DISCRIMINATIVELY TRAINED CONTEXT-DEPENDENT DURATION-BIGRAM MODELS FOR KOREAN DIGIT RECOGNITION

Daniel Willett, Franz Gerl, Raymond Brueckner

Harman/Becker Automotive Systems, Speech Dialog Systems, Ulm, Germany

dwillett@harmanbecker.com

ABSTRACT

The recognition of continuously spoken Korean digits is well known to be a particularly challenging task among small vocabulary recognition problems. In this paper, we review and evaluate our acoustic modeling efforts for the purpose of efficient high-accuracy recognition of Korean digit strings for in-car applications. The measures comprise context-dependent word models, duration-dependent distribution functions, error-weighted discriminative training as well as a compressed bigram model that strongly constrains the HMM state durations. Finally, an average word error rate across multiple channel and noise conditions of below 3% is achieved, which is a relative reduction of 62% over the error observed with traditional context-independent digit modeling techniques and about 36% relative error reduction compared to ML-trained context-dependent digit models of ordinary linear topology. Fast unsupervised model adaptation during decoding yields additional 10% of relative improvement.

1. INTRODUCTION

The recognition of continuously spoken Sino-Korean[1] digits is a very difficult task for a couple of reasons. The first is that Sino-Korean digits are monosyllabic, two of them even mono-phonemic. Thus, they are very short. Moreover, a couple of digits are very similar to each other and some are phonetic substrings of another digit. Table 1 gives an idea. Furthermore, when uttered in continuous strings, Korean digits are strongly influenced by coarticulation.

1	2	3	4	5	6	7	8	9	0
il	i	sam	sa	o	yuk lyuk	chil	pal	ku	kong yeong

Table 1. The 12 Sino-Korean digit words

For these reasons, it is particularly challenging to achieve acceptable system performance for Korean digit recognition tasks. State-of-the-art techniques such as whole word digit models and MFCC front-ends that potentiate word error rates (WERs) below and far below 5% on other languages with adequate training data, end up at WERs of around 10% for Korean digits [8, 7]. In this regard, Korean digit modeling is a particularly well suited area for the evaluation and the application of advanced modeling and parameter estimation techniques. A couple of our techniques are outlined in the following paragraphs and evaluated jointly in a Korean digit recognition scenario.

[1]The Korean language actually has two ways of speaking digits, the Korean and the Sino-Korean digit words. The original purely Korean digit words (hana, tul, set, ...) are rarely used and do not pose a particularly hard recognition problem. In this text, the term Korean digits refers to the by far more frequently used Sino-Korean digits.

2. CONTEXT-DEPENDENT DIGIT MODELING

The modeling of coarticulation effects between digits has been addressed in multiple studies [1, 3, 4, 16, 2]. Besides the straightforward application of cross-word context-dependent phonetic models [2, 15], which might be word-specific or not, so-called Head-Body-Tail models have been proposed which split digits into three parts independent of the number of phonetic units [1, 3, 16]. In this study, we make use of a different kind of model structure, which we will refer to as Final-Initial models. 12 distinct models $\texttt{B}n$ ($n = 1, \ldots 12$) model the initial part of each of the 12 digit versions at the beginning of utterances or after pauses and another 12 models $n\texttt{E}$ do model the final part of digits at the end of utterances or before pauses. Besides that, 12^2 models nm ($n = 1 \ldots 12, m = 1 \ldots 12$) are set up for the final part of each digit in coarticulation with the initial part of each digit. For example, the digit sequence "123" is modeled by four models concatenated to the sequence `B1 12 23 3E`. Prior to bigram duration modeling all digit models have linear topology with the internal Final-Initial models nm having 9 states and the start models $\texttt{B}n$ and end models $n\texttt{E}$ having 5 states each. We do not claim that this type of context-dependent digit modeling is superior to other approaches. In the end, the best setup largely depends on the amount of available training data. The Final-Initial models make particularly little usage of parameter tying across models, which in the other approaches is achieved by sharing the body-model across context-dependent versions of the same digit or by tying phonetic units of similar contexts. In this respect, our approach requires large amounts of training data and is capable of estimating very accurate models on such large amounts. We chose this kind of model topology for this study, as it is particularly useful when it comes to bigram-duration modeling. For reasons of computational complexity the bigram-duration modeling is only applied model internally and therefore benefits from single models covering longer periods of the speech signal.

Also, it should be noted that having dedicated models for the final digit parts before pauses and at the end of utterances realizes an explicit modeling of the major part of final digits in digit sequences. It has been found that final digits are often uttered somewhat slower with more determination and clarity [10]. The independent end models learn this circumstance and they are capable of learning and representing the final digits' very own output distributions and durational statistics, especially so with the bigram duration model.

3. DURATION-BIGRAM MODELS AND DURATION-DEPENDENT OUTPUT DISTRIBUTIONS

We have proposed a context-dependent duration model and discussed its efficient integration into a first-order hidden Markov model-based speech recognizer in [18]. The principal approach is the introduction of a durational bigram model that conditions the du-

rational probability of an HMM state on the duration of the previous state. This higher order model can be implemented in a first-order model-based speech recognizer by expanding each HMM state into an automaton of distinct rows of sub-states that represent the possible state duration, as depicted in Fig. 1.

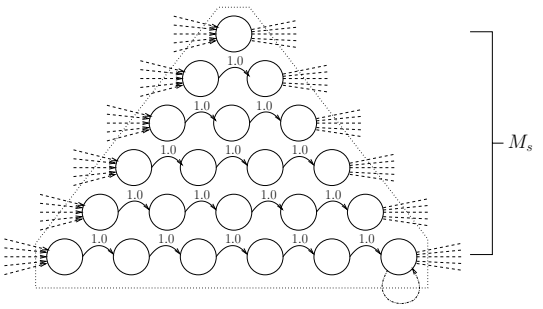

Fig. 1. Context-dependent ESHMM (CDESHMM) superstate

With each model state expanded into this kind of substate structure, the transition probabilities between expanded states automatically represent durational bigram probabilities. They can be estimated using forward-backward training and applied in ordinary Viterbi decoders. As indicated in Fig. 2, the transitions between superstates can be restricted to $reasonable$ transitions in order to prevent particularly unusual duration relations, such as rushing through one state after having remained in the previous state for a long time. In [18] we found imposing this kind of reasonable transition structure to be beneficial in terms of recognition performance. Similarly, it turned out to be advantageous to fully refrain from self-loopable states in this new model topology. Thus, in this current study, none of the states of the expanded state topology allows a self-transition. This way the valid paths through the expanded HMMs are highly constrained. In fact, having no self-looped states signifies that the time-warping ability of standard HMMs is fully replaced by the bigram-model based new topology.

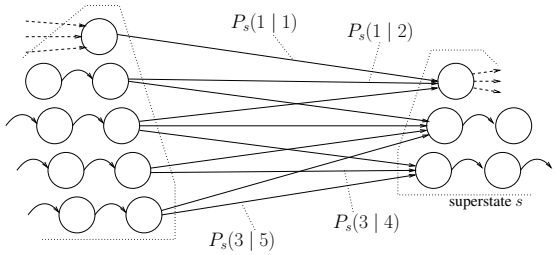

Fig. 2. CDESHMM superstate transitions (reduced connectivity)

In principle, all substates of an expanded state are assumed to share a single distribution function, so that the expanded topology purely realizes a more refined duration model. With each state duration represented in a distinct row of substates, however, the expanded state representation allows the straightforward incorporation of duration-dependent distribution functions. The left hand side of Fig. 3 illustrates an expanded state with the relaxed substate tying indicated through different shadings.

As a final measure of model compression, we here also introduce and apply a merging of those substate rows which on the one hand share the same distribution function and on the other hand have similar entry probabilities. These substate rows can be merged with a neglectable loss in (bigram-) model resolution. The procedure is

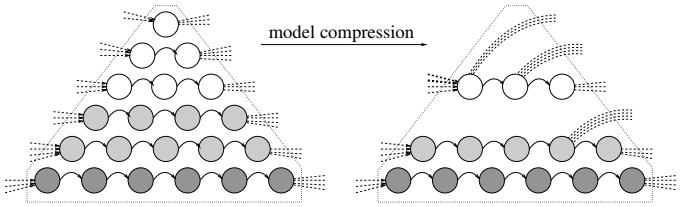

Fig. 3. Relaxed substate tying and model compression

illustrated in Fig. 3. The experimental section will demonstrate that the resulting vast reduction of the number of states comes along with almost no additional error.

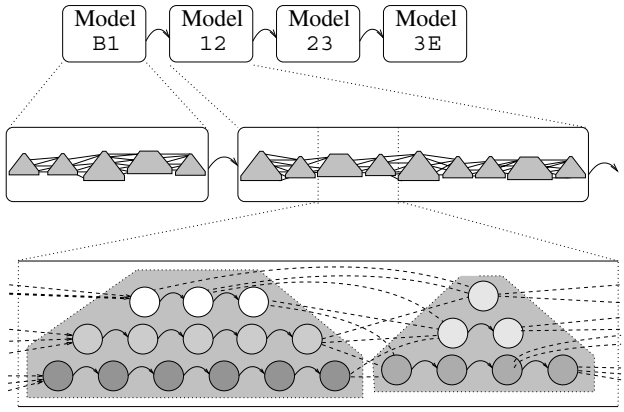

Fig. 4. Model topology representing the digit sequence "123"

Fig. 4 finally illustrates the resulting concatenated model that represents the digit sequence "123".

4. ERROR-WEIGHTED DISCRIMINATIVE TRAINING

Most commonly, the parameter set ϕ of HMM-based speech recognition systems is estimated according to the Maximum Likelihood (ML) objective function. Given U utterances with acoustic observation vectors $X_1 ... X_U$ and transcriptions $W_1 ... W_U$, this yields

$$\phi_{\mathbf{ML}} = \underset{\phi \in \Phi}{\operatorname{argmax}} \prod_{u=1}^{U} p_\phi(X_u \mid W_u) \qquad (1)$$

with Φ representing all possible parameter settings. The Baum-Welch algorithm is a cheap and reliable optimization method for this objective. Nonetheless, whenever the model is unable to learn the true distribution of the data parameter estimates gained from more discriminative objectives outperform Maximum Likelihood (ML) parameter estimates in terms of recognition accuracy. This has been confirmed in numerous studies [12, 17, 20, 11, 14, 19]. However, the optimization of discriminative objectives is far more expensive and much more difficult to apply. Even in the Extended Baum-Welch training as formulated in [17, 14] some parameters are left for manual tuning and convergence cannot be guaranteed.

Here, we review and apply an approach called ML-preferred Error-Weighted MMI-Training (MLpEWMMI) that was originally proposed in [19]. In analogy to an approach in [5] of Error-Weighted Maximum Likelihood training, it augments the Maximum Mutual Information criterion of discriminative training with tunable weights

that control the influence of each training utterance on the training objective. This way it was found that it is possible to obtain parameter estimates that come along with little error on the training data and which generalize better than ordinary MMI estimates on unseen data due to their stronger similarity to ML estimates.

The objective function of Maximum Mutual Information (MMI) training is often formulated as

$$\phi_{\mathbf{MMI}} = \underset{\phi \in \Phi}{\operatorname{argmax}} \prod_{u=1}^{U} \frac{p_\phi(X_u \mid W_u)}{p_\phi(X_u)} \tag{2}$$

in which $p_\phi(X_u)$ represents the overall likelihood of the acoustic observation in the model. The Soft Error-Weighted MMI (softEWMMI) objective introduces an additional weighting factor α_u for each utterance u weighting both numerator and denominator.

$$\phi_{\mathbf{softEWMMI}} = \underset{\phi \in \Phi}{\operatorname{argmax}} \prod_{u=1}^{U} \frac{p_\phi(X_u \mid W_u)^{\alpha_u+1}}{p_\phi(X_u)^{\alpha_u}} \tag{3}$$

With α_u initialized with 0.0 for all utterances u, the training objective at first resembles the ML criterion. After each EM iteration, however, the weighting factors of misrecognized utterances are raised and those of correctly recognized utterances are lowered in a step-size adjusting scheme inspired by RProp [13]. This way, the denominator term is only introduced for utterances which get misrecognized and the overall influence of an utterance on the parameter estimates is slowly increased until the utterance is being recognized correctly (with some confidence) or until a maximum threshold α_{max} has been reached.

The final MLpEWMMI objective introduces additional weighting factors β_u in order to allow for a different weighting of numerators and denominators.

$$\phi_{\mathbf{MLpEWMMI}} = \underset{\phi \in \Phi}{\operatorname{argmax}} \prod_{u=1}^{U} \frac{p_\phi(X_u \mid W_u)^{\alpha_u+1}}{p_\phi(X_u)^{\beta_u}} \tag{4}$$

With both α_u and β_u initialized with 0.0 for all utterances, the training objective at first again resembles the ML criterion. Now, in the beginning, only the numerator weights α_u are increased. Only once an α_u has reached the threshold maximum value without the misrecognition of the respective utterance u being corrected, α_u is retained at that value and the procedure starts introducing u's denominator term by increasing its β_u. Doing so, parameter estimates gained by MLpEWMMI optimization even stronger resemble Maximum Likelihood estimates and are believed to yield even better generalization on unseen data. In [19], the objectives softEWMMI and MLpEWMMI clearly outperformed the common MMI criterion in terms of recognition accuracy of the resulting models on unseen test data. However, the performance difference among the two was hardly measurable.

5. EXPERIMENTS

The experiments were performed using the SiTec Databases Car01, Car02 and Car03 of Korean speech recorded in stationary and moving vehicles [9]. The digit string utterances of the 800 speakers of Car02 and Car03 were used for training, the 100 speakers of the Car01 collection were used for testing. In order to achieve a good degree of channel robustness, the data recorded via headset microphone, seat-belt microphone, sun-visor microphone and rear-view mirror microphone were used jointly in training. From Car01, the three channels headset, seat-belt and sun-visor microphone were merged into a single test set. Overall this comprises over 110 hours of digit string data for training and about 4 hours of digit string data for testing. The front-end includes spectral subtraction to eliminate stationary noise and computes 11 MFCC coefficients. These are fed into an LDA transformation before being modeled with a system of semi-continuous Gaussian mixture distributions.

Table 2 compares the baseline performance of context-independent (12 models of 9 states each) versus context-dependent digit models (168 models with 5 and 9 states as described in Section 2), each trained according to ML.

configuration	training data	test data
context-independent	8.00	7.80
context-dependent	5.19	4.57

Table 2. WER [%] of CI vs. CD digit models

Starting from the context-dependent models of linear topology, we introduce the bigram duration model, then the duration-dependent distribution functions as described in [18] and afterwards perform model compression by merging rows of similar entry probabilities. Table 3 lists the number of states and distributions after these processing steps. Finally, with all 168 models expanded into this dedicated topology, with the tying of distribution functions being relaxed and with the the model rows slightly compressed we end up at 168 models with 8229 states in total. These share 1632 distribution functions.

configuration	number of models	number of states	number of pdfs
context-independent	12	108	108
context-dependent	168	1416	1416
+ bigram-duration model	168	13658	1416
+ duration-dependent pdfs	168	13658	1632
+ state compression	168	8229	1632

Table 3. Size of the digit models

Obviously, the increase in distribution functions is very moderate. Due to the training data being spread over a large number (144) of internal Final-Initial digit models, there is too little data per internal model to justify adding independent distributions. In the data-driven procedure that relaxes the state tying dependent on the amount of respective training data, nearly all internal Final-Initial models remain untouched at only a single distribution function.

Table 4 lists the recognition performance after each of the various processing steps and also states the final recognition performance with MLpEWMMI training performed on the state compressed bigram-duration models. The WER of 2.91% is equivalent to a relative reduction of 62% compared to the context-independent digit models and of about 36% compared to the baseline context-dependent ones.

The last row of Table 4 shows the recognition performance when applying a cheap unsupervised adaptation scheme during decoding on top of that. For speaker- and channel-adaptive decoding we utilize a Maximum Likelihood based method that has been optimized for a semi-continuous HMM recognizer running on limited resources [6]. The basic idea is to transform all Gaussian mean vectors using a linear transformation similar to MLLR. In order to estimate this matrix we utilize information collected only in previous utterances. This way, the approach is advantageous whenever decoding multiple utterances from the same speaker or channel in sequence. In order to

meet run-time constraints, we refrain from iterating adaptation and recognition on each utterance in this adaptive decoding approach. Adaptive decoding achieves an additional 10% of WER reduction. This is despite the circumstance that it is an ML-based optimization applied on discriminatively trained models.

configuration	WER [%] on training data	WER [%] on test data	run-time [RTF]
context-independent	8.00	**7.80**	0.069
context-dependent	5.19	**4.57**	0.081
+ bigram-duration model	4.51	**3.91**	0.108
+ duration dependent pdfs	4.22	**3.65**	0.109
+ state compression	4.21	**3.64**	0.097
+ discr. training (MLpEWMMI)	2.86	**2.91**	0.099
+ unsupervised adaptation	2.34	**2.67**	0.110

Table 4. Word error rates and run-time [Real-Time Factor] after the various processing steps

The right hand column of Table 4 lists the run-times in the different model setups for processing the 24000 test utterances using a beam-search Viterbi decoder on a 3.0GHz (Xeon) workstation. The Real-Time Factor (RTF) of 0.069 for the context-independent acoustic models indicates that decoding of the test data (including all front-end processing) that results in the WER of 7.8% consumes 0.069 times real-time. This means that decoding is 1/0.069 (=14.5) times faster than real-time.

Obviously, run-time increases moderately from RTF 0.069 to 0.081 with the context-dependent digit models compared to the context-independent ones. The additional increase in run-time induced by the introduction of the bigram-duration model and duration-dependent distribution functions is rather moderate as well and, not surprisingly, run-time slightly gains from model compression. The effect of discriminative training on the decoding cost is hardly measurable.

With the computationally cheap implementation of unsupervised adaptation applied on top of that, we end up at a RTF of 0.11, which indicates that this overall setup is well suited for embedded devices and allows real-time response even on machines which are a magnitude slower than the workstation used in these evaluations.

6. CONCLUSION

The paper has outlined various advanced modeling and parameter estimation techniques and evaluated them jointly for the recognition of continuously spoken Korean digits. The application of error-weighted discriminative training, context-dependent digit models, the introduction of a bigram-duration model as well as duration-specific distribution functions and the application of unsupervised speaker- and channel-adaptation resulted in massive word error reductions. Run-time of Viterbi decoding was only very moderately affected by the additional degree of model complexity. The final WER of 2.67% on independent mixed-channel evaluation data with a reasonably small system setup facilitates Korean digit recognition applications on embedded devices that seemed impracticable so far, such as number dialing and postal code-based navigation.

7. REFERENCES

[1] W. Chou, C. H. Lee, B. H. Juang, "Minimum Error Rate Training of Inter-Word Context-Dependent Acoustic Model Units in Speech Recognition", ICSLP'94, Yokohama, Japan, 1994.

[2] D. N. Duc, J.-P. Hosom, L. C. Mai, "HMM/ANN System for Vietnamese Continuous Digit Recognition", IEA/AIE'03, Laughborough, UK, 2003.

[3] M. B. Gandhi, J. Jacob, "Natural Number Recognition Using MCE Trained Inter-Word Context Dependent Acoustic Models" IEEE ICASSP'98, Seattle, ISA, 1998.

[4] J. J. Godfrey, A. Ganapathiraju, C. S. Ramalingam, J. Picone, "Microsegment-Based Connected Digit Recognition", IEEE ICASSP'97, Munich, Germany, 1997.

[5] V. Goel, S. Axelrod, R. Gopinath, P. Olsen, K. Visweswariah, "Discriminative Estimation of Subspace Precision and Mean (SPAM) Models", Eurospeech'03, Geneva, Switzerland, 2003.

[6] U. Haiber et al., "Robust Recognition of Spontaneous Speech", Verbmobil: Foundations of Speech-to-Speech Translation, Springer, 2000, pp. 46–62.

[7] H. B. Jeon, D. K. Kim, "Maximum a Posteriori Eigenvoice Speaker Adaptation for Korean Connected Digit Recognition", IEEE ICASSP'04, Jeju Island, Korea, 2004.

[8] O. W. Kwon, C. K. Un, "Context-Dependent Word Duration Modelling for Korean Connected Digit Recognition", Electronics Letters, Vol.31, No.19, 1995.

[9] Y.-J. Lee, B.-W. Kim, Y.-I. Kim, D.-L. Choi, K.-H. Lee, Y. Um, "Creation and Assessment of Korean Speech and Noise DB in Car Environment", LREC2004, Lisbon, Portugal, 2004.

[10] N. Ma, P. Green, "Context-Dependent Word Duration Modelling for Robust Speech Recognition", Eurospeech'05, Lisbon, Portugal, 2005.

[11] E. McDermott, S. Katagiri, "Prototype Based Discriminative Training for Various Speech Units", Computer Speech and Language, pp. 8:351–368, 1994.

[12] Y. Normandin, "An Improved MMIE Training Algorithm for Speaker-Independent, Small Vocabulary, Continuous Speech Recognition", IEEE ICASSP'91, Totonto, Canada, 1991.

[13] M. Riedmiller, H. Braun, "A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP algorithm", IEEE International Conference on Neural Networks 1993.

[14] R. Schlüter et al., "Comparison of discriminative training criteria and optimization methods for speech recognition", Speech Communication 34, pp. 287–310, 2001.

[15] A. Sixtus, "Across-Word Phoneme Models for Large Vocabulary Continuous Speech Recognition", Ph.D. Thesis, RWTH Aachen, Germany, 2003.

[16] J. Sturm, E. Sanders, "Modelling Phonetic Context using Head-Body-Tail Models for Connected Digit Recognition", ICSLP'00, Beijing, China, 2000.

[17] V. Valtchev, "Discriminative Methods in HMM-based Speech Recognition", Ph.D. thesis, Cambridge University Engineering Department, UK, 1995.

[18] D. Willett, "Context-Dependent Duration Modeling", IEEE ICASSP'05, Philadelphia, USA, 2005.

[19] D. Willett, "Error-Weighted Discriminative Training for HMM Parameter Estimation", ICSLP'04, Jeju, Korea, 2004.

[20] P. C. Woodland, D. Povey, "Large Scale Discriminative Training for Speech Recognition", ISCA ITRW ASR 2000 Workshop: Challenges for the Millennium, pp. 7–16, Paris, 2000.