ neighbouring phoneme labels;
- Current and neighbouring phoneme types, e.g., vowel, nasal, fricative, etc;
- Is the current/neighbouring phone first or last in the word (morph for Finnish) or utterance?
- Is the current/neighbouring phone phonetically long (Estonian and Finnish only)?
- Does the current/neighbouring phone carry primary or secondary stress (English only)?
- Position of the corresponding syllable in the word, for current/neighbouring phone (ordinal, omitted for Finnish, as Finnish use morphs as basic units);
- Duration of the previous K phones (ordinal, real after normalization).

As with language models, duration model performance can be evaluated based on how well it predicts unseen data. We compared duration models with increasing context size using development set perplexity, defined as:

$$\texttt{PPL} = \exp\left(-\sum_{i=1}^{N} \frac{1}{N} \log \hat{p}(d_i|x_i)\right)$$

Table 1: Perplexity of the duration model of increasing context sizes (K), with (*+DF*) and without (*-DF*) using the durations of previous K phones as additional features.

	Estonian		English		Finnish	
Context	-DF	+DF	-DF	+DF	-DF	+DF
± 0	10.8		14.1		10.5	
± 1	8.5	7.8	10.5	10.0	8.2	7.9
± 2	8.1	7.3	10.0	9.4	7.8	7.5
± 3	7.9	7.1	10.0	9.4	7.7	7.4

Table 1 shows how the perplexity of the model decreases when features from a wider context are used. We compared models with and without using the durations of previous K phones as inputs to the model. The table shows that the duration features are highly important: a model with a three phone context window (± 1 phones) that uses the duration feature achieves about the same performance as a model using a context window of 7 phones ($\pm$ 3) and no duration features. In the speech recognition experiments, duration features were thus used in all models.

3.3. Speech recognition experiments

We used the Kaldi toolkit [15] for training acoustic models and generating the baseline recognition lattices. Estonian and English acoustic models were trained using a similar strategy which is based on the Kaldi cross-entropy trained DNN training recipe, documented in [16, 3.2]. Acoustic feature vectors for the GMM-HMMs were obtained by splicing together 7 frames of 13-dimensional MFCCs and projecting them down to 40 dimensions using LDA. Speaker-based cepstral mean normalization was applied to the MFCCs. Speaker adaptive training (SAT) was done by estimating a fMLLR transform for each speaker. In the DNN-HMM hybrid system, the DNN was trained to provide posterior probability estimates for the HMM states. The DNNs have four hidden layers. The Estonian DNN has 1367 neurons in each hidden layer and 3166 output units, and the English DNN 1051 and 1560 units, respectively. For the DNNs, a nine frame context window (4 frames at each side) is used at input, followed by a second LDA transform that keeps 250 dimensions out of the 360. DNNs were trained using cross-entropy cost and stochastic gradient descent.

For Finnish, we used GMM-HMMs as the main models, trained using SAT and boosted MMI [17]. The models contain 25 0000 Gaussians in 2000 mixtures.

[1] http://www.voxforge.org

Table 2: Word error rates for two Estonian, one English and one Finnish transcription task(s) before and after rescoring with duration models of increasing context size. Average relative improvement over the baseline across all evaluation sets is given in the last column.

	Estonian				English		Finnish		
	BC		BN		VoxForge		Speecon		Average **relative**
Duration model context	Dev	Eval	Dev	Eval	Dev	Eval	Dev	Eval	improvement (%)
No duration model	27.6	17.9	9.6	9.0	29.3	25.5	12.6	14.4	
± 0	26.5	17.7	9.4	8.6	28.4	25.4	12.3	14.3	2.0
± 1	25.9	17.0	9.2	8.3	27.2	25.0	11.7	14.0	5.2
± 2	25.3	16.5	9.1	8.2	26.9	24.4	11.4	13.3	7.2
± 3	25.2	16.5	9.1	8.2	26.6	24.2	11.5	13.4	7.2

Language model (LM) training data for the Estonian system contains around 300M words from various sources (newspapers, popular magazines, web, parliament transcripts, fiction, manual transcriptions of broadcast news, broadcast conversations and lectures). As Estonian is a heavily compounding and inflective language, the lexical variety of the language is very high. To reduce the out-of-vocabulary (OOV) rate of the LM, compound words were decomposed into compound segments, using a morphological analyzer [18]. After decoding, compound words were reconstructed from the recognized tokens using a hidden event LM, as described in [19]. We used a vocabulary size of 200K units. For both of the domains used in the experiments, we created an optimized 4-gram LM by interpolating source-specific LMs estimated using interpolated modified Kneser-Ney discounting. The LM to be used in decoding was heavily pruned, and a larger LM was created for the final LM rescoring pass. A rule-based system was used for deriving the pronunciations for words in the LM lexicon.

Language model for the English system was estimated from the VoxForge transcripts that were not present in the development or test set, resulting in a corpus of 171K words. A Kneser-Ney smoothed trigram LM with a vocabulary of 13151 words was built. The CMU pronunciation dictionary was used, supplemented by a G2P system for words not present in lexicon.

A morph-based language model was used in the Finnish recognition experiment, trained on 150M words from the Kielipankki[2] corpus which contains texts from books, magazines and newspapers. The words in the texts were decomposed into shorter data-driven units using the Morfessor software [20]. A lexicon of 46K morphs was used. After segmentation, the text corpus contained 233M morphs. A 4-gram LM was estimated, using modified Kneser-Ney discounting. For speech recognition experiments, the LM was pruned to about 1/10 in size. Words were reconstructed from the recognized morphs using a separate 4-gram LM that models word break as a hidden token. A grapheme-based pronunciation lexicon was used.

For the Estonian experiment, the transcription system first performed speech/non-speech detection and speaker diarization. English and Finnish data was already segmented into utterances and grouped by speakers. We performed three decoding rounds: using a speaker-independent GMM-HMM system, using a SAT-trained GMM-HMM system and finally using the DNN-HMM system (the final phase is omitted for Finnish). Between the passes, fMLLR transforms were re-calculated for all speakers. The Estonian system also performed a final LM rescoring pass using a larger LM than was used for decoding.

In the final decoding pass, we generated word lattices that contain phone segmentation information for each word edge. We then used duration model to add log duration scores for each word edge in the lattice, by summing the log duration likelihoods of the word's phonemes. An additional constant equal to the number of phones in the word was added to each word edge, to be used as a phone insertion penalty. For edges representing silence or noise, the duration and phone penalty scores were set to zero. Finally, duration model score, phone penalty, LM score and AM score were combined, using weights that were optimized on the corresponding development sets using random search. Finally, the rescored lattices were decoded.

Table 2 lists word error rates (WER) before and after rescoring the lattices with DNN duration models of different phone context size. Due to the well optimized DNN-based acoustic model, the baseline system was 35-45% relatively more accurate than in our previous experiment [8], where the WER for the Estonian BC and BN evaluation sets was 32.4 and 13.8, respectively. However, rescoring the lattices with the duration model decreased WER relatively more than in the previous study. The relative reduction of WER brought by the full duration model was larger for Estonian (7.8% and 8.9% for BC and BN, respectively) and Finnish (7.6% at context size 2) than for English (5.1%). For all four evaluation sets, the improvements were statistically significant based on the Wilcoxon test ($p \leq 0.05$).

The larger improvements in WER and the smaller DM perplexities in Table 1 for Estonian and Finnish might be explained by the opposition of short and long phonemes in the Estonian and Finnish phonemic inventory, whereas there is no such opposition in English. Therefore, as duration in Estonian and Finnish carries more discriminative linguistic information than in English, the durations can be expected to be more stable, which increases the usefulness of a dedicated duration model.

4. Conclusion

This paper presented a neural network based phone duration model for speech recognition. The model is trained to estimate a phone duration PDF based on its neighbouring phones and the already observed durations of the previous phones in the context window. The duration model is used for rescoring lattices that include phone segmentations for each word edge. Experimental results on Estonian, English and Finnish recognition tasks show that the method results in a consistent drop in WER, even when using carefully optimized speaker-adapted DNN-based acoustic models as baseline. A 5.1-8.9% relative improvement is observed across four different test sets.

5. Acknowledgements

This research was supported by the European Union through the European Regional Development Fund and by the Estonian Ministry of Education and Research target-financed research theme No. 0140007s12.

[2] http://www.csc.fi/kielipankki/

6. References

[1] S. E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Computer Speech and Language*, vol. 1, no. 1, pp. 29–45, 1986.

[2] M. J. Russell and R. K. Moore, "Explicit modeling of state occupancy in hidden Markov models for automatic speech recognition," in *ICASSP 1985*, 1985.

[3] A. Anastasakos, R. Schwartz, and H. Shu, "Duration modeling in large vocabulary speech recognition," in *ICASSP 1995*, vol. 1, 1995, pp. 628–631.

[4] V. R. R. Gadde, "Modeling word duration for better speech recognition," in *Speech Transcription Workshop*, Maryland, 2000.

[5] D. Povey, "Phone duration modeling for LVCSR," in *ICASSP 2004*, vol. 1, 2004, pp. 829–832.

[6] D. Seppi, D. Falavigna, G. Stemmer, and R. Gretter, "Word duration modeling for word graph rescoring in LVCSR," in *Interspeech 2007*, 2007, pp. 1805–1808.

[7] N. Jennequin and J.-L. Gauvain, "Modeling duration via lattice rescoring," in *ICASSP 2007*, vol. 4, Honolulu, HI, USA, 2007, pp. 641–644.

[8] T. Alumäe and R. Nemoto, "Phone duration modeling using clustering of rich contexts," in *Interspeech 2013*, Lyon, France, 2013.

[9] O. Karaali, G. Corrigan, and I. A. Gerson, "Speech synthesis with neural networks," in *World Congress on Neural Networks*, San Diego, USA, 1996.

[10] R. Neuneier, F. Hergert, W. Finnoff, and D. Ormoneit, "Estimation of conditional densities: A comparison of neural network approaches," in *ICANN 1994*. Berlin: Springer, 1994, pp. 689–692. [Online]. Available: http://dx.doi.org/10.1007/978-1-4471-2097-1_162

[11] W. W. Hsieh, *Machine Learning Methods in the Environmental Sciences: Neural Networks and Kernels*, 1st ed. Cambridge University Press, 2009. [Online]. Available: http://www.ocgy.ubc.ca/hsieh/book/

[12] T. Alumäe, "Recent improvements in Estonian LVCSR," in *SLTU 2014*, Saint Petersburg, Russia, 2014.

[13] D. J. Iskra, B. Grosskopf, K. Marasek, H. van den Heuvel, F. Diehl, and A. Kießling, "Speecon - speech databases for consumer devices: Database specification and validation," in *LREC 2002*, Las Palmas, Spain, 2002.

[14] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," *ArXiv e-prints*, Feb. 2013.

[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE ASRU Workshop*, Dec. 2011.

[16] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Interspeech 2013*, Lyon, France, August 2013.

[17] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *ICASSP 2008*, 2008, pp. 4057–4060.

[18] H.-J. Kaalep and T. Vaino, "Complete morphological analysis in the linguist's toolbox," in *Congressus Nonus Internationalis Fenno-Ugristarum Pars V*, Tartu, Estonia, 2001, pp. 9–16.

[19] T. Alumäe, "Automatic compound word reconstruction for speech recognition of compounding languages," in *NODALIDA 2007*, Tartu, Estonia, 2007, pp. 5–12.

[20] M. Creutz, K. Lagus, and S. Virpioja, "Unsupervised morphology induction using Morfessor," in *FSMNLP 2005*, Helsinki, Finland, 2005, pp. 300–301.