Advances in Word based Dialect/Accent Classification

Rongqing Huang, John H.L. Hansen

Robust Speech Processing Group, Center for Spoken Language Research
University of Colorado, Boulder, CO, USA
{huangr,jhlh}@cslr.colorado.edu

10.21437/Interspeech.2005-709

Abstract

In an earlier study, we proposed a very effective dialect/accent classification algorithm, which is named Word based Dialect Classification (WDC). The WDC works well for large size corpora and significantly outperforms traditional Large Vocabulary Continuous Speech Recognition (LVCSR) based systems, which is claimed to be the best performing system for language identification. For a small training corpus, however, it is difficult to obtain a robust statistical model for each word and each dialect. Therefore, a Context Adapted Training (CAT) algorithm is formulated here, which adapts the universal phoneme GMMs to dialect-dependent word HMMs via linear regression. Employing on a 8-dialect British English corpus–IViE, the CAT algorithm trained WDC system obtains a 35.5% relative classification error reduction from the baseline LVCSR system, and a 20.2% relative classification error reduction from the basic WDC system.

1. Introduction

Dialect/Accent is a pattern of pronunciation and/or vocabulary of a language used by the community of *native/non-native* speakers belonging to some geographical region. Some researchers have slightly different definition of dialect and accent. In this study, we use "dialect" and "accent" interchangeably. Our focus is to detect the dialect of an unrestricted (i.e., speaker-independent, transcript unknown) audio utterance from pre-defined set of N dialect classes. Accent detection, or as it is sometimes referred to as accent classification, is an emerging topic of interest in the Automatic Speech Recognition (ASR) community since accent is one of the most important factors next to gender that influence ASR performance [1, 6, 7]. Accent knowledge could be used in various components of the ASR system such as pronunciation modeling [12], lexicon adaptation [17], and acoustic model training [9] and adaptation [3].

In our study, we focus our attention on dialect of English, since application to other languages is straightforward. From an linguistic point of view [15, 16, 18], a word may be the best unit to classify dialects. However, for an automatic speech based classification system, it is impossible to construct statistical models for all possible words from even a small subset of dialects. Fortunately, the words in a language are very unevenly distributed. The 100 most common words account for 40% of the occurrences in the Wall Street Journal (WSJ) corpus [13], which has 20K distinct words, and account for 66% in the SwitchBoard corpus [5], which has 26K distinct words. So, only a small set of words are required for dialect classification based modeling.

In our previous study [8], we have proposed a very effective dialect/accent classification algorithm, which is named as Word based Dialect Classification (WDC). The WDC turns the single text-independent decision problem at the utterance level into a combination of text-dependent decision problems at the word level. The WDC framework also provides options for further modeling and decision space improvement. In [8], for example, we applied AdaBoost and dialect dependency algorithms to boost the word classifier and utterance classifier respectively. WDC works very well for large size corpora and significantly outperforms a Large Vocabulary Continuous Speech Recognition (LVCSR) based system, which is claimed to be the best performing system in language identification [19]. On the 2-dialect WSJ corpus and 4-dialect NATO N4 corpus, the WDC system obtains 1.9% and 1.6% classification error respectively; on the contrary, the LVCSR based system obtains 5.9% and 5.5% classification error respectively. The length of each test utterance is 9 seconds in duration. The relative error reduction is 69% on the average. For a small training corpus, however, it is difficult to obtain a robust statistical model for each word and each dialect. Therefore, a Context Adapted Training (CAT) algorithm is formulated here, which adapts the universal phoneme GMMs to dialect-dependent word HMMs via linear regression. The CAT trained word HMMs are then applied in the regular WDC system for dialect classification.

The remainder of this paper is organized as follows: we briefly review the LVCSR based system and basic WDC algorithm in Sec. 2; the CAT word model training algorithm is formulated in Sec. 3; the experimental results are shown in Sec. 4; finally, conclusions are presented in Sec. 5.

2. Background

2.1. LVCSR based Dialect Classification System

Given audio data and corresponding transcripts, it is straightforward to train acoustic and language models for each dialect. Fig. 1 shows a block digram of the LVCSR based classification system, where N is the number of dialects, and L_i is the likelihood of dialect i. During the test phase, N recognizers are employed in parallel, and the recognizer with the highest likelihood is selected as the dialect class. The LVCSR based system achieves high classification accuracy since it uses knowledge from the phoneme to phoneme sequence through word to word sequence and it has been to be the best performing system in the language identification [19].

2.2. Basic WDC Algorithm

The idea behind WDC is to make a multiple text-dependent decision at the word level instead of making a single text-independent decision at the utterance level. It's clear that a text-dependent decision would be more accurate than a text-independent decision. Also, it's reasonable to apply a weighting strategy on the multiple decision problem as shown in our previ-

This work was supported by US Air Force (FA8750-04-1-0058).

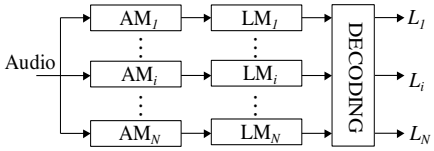

Figure 1: *LVCSR-based Dialect Classification System*

ous study [8]. Fig. 2 shows the block diagram of WDC training. In the training phase, we find all common words across the dialects and train an HMM for each word and dialect. Fig. 3 shows the block diagram of WDC test. During the test phase, a recognizer is applied to generate the word sequence, then each word which has a trained model is used in a word classifier. The final decision at the utterance level is based on voting from all the word classifiers.

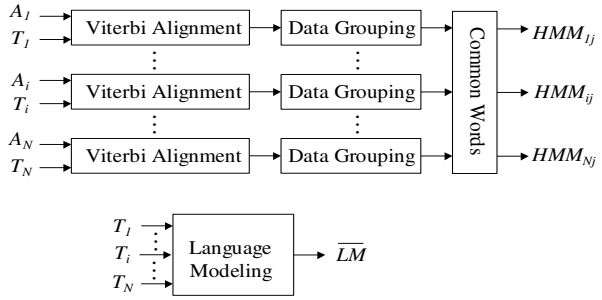

Figure 2: *Block Diagram of WDC Training*

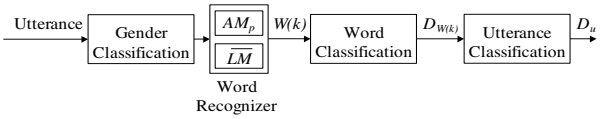

Figure 3: *Block Diagram of WDC Test*

3. Context Adaptive Training (CAT) on Small Data Set

If the size of the training set is small or there are many dialects for a limited data set, it becomes a challenge to train a robust HMM for each word and each dialect, and therefore model adaptation techniques should be applied. In our study [8], we set the number of states in the word HMM equal to the number of phonemes contained in the word. Therefore the word HMMs can be adapted from the phoneme models, which can be trained using data from all the dialects, or even data that is independent of the dialect data set. The proposed adaptation scheme here is motivated by the well established Maximum Likelihood Linear Regression (MLLR) [11] method. To begin with, we define the following notation:

T : the total number of frames for a word in a dialect;

T_w : the total number of training samples for a word in a dialect;

S : the number of states (or phonemes) for a word in a dialect;

M : the number of Gaussian mixtures for each state in the HMM;

$N_{i,j}$: the number of frames for the j^{th} training sample in state i;

$\mathbf{o}_t$: the t^{th} frame vector, where the dimension of the feature is n;

$\mu_{s,m}$, $\Sigma_{s,m}$: the mean vector and the diagonal covariance matrix of the m^{th} Gaussian mixture in the s^{th} state, where $diag(\Sigma_{s,m}) = [\sigma^2_{s,m,1}\ \ \sigma^2_{s,m,2}\ \cdots\ \sigma^2_{s,m,n}]'$;

ω_m : the mixture weight of the m^{th} Gaussian mixture (in state s);

$a_{i,j}$: the transition probability from state i to state j;

π_i : the initial probability of state i;

λ : the parameter set of the HMM ;

$\mathbf{W}$: the $n \times (n+1)$ transformation matrix which must be estimated;

$v, \hat{v}$: v is an original variable, and $\hat{v}$ is the updated/estimated variable of v.

Using this notation, the mean vector is updated through [11] as,

$$\hat{\mu}_{s,m} = \mathbf{W}\xi_{s,m}, \tag{1}$$

where $\xi_{s,m} = [1\ \mu'_{s,m}]'$, and

$$\mathbf{W}'_i = (\mathbf{G}^{(i)})^{-1}\mathbf{Z}'_i, \tag{2}$$

where $\mathbf{W}_i, \mathbf{Z}_i$ are the i^{th} row of $\mathbf{W}$ and $\mathbf{Z}$ ($\mathbf{Z}$ is a $n \times (n+1)$ matrix) respectively, and

$$\mathbf{Z} = \sum_{t=1}^{T}\sum_{s=1}^{S}\sum_{m=1}^{M}\gamma_{s,m}(\mathbf{o}_t)\Sigma^{-1}_{s,m}\mathbf{o}_t\xi'_{s,m}, \tag{3}$$

$$\mathbf{G}^{(i)} = \left[g^{(i)}_{jq}\right]_{(n+1)\times(n+1)}, \tag{4}$$

$$g^{(i)}_{jq} = \sum_{s=1}^{S}\sum_{m=1}^{M} v^{(s,m)}_{ii} d^{(s,m)}_{jq}, \tag{5}$$

$$\mathbf{V}^{(s,m)} = \left[v^{(s,m)}_{ii}\right]_{n\times n} = \sum_{t=1}^{T}\gamma_{s,m}(\mathbf{o}_t)\Sigma^{-1}_{s,m}, \tag{6}$$

$$\mathbf{D}^{(s,m)} = \left[d^{(s,m)}_{jq}\right]_{(n+1)\times(n+1)} = \xi_{s,m}\xi'_{s,m}. \tag{7}$$

Based on previous studies [4], we choose the diagonal covariance matrix to update as follows,

$$\hat{\sigma}^2_{s,m,l} = R_l\sigma^2_{s,m,l},\ l = 1, 2, \ldots, n, \tag{8}$$

where

$$R_l = \frac{\sum_{t=1}^{T}\sum_{s=1}^{S}\sum_{m=1}^{M}\gamma_{s,m}(\mathbf{o}_t)\left(\frac{(o_{t,l}-\hat{\mu}_{s,m,l})^2}{\sigma^2_{s,m,l}}\right)}{\sum_{t=1}^{T}\sum_{s=1}^{S}\sum_{m=1}^{M}\gamma_{s,m}(\mathbf{o}_t)}. \tag{9}$$

The term $\gamma_{s,m}(\mathbf{o}_t)$ denotes the probability of the t^{th} frame being observed in the m^{th} mixture of the s^{th} state of the HMM. In the original formulation of MLLR, this term is computed through the Forward-Backward algorithm, here the Viterbi algorithm is used. It is defined as

$$\gamma_{s,m}(\mathbf{o}_t) = \begin{cases} P(m|\mathbf{o}_t,\lambda) = \frac{\omega_m b_m(\mathbf{o}_t)}{\sum_{k=1}^{M}\omega_k b_k(\mathbf{o}_t)} & if\ S(\mathbf{o}_t) = s \\ 0 & else \end{cases}, \tag{10}$$

where $S(\mathbf{o}_t)$ is the state which generates the frame $\mathbf{o}_t$, $b_m(\mathbf{o}_t)$ is the observation probability of the m^{th} Gaussian mixture (in state s),

$$b_m(\mathbf{o}_t) = (2\pi)^{-n/2} \left(\Pi_{l=1}^{n} \sigma_{m,l}^2\right)^{-1/2} \exp\left[-\frac{1}{2}\sum_{l=1}^{n} \frac{(o_{t,l} - \mu_{m,l})^2}{\sigma_{m,l}^2}\right]. \quad (11)$$

Since the states of the word HMM is the phoneme sequence of the word obtained from a pronunciation dictionary (the CMU 125K American English dictionary [2] is used in our study), the HMM structure should be left-to-right. Also, the one-state-skip structure can address the phoneme deletion (e.g., *farm* pronounced as /F AA R M/ in the CMU dictionary, which is actually pronounced as /F AA M/ in British English). Because it is very hard to obtain a pronunciation dictionary which includes the pronunciation variations of all dialects, a phoneme recognizer may be applied to decode the phoneme sequence in order to capture the phoneme substitution, phoneme deletion and phoneme insertion. Therefore, we define 3 HMM structures in the CAT training: (i) CAT1-a: a no-skip left-to-right structure, where the phoneme sequence is obtained from the CMU pronunciation dictionary; (ii) CAT1-b: a no-skip left-to-right structure, where the phoneme sequence is obtained from the phoneme recognizer; (iii) CAT2: a one-state-skip left-to-right structure, where the phoneme sequence is obtained from the CMU pronunciation dictionary.

The steps employed for CAT training are summarized as follows,

1. Given the audio data and word-level transcripts, find the training samples for the words and phonemes using Viterbi forced alignment.
2. Train the universal gender-dependent and/or gender-independent M mixture based phoneme GMMs using all the dialect data.
3. For each word and each dialect, do:

   (a) Initialize the word HMM. The corresponding S phoneme GMMs are concatenated to form an S state word HMM. The phoneme sequence of the word can be obtained by the pronunciation dictionary or by a phoneme recognizer. The initial state probabilities are set as

   $$\pi_1 = 1,\ \pi_i = 0,\ i = 2, 3, \ldots, S. \quad (12)$$

   If a no-skip left-to-right HMM structure (CAT1-a, CAT1-b) is used, the initial transition probabilities are set as follows,

   $$\begin{aligned} &a_{i,i} = a_{i,i+1} = 0.5,\ if\ i = 1, 2, \ldots, S-1 \\ &a_{i,i} = 1,\ if\ i = S \\ &a_{i,j} = 0,\ if\ i = 1, 2, \ldots, S,\ j \notin \{i, i+1\}. \end{aligned} \quad (13)$$

   If a one-state-skip left-to-right HMM structure (CAT2) is used, the initial transition probabilities are set as follows,

   $$\begin{aligned} &a_{i,i} = a_{i,i+1} = a_{i,i+2} = 1/3,\ if\ i = 1, 2, \ldots, S-2 \\ &a_{i,i} = a_{i,i+1} = 0.5,\ if\ i = S-1 \\ &a_{i,i} = 1,\ if\ i = S \\ &a_{i,j} = 0,\ if\ i = 1, 2, \ldots, S,\ j \notin \{i, i+1, i+2\}. \end{aligned} \quad (14)$$

   (b) Viterbi forced alignment is used to obtain the state and mixture sequences for each training sample.

   (c) Update the HMM parameters as follows:

   i. Use Eq. 1 to update $\mu_{s,m}$

   ii. Use Eq. 8 to update $\Sigma_{s,m}$

   iii. The mixture weights are updated through,

   $$\hat{\omega}_{s,m} = \frac{1}{\sum_{j=1}^{T_w} N_{s,j}} \sum_{j=1}^{T_w} \sum_{k=1}^{N_{s,j}} P(m|\mathbf{o}_{j(k)}, \hat{\lambda}), \quad s = 1, 2, \ldots, S,\ m = 1, 2, \ldots, M. \quad (15)$$

   iv. Here, three alternate methods are used for Context Adaptive Training (CAT1-a, CAT1-b, and CAT2):
   for CAT1-a and CAT1-b, the transition probabilities are updated through,

   $$\begin{aligned} \hat{a}_{i,i} &= \frac{\sum_{j=1}^{T_w} N_{i,j}}{\sum_{j=1}^{T_w} (N_{i,j} + 1)}, \\ \hat{a}_{i,i+1} &= 1 - \hat{a}_{i,i}, \\ & i = 1, 2, \ldots, S-1; \end{aligned} \quad (16)$$

   for CAT2, the transition probabilities are updated through,

   $$\begin{aligned} \hat{a}_{i,i} &= \frac{\sum_{j=1}^{T_w} N_{i,j}}{\sum_{j=1}^{T_w} (N_{i,j} + 1)}, \\ \hat{a}_{i,i+1} &= \frac{\sum_{j=1}^{T_w} \mathcal{I}(N_{i+1,j} \geq 1)}{\sum_{j=1}^{T_w} (N_{i,j} + 1)}, \\ \hat{a}_{i,i+2} &= 1 - \hat{a}_{i,i} - \hat{a}_{i,i+1}, \\ & i = 1, \ldots, S-2,\ and \\ \hat{a}_{S-1,S-1} &= \frac{\sum_{j=1}^{T_w} N_{S-1,j}}{\sum_{j=1}^{T_w} (N_{S-1,j} + 1)}, \\ \hat{a}_{S-1,S} &= 1 - \hat{a}_{S-1,S-1}. \end{aligned} \quad (17)$$

   Where $\mathcal{I}(\rho)$ is the indicator function and defined as $\mathcal{I}(\rho) = 1$, if ρ is true and $\mathcal{I}(\rho) = 0$, if ρ is false.

   (d) Iterate between step (b) and (c) until a preselected stopping iteration is reached or a model change threshold is achieved.

4. Experiments

The speech recognizer used in our study is the Sonic system [14], which employs a decision-tree triphone acoustic model and back-off trigram language model. The acoustic model is trained using the WSJ (Wall Street Journal) American English data, which is represented as the AM_p in Fig. 3. The feature used in our study is a 39-dimensional MFCC vector (static, delta, double delta).

The corpus used is the IViE British 8-dialect corpus [10]. Table 1 shows a summary of the training and test sets used from the corpus. The length of each test utterance is 9 seconds in duration. From Table 1, it is observed that there is on the average less than 40 minutes of training data for each dialect in the IViE corpus. There are only 10 training samples for each word and each dialect on the average. So, it is hard to train a robust HMM for each word and each dialect. The CAT algorithm therefore can be applied for this limited size corpus.

Table 1: *The used IViE corpus*

Data	Total Training Set				Total Test Set			Dialects/ Accents
	Vocab.	Spkrs	Size	style	Spkrs	Size	Style	
IViE	320	64	5 hours	read	32	86 minutes	Spontaneous	8 British dialects (Belfast, Bradford,Cambridge, Cardiff, Leeds, Liverpool, London, Newcastle)

As shown in Table 1, there are 96 speakers in total, with each speaker producing read and spontaneous speech. We use the read speech of 64 speakers as the training data. The read speech of the remaining 32 speakers is used to test the word classification error of CAT and baseline training algorithms, the result is shown in Fig. 4. The spontaneous speech of the remaining 32 speakers is used in the utterance dialect classification evaluation, with the results shown in Table 2.

Fig. 4 shows the word dialect classification error of the 3 CAT structures versus the baseline WDC training algorithm for the 8-dialect IViE corpus. From Fig. 4, we see that all three CAT based methods outperform the baseline WDC training algorithm significantly on the words with 3-or-more phonemes, and the 3 CAT structures achieve almost the same performance for word classification. Since all 8 dialects are from the United Kingdom, there are fewer differences across each dialect for phoneme deletion and phoneme insertion. If the dialects were a mixture from the United Kingdom and the United States, we would expect more differences. This may be the reason why CAT1-b and CAT2 do not outperform the CAT1-a structure. The CAT1-a is used in the following experiment. Table 2 shows the utterance classification errors of different algorithms. "WDC+CAT" means that the word models are trained by the CAT algorithm (CAT1-a is used). "WDC" is the basic WDC algorithm as in Sec. 2.2. When using CAT as the training algorithm in the WDC system, we obtain a 20.2% error reduction on utterance dialect classification. The relative error reduction from the baseline LVCSR system to that with "WDC+CAT" is 35.5%.

Table 2: *Utterance dialect classification error(%) of algorithms using the 8-dialect IViE corpus*

LVCSR	WDC	WDC+CAT
32.4	26.2	20.9

5. Conclusions

In this paper, we extended our previously proposed Word based Dialect Classification (WDC) algorithm. The original WDC works well on the large size corpus. For a small size corpus, however, it is difficult to train a robust HMM for each word and each dialect due to the limited size of the training data. We proposed a Context Adaptive Training (CAT) algorithm which can adapt the universal phoneme GMMs to dialect-dependent word HMMs. Using the CAT training algorithm on the basic WDC system, we obtain a 20.2% relative utterance classification error reduction. The error reduction from the baseline LVCSR system to the "WDC+CAT" dialect classification system is 35.5%.

6. References

[1] P. Angkititrakul and J.H.L. Hansen, "Advances in Phone-Based Modeling for Automatic Accent Classification", accepted to *IEEE Trans. Speech & Audio Processing*, will appear in 2005

[2] The CMU Pronunciation Dictionary, <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>

[3] V. Diakoloukas, V. Digalakis, L. Neumeyer, and J.Kaja, "Development of Dialect-Specific Speech Recognizers using Adaptation Methods", in *ICASSP* , vol.2, pp.1455-1458, Munich, Germany, April, 1997

[4] M. J. F. Gales and P. C. Woodland, "Mean and Variance Adaptation within the MLLR Framework", in *Computer Speech and Language*, vol.10, pp.249-264, 1996

[5] S. Greenberg, "On the Origins of Speech Intelligibility in the Real World", in *Proc. ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels*, vol.1, pp.23-32, France, 1997

[6] V. Gupta and P. Mermelstein, "Effect of Speaker Accent on the Performance of a Speaker-Independent, Isolated Word Recognizer", in *Journal of Acoustic Society of America*, vol.71, pp.1581-1587, 1982

[7] C. Huang, T. Chen, S. Li, E. Chang and J. L. Zhou, "Analysis of Speaker Variablility", in *Proc. EuroSpeech*, vol.2, pp.1377-1380, Sep, 2001

[8] R. Huang and J.H.L. Hansen, "Dialect/Accent Classification via Boosted Word Modeling", in *ICASSP*, Philadelphia, USA, March, 2005

[9] J. J. Humphries and P. C. Woodland, "The Use of Accent-Specific Pronunciation Dictionaries in Acoustic Model Training", in *ICASSP*, vol.1, pp.317-320, Seattle, USA, May, 1998

[10] IViE corpus, <http://www.phon.ox.ac.uk/∼esther /ivyweb/>

[11] C. J. Leggetter and P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", in *Computer Speech and Language*, vol.9, pp.171-185, 1995

[12] M. K. Liu, B. Xu, T. Y. Huang, Y. G. Deng, and C. R. Li, "Mandarin Accent Adaptation based on Context-Independent/ Context-Dependent Pronunciation Modeling", in *ICASSP*, vol.2, pp.1025-1028, Istanbul, Turkey, 2000

[13] D. Matrouf, M. Adda-Decker, L. F. Lamel and J. L. Gauvain, "Language Identification Incorporating Lexical Information", in *ICSLP*, vol.1, pp.181-185, Sydney, Australia, December, 1998

[14] B.Pellom, "Sonic: The University of Colorado Continuous Speech Recognizer", *Tech. Report TR-CSLR-2001-01*, Univ. of Colorado, USA, 2001.

[15] T. Purnell, W. Idsardi and J. Baugh, "Perceptual and Phonetic Experiments on American English Dialect Identification", in *Journal of Language and Social Psychology*, vol.18, no.1, pp.10-30, March, 1999

[16] P. Trudgill, "The Dialects of England", 2nd edition, *Blackwell Publishers Ltd*, Oxford, UK, 1999

[17] W. Ward, H. Krech, X. Yu, K. Herold, G. Figgs, A. Ikeno, D. Jurafsky, and W. Byrne, "Lexicon Adaptation for LVCSR: Speaker Idiosyncracies, Non-Native Speakers, and Pronunciation Choice", in *Proc. ISCA Workshop on Pronunciation Modeling and Lexicon Adaptation*, Colorado, Sep, 2002

[18] J. C. Wells, "Accents of English", vol. I, II, III, *Cambridge University Press*, Cambridge, UK, 1982

[19] M. A. Zissman and K. M. Berkling, "Automatic Language Identification", in *Speech Communication*, vol.35, pp.115-124, 2001

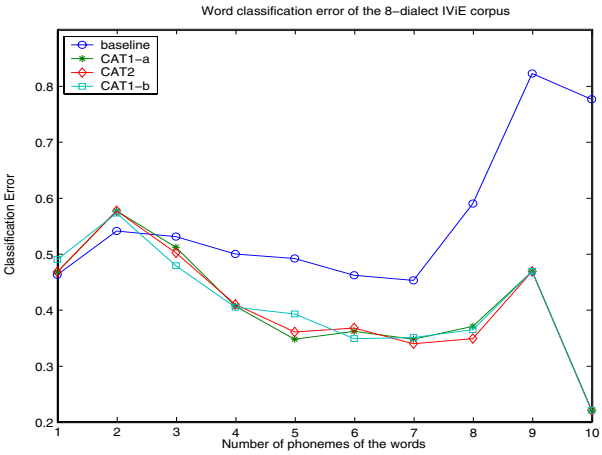

Figure 4: *The 3 CAT and baseline WDC word classification errors versus the number of phonemes in the word.*