Data balancing for efficient training of Hybrid ANN/HMM Automatic Speech Recognition systems

Ana Isabel García-Moral, *Member, IEEE,* Rubén Solera-Ureña, *Student Member, IEEE,* Carmen Peláez-Moreno*, *Member, IEEE,* and Fernando Díaz-de-María, *Member, IEEE*

***Abstract*—Hybrid speech recognizers, where the estimation of the emission pdf of the states of Hidden Markov Models (HMMs), usually carried out using Gaussian Mixture Models (GMMs), is substituted by Artificial Neural Networks (ANNs) have several advantages over the classical systems. However, to obtain performance improvements, the computational requirements are heavily increased because of the need to train the ANN.**

Departing from the observation of the remarkable skewness of speech data, this paper proposes sifting out the training set and balancing the amount of samples per class. With this method the training time has been reduced 18 times while obtaining performances similar to or even better than those with the whole database, especially in noisy environments.

However, the application of these reduced sets is not straightforward. To avoid the mismatch between training and testing conditions created by the modification of the distribution of the training data, a proper scaling of the *a posteriori* probabilities obtained and a resizing of the context window need to be performed as demonstrated in the paper.

***Index Terms*—Robust ASR, Additive noise, Machine Learning, Hybrid ASR, Artificial Neural Networks, Multilayer Perceptrons, Hidden Markov Models, Active Learning, ANN/HMM, MLP/HMM.**

I. Introduction

HIDDEN Markov Models (HMMs) have become the most employed core technique for Automatic Speech Recognition (ASR). After several decades of intense research in the field, the HMM-based ASR systems seem to be close to reaching their limit of performance. Some alternative approaches, most of them based on Artificial Neural Networks (ANNs), were proposed during the late 1980s and early 1990s. However, two main difficulties have prevented them from becoming mainstream: their inability to cope with the variable time duration of speech acoustic units and their increased training computational requirements.

Nonetheless, hybrid ANN/HMM systems (see [1] for an overview), and especially those based on Multilayer Perceptrons (MLPs), have found a place in the development of recognizers given that modern computers are becoming increasingly capable of coping with their computational requirements. However, their training with very large databases requires several adjustments and great doses of *know-how* [2]–[4].

Signal Processing and Communications Department, University Carlos III Madrid, Legans, 28911 Spain e-mail: (see http://gpm.tsc.uc3m.es).

Manuscript received February 5th, 2009; revised July, 3rd, 2009.

*Corresponding author.

Recent introduction of other types of models, such as Support Vector Machines (SVMs) in the hybrid architecture [5]–[7] have been found to be advantageous in noisy environments. Unfortunately, the computational demands of these machines are much greater than those of the MLPs (between two and three orders of magnitude, depending on the recognition task and the system architecture). SVMs being relatively new in the ASR field, there is a lack of *know-how* necessary to make them competitive and it becomes extremely difficult to test and tune these models owing to the large turnaround time for a single test. However, in state-of-the-art ASR experiments, large databases have become the only warrant of relevant and statistically reliable results.

To alleviate the problem of increased computational training requirements in hybrid systems, we propose the use of balanced training sets which contain the same number of samples per class. We have investigated this proposal under a hybrid MLP/HMM setup more appropriate and quick for research. Its application to SVMs is beyond the scope of this paper, but in our opinion, the conclusions are sufficiently clear and generic to be translated, to some extent.

The rationale for this proposal is that the nature of speech being extremely unbalanced, not all samples of a given database are equally informative, as will be exposed in section III. The problem of skewed or imbalanced data is receiving great attention in the machine learning community (see for example, [8]–[10]) and its solution is still under debate. Nevertheless, given that the speech databases are, by far, much more populated than the ones in the standard experimental frameworks in machine learning, barely any applications of these methods have been employed in ASR, with the notable exception of [11]. These solutions, however, do not take into account their computational requirements, which is our primary goal in this paper.

On the other hand, data selection techniques for active learning are becoming more popular in the ASR field. Nonetheless, they are not intended for reducing the computational cost of the traning stage, but for improving the performance. These techniques make use of different metrics to determine the worth of a given sample in the training process selecting the most informative. In conventional HMM systems, the selection unit is utterance (though the metric can be obtained at several levels [12]–[16]). In the case of hybrid systems, the computational effort needed to train the nets is typically much greater than that for the rest of the system and, therefore, we have focused our efforts on this task. This allows a much finer selection of the training data that, in this case, has been

performed on a frame basis. By using approximately 13% of the database, we achieve a similar performance as with the whole database in clean conditions and with a significant improvement in noisy conditions. The training of the proposed system is approximately 18 times faster.

Balancing the training data creates a mismatch between training and testing conditions since the learnt distribution of the data does not match those of the test. Thus, a proper scaling of the *a posteriori* probabilities and an adaptation of the word insertion probability need to be performed as demonstrated in the paper. Although this issue have received some attention in many relevant papers [11], [18], [20], [25], we found out that, in the ASR field, it has not been systematically addressed. Furthermore, depending on the task, the researchers choose to approach the problem in one way or another, finding in some cases that the experiments contradict the theoretical facts.

Finally, another side-effect of the modification of the original distribution of the training data can be observed in the selection of the optimal context window in the MLP. The use of extended context windows being a fundamental advantage of hybrid systems and an important source of improvement over conventional HMM systems, we have drawn the evolution of the balanced and unbalanced systems' performances with the context window. We found that the former achieve better results with shorter windows which, in turn, becomes another source of reduction of the computational demands.

The rest of this paper is organized as follows: Hybrid ANN/HMM systems are presented in section II with especial attention to the computation of the required likelihoods from the outputs of MLPs. Some notions of the data selection methods proposed in the literature are sketched in section III with an emphasis on the treatment of skewed data, which we have illustrated with speech data examples. Next, our proposal is described also in section III. Finally, experiments and results are presented followed by conclusions and suggested future lines of research.

II. Hybrid ANN/HMM systems for ASR

A. Motivation

As a result of the difficulties faced in the application of ANNs to speech recognition, mostly motivated by the duration variability of the speech instances corresponding to the same class, a variety of different architectures and novel training algorithms that combined both HMMs with ANNs were proposed in the late 1980s and early 1990s. The fundamental advantage of this approach is that it introduces a discriminative technique (ANN) into a generative system (HMM) while retaining its ability to handle the temporal variability of the speech signal. For a comprehensive survey of these techniques, see [1]. Besides, these hybrid architectures are very flexible allowing, for example, the introduction of long-term information into the feature vectors [2], [17].

In this paper, we have focused on systems that employ ANNs to estimate the HMM state posterior probabilities proposed by Bourlard and Morgan [18], [19]. Though at the time this approach was suggested, the use of ANN in speech recognition was still a challenging issue from a computational point of view, modern computers have certainly made it attractive. As a result, many recent papers make use of this technique [20]–[23] even substituting HMMs for more complex DBNs (Dynamic Bayesian Networks) as in [24].

The following are among the significant advantages of using hybrid approaches (from [25]):

- Model accuracy: ANNs have greater flexibility to provide more accurate acoustic models including the possibility of using different combinations of features along with different sizes of context. Features do not need to be uncorrelated because the network learns the local correlation between its input units. Therefore we can concatenate different types of inputs into the same input vector or even patch several consecutive feature vectors to represent the context. This has been used to include alternative features such as spectral parameters obtained by frequency filtering (FF) [20] or articulatory features [24] in the speech recognizer.
- Local discrimination ability (at a frame level). MLPs are trained to obtain class boundaries instead of providing an accurate (generative) model for each particular class.
- Parsimonious use of parameters: all the classes share the same ANN parameters (this does not hold for every ANN, but it does for MLPs).
- HMMs and ANNs exhibit complementary abilities for ASR tasks, which lead to higher recognition rates, especially under noisy conditions.
- Adaptation techniques have also been proposed (for example, speaker adaptation as in [22], [26], [27]).

As a drawback, we can mention that these implementations rely on an initial segmentation of the training set at the level of the classes considered by the ANN. That is, if the target of the ANN is phoneme classification, each training frame must have its corresponding phoneme label. However, large databases are rarely manually labeled at a phoneme level because of the enormous human effort necessary for the task. Therefore, most state-of-the-art hybrid recognizers perform an initial forced alignment with conventional HMM. This alignment becomes the ground truth for the training of the ANN. We have made use of this approach and further subdivided the phonemes into three sections (initial, middle, and final) making a finer segmentation attending to the distribution of the frames into the states of the HMM employed for forced alignment. We have used this alignment to illustrate important characteristics of the speech signal in section III.

B. Estimating the class likelihood for the HMMs

The starting point for the hybrid approach is the well-known capability of feed-forward networks, such as Multilayer Perceptrons, of estimating *a posteriori* probabilities, $P(q_l|x_t)$, of a certain class q_l, given an input feature vector x_t, when the system is trained in classification mode (see [28] for the fundamentals of MLPs).

Adopting a Maximum a Posteriori (MAP) criterion, the speech recognition problem can be stated as *finding the sequence of words $\tilde{W}$ that maximizes the quantity $P(W|X)$* where $X = x_1, ..., x_T$ is the sequence of input observation

features. However, to solve this problem, it is usually factorized using the Bayes theorem as

$$P(W|X) \propto P(X|W)P(W) \tag{1}$$

where the *a priori* probabilities $P(W)$ are usually modeled using a language model and the likelihoods $P(X|W)$ are estimated by the HMMs. In this context, W is modeled as a sequence of states $W = q_1, ..., q_L$ where each state describes the probability of occurrence of some feature vectors $p(x_t|q_l)$ (emission probability density function)[1]. The probability of the initial state $P(q_1)$ and the probability of transition between states $P(q_j|q_i)$ complete the model.

GMMs are mostly used to model the emission pdfs while in the hybrid formulation, the outputs of the ANN substitute these models. In the next section, we evaluate the necessary transformation that MLP outputs need to undertake.

C. MLP: posteriors and scaled likelihoods

To obtain the true emission (likelihood) pdfs from the outputs of the MLPs, we must use Bayes' rule once more:

$$\frac{p(x_t|q_l)}{p(x_t)} = \frac{P(q_l|x_t)}{P(q_l)} \tag{2}$$

Given that, in the decoding stage, the scaling factor $p(x_t)$ remains constant for every class, we can drop it from the equation. Therefore, the *a posteriori* probabilities should be normalized by the class priors to obtain what is called *scaled likelihoods*. Thus, systems of this type continue to be locally discriminant given that the ANN was trained to estimate *a posteriori* probabilities [29].

On the other hand, it can also be shown that, theoretically, HMMs can be trained using local posterior probabilities as emission probabilities, resulting in models that are both locally and globally discriminant. This fact was further reinforced by posterior theoretical developments in the search of a global ANN optimization procedure [25].

Nevertheless, in practice, there generally are mismatches between the prior class probabilities implicit to the training data and the priors that are implicit to the lexical and syntactic models used in recognition. In fact, some experimental results show that for certain cases, division by priors is not necessary [25], leaving a choice for empirical assessment over the particular task considered [20], [30].

In this paper we have further investigated in this direction postulating that the balancing of the training set adds on the advantage of producing the adequate (scaled) likelihoods without the need of applying any corrections irrespective of the different lexical and syntactic structures of the test set. In [11], the theoretical foundations for this assertion were already laid down although the experiments presented did not fully comply with the expectations.

[1]We denote the HMM states as q_l because, in this work, these states are synonymous of classes.

III. Data selection and the problem of class imbalance in speech

In the ASR community, there has been a long-standing saying that goes "there is no data like more data". It recognizes the empirical observation that one of the most influential factors in the quality of a recognizer is the size of its training database. However, with the growth of databases in the last few years, a question about the eventual saturation of that lemma has been raised [15] adding a preoccupation about an adequate treatment of erroneously labeled samples. Several active learning solutions implying a selection of the data have been proposed not only in ASR but also in Spoken Language Processing [31], [32], emotion recognition [33], or language identification [34] among others. It is, however, worth noting that the main goal of active learning is the improvement in the precision of the target classifier disregarding, most of the time, the computational costs. In other words, using a sample selection method may increase the overall computational cost of the complete system.

On the other hand, there has been an increased interest among the machine learning community in assessing the influence of *training classes imbalance* and *overlapping* in a variety of classification techniques. Though these two characteristics are a well-known fact for speech practitioners, they have not been fully explored in conventional speech recognition given the need for treating speech utterances as the smallest unit for selection in sequence models like the classical HMMs. This is not the case of hybrid recognizers as each frame is presented individually (and in fact in random order) to the ANN. However, a probabilistic sampling method aimed at changing the phoneme distribution of the training set in a hybrid framework [11] is, to our knowledge, the only contribution in this field. Once more, these methods do not treat the problem of reducing the computational burden of the classification algorithm.

In this paper, we propose a simple selection method that takes into account the class imbalance problem of speech data, which consists of downsampling the majority classes so as to come out with fully balanced sets. This reduces the computational demands significantly without producing corresponding reductions in performance and even improves the results in noisy environments.

With the purpose of contextualizing our proposal, the next section introduces the basics of data selection and active learning, and is followed by a presentation of the problem of class imbalance, which we have illustrated with speech data examples. Finally, we present our proposal with emphasis on the implications of the downsampling method for the training of the hybrid speech recognizer.

A. Generic data selection

Several data selection techniques have been proposed in the literature and variations of these can be found under names such as novelty detection, selective sampling or active learning. In general, these methods can be classified into two groups [35]:

- **Generative methods** aim at selecting the best samples from unlabeled data to maximize data labeling investment returns. Though generative methods have important applications in speech recognition to avoid the expensive process of speech labeling and tagging [36]–[38], we are primarily concerned with the selection of labeled data with the purpose of relieving the training burden of the ANNs that estimates the pdf of each state of the underlying HMM.
- **Selective methods** try to select an adequate subset from labeled data to maximize performance or reduce the computational effort while maintaining a similar performance. Here, we can further distinguish between *wrapper* and *filter* approaches. The former employs a statistical resampling technique (such as cross-validation) and uses the actual target learning algorithm to estimate the accuracy of the subsets. Its disadvantage is its high cost as the learning algorithm has to be invoked repeatedly and, therefore, is not adequate for our goals. The later approach has been employed in several papers aiming at selecting entire sentences to allow the selection to be applied to a conventional HMM system. In those works, different metrics at various levels (frame, phone or utterance) were used [14], [15] recognizing that the lenght of the training utterance is an important factor [16]. Utterance selection was also used for discriminative training [39] and utterance verification [40].

In [41] we used the *filter* approach at a frame level to alleviate the computational training load of a system which employed ANN to produce acoustic features for a conventional HMM system. In that work, the recognizer was not hybrid in the sense we are using in this paper, as the outputs of the MLPs were not used to estimate the pdf of the states of the HMMs. Our aim was then to reduce the number of samples using a selection metric based on the entropy of the outputs of a downsized ANN which acted as the *filter*. Its design and dimensioning was an important issue in order to obtain an efficient solution in terms of computational costs. The class imbalance problem, noise robustness and the fact that, in hybrid recognizers, the outputs of the ANN need to be massaged to represent the likelihoods of the different classes (or acoustic units) and the influence of the size of the context window were not investigated at that point.

B. The problems of skewed and overlapped data

The class imbalance (or skewed data) is a known problem in machine learning still under debate (see the special issues [8], [9], [42]). Although most learning systems assume that the training data sets are balanced, this is not always the case in real-world data where several classes might be represented by a large number of examples, but others by only a few.

This is certainly the case of speech data where we can identify two sources for this lack of balance: first, the natural distribution of the sounds of a given language is not uniform. Moreover, this distribution depends on the task for which the ASR system is being designed. Second, owing to the time uniform sampling of the speech waveforms, those phonemes with longer temporal durations produce a larger amount of samples.

In Figs. 1 and 2 we show these distributions clearly exhibiting the skewness of speech data. Full details of the experiments configuration are provided in section IV but we have considered it useful to illustrate this section with this example. Recall from II-A that we need to perform a preliminar forced alignment to obtain a partition of the training frames into the classes considered by the frame-level classifier implemented by the ANN.

The black series of Fig. 1 shows the relative frequency of appearance of the (active) states of the initial Hidden Markov Model that produced the forced alignment we used for obtaining ANN labels (in percentage) in the training data set. Thus it represents the first of the two sources of skewness mentioned above. Note that due to the topology of the HMMs employed (see section IV) the three black bars corresponding to the same phoneme are equally tall with the exception of the *silence* (*/sil/*) whose topology is slightly different to account for the distinction between short and long pauses. Therefore, the relative frequency of the phonemes in the database can be computed as the sum of those three bars.

The white series of bars represent how the frames are distributed among these states, demonstrating the second source of skewness. To make it clearer, Fig. 2 shows the average number of frames the recognizer spends in each state. We can see that, with the notable exception of */sil/* and */tS/*, the average duration is between 2 and 4 frames which is, however, an important variation given that if, for example, a training set contained the same number of */r/* and */s/*, there would be twice as many samples of the former as the later.

At this point we find it useful to explain that long pauses at the beginning and the end of the utterances in the initial available SpeechDat database [43] have been removed in all the experiments and discussions of this paper, leaving only short pauses between words. This can be clearly observed in Fig. 2 where the central state of */sil/*, mostly devoted to model inter-word short pauses, received considerably longer sequences of frames than the side ones that exclusively take in pre and post utterance silences that have been cut out. This is usual practice in ASR but it is even more important for our experiments given that, even with the reduction, it keeps being a majority class.

Another key observation of skewed data difficulties is that highly imbalanced problems generally have highly non-uniform error costs that heavily penalize the overall performance when errors ocurr in the minority classes. The case of ASR is one such example, since many times the short phonemes are the key to distinguishing among confusable sets of words, and therefore, are more informative. However, it is not easy to effectively quantify these costs since they depend on the confusability of the vocabulary of a particular task. Here, it is not our intention to adapt our selection to a certain task, but to extract more general conclusions.

On the other hand, a comparative analysis of techniques to alleviate the problem of imbalanced training sets [44] highlights the fact that class imbalance does not just hinder the performance of the learning systems; a major point of

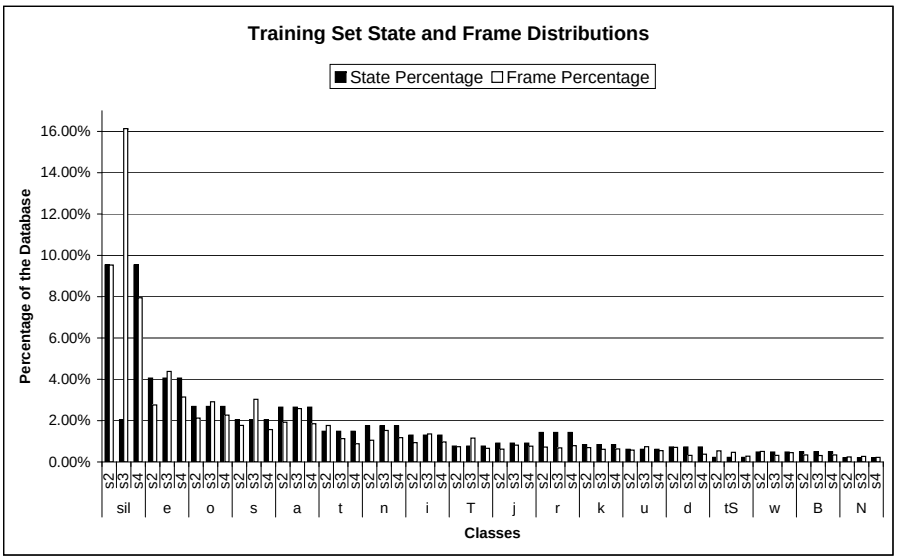

Fig. 1. *State and frame distributions in the training dataset using a plain HMM recognizer employed (three emitting states per phone labeled s2-s4 in the figure) to produce the forced alignment for the MLPs training data labeling. Phonemes not present in the test set have been omitted for simplicity though the same conclusions hold for them. The database employed in our experiments is the well-known SpeechDat Spanish database [43]. This large vocabulary (more than 24,000 words) continuous speech recognition database comprises recordings from 4,000 Spanish speakers recorded at 8 KHz over the PSTN using an E-1 interface, in a noiseless office environment.*

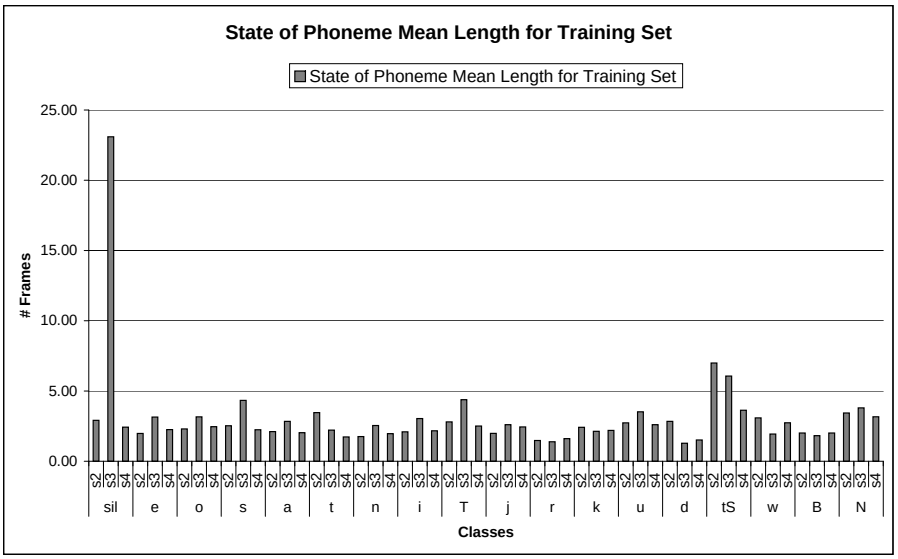

Fig. 2. *Mean length of the states of the phonemes in the training dataset using a plain HMM recognizer employed to produce the forced alignment for the MLPs training data labeling as in figure 1. Phonemes not present in the test set have been omitted for simplicity though the same conclusions hold for them.*

concern is that the scarce availability of training samples of the minority classes can be devastating in the presence of complicating factors such as class overlapping [10]. The difficulties also arise when the under-represented classes can be internally partitioned into several subclasses [45] making these subclasses unevenly represented in the training data.

We believe this is the case for speech data given that effects such as coarticulation make the border of the classes quite blurry and the mechanisms of production of certain classes of phonemes (for example, plosives) produce non-stationary signals. These are mainly the reasons for using subphonetic units in speech recognition as we do (see section IV-B3), but these units still tend to be quite overlapped owing to the enormous difficulty of obtaining an accurate segmentation, which most of the times is done automatically (at present, very few databases are phonetically manually labeled).

Techniques proposed to alleviate the problem of imbalance can be separated into the ones that concentrate on the input training data and those that adapt the training algorithm [46], [47]. According to [46], in most cases the former are preferable though some methods based on modifications of certain classification algorithms (for example, Support Vector Machines) are advantageous [48]. Besides, when *a posteriori* probabilities are available, as in hybrid ASR, they can be

further manipulated to account for different class error costs. This process is more flexible than wiring these costs into the classification algorithm because it allows adaptation to change or unavailability during training costs [49]. We shall concentrate on these techniques in the remainder of this paper.

Techniques applied to the selection of the input data can be further classified into those based on oversampling of the minority classes and those based on undersampling of the majority ones. In principle, the advantage of random undersampling is the avoidance of the overfitting that often occurs with random oversampling. On the other hand, the possibility of discarding some informative samples from the majority classes is its major drawback. To overcome these limitations techniques such as Kubat's [50] for undersampling or SMOTE (Synthetic Minority Over-sampling Technique) for oversampling [51] have been proposed. In [8] several methods based on ensembles and cascades of classifiers are designed to make use of all the samples without biasing the classifier.

However, the best solution is generally a combination of both through a *wrapper* approach [47], [49], [52] where the amounts of under and over sampling are determined based on performance evaluations of some partitions of the data (sometimes referred as *learning by recognition*). This proposal heavily depends on the performance metric used for the evaluations, which is a crucial notion in data selection. Examples of such metrics include AUC (Area under the ROC curve), f-measure (or β varied f-measure) [44], [46]. On the other hand, [53] acknowledges difficulties when methods design for binary classifications are to be exported to multiclass proposing techniques as threshold moving to balance the misclassifying costs of the different classes.

C. Balancing speech data

When dealing with speech data for ASR it is very important to realize that the magnitudes of the databases employed in order to obtain relevant results are usually many times bigger than those employed in machine learning. For example, the most recent papers reviewed in the previous subsection ([8], [49]) employ databases of thousands of examples (the largest containing 20,000 samples), while in this paper our training data set comprises 16 million speech samples. When testing the benefits of undersampling techniques as in this paper, the use of these large data sets is required in order to obtain statistically significant results capable of corroborating the theoretical hypothesis and relevant in the sense of demonstrating their effectiveness in a real situation when the implementation of the recognizer with the full database is really challenging.

Therefore and owing to the enormous computational burden involved in the wrapper method, we have ruled it out for our recognizer. In an effort to keep it as simple as possible, and in view of the fact that the (a priori) determination of the amounts of undersampling needed for each class is very difficult given its dependency on the particular database structure, we have adopted what we call the *balanced* solution where each class is equally represented in terms of number of samples. This very simple but certainly efficient solution has the additional advantage of producing the desired *likelihoods* as the outputs of our MLPs as will be demonstrated in section IV-C.

Although frame accuracy is the only suitable figure of merit to evaluate the performance of the MLP in the hybrid archictecture, it is not an appropriate metric of the performance of the complete ASR system when the database is skewed, because it is biased toward the majority classes which, in our problem, is clearly (but not exclusively) the *silence*. This means that if not properly balanced, the MLP tends to model these classes very accurately because they produce the highest reward in terms of accuracy. Then, in an ASR problem, as we know, the most unbiased metric is the WER (Word Error Rate), which is our ultimate goal. Note that if this metric is used under a wrapper approach for the determination of optimum distribution of data for each iteration, the full recognizer should be evaluated several times.

In section IV-D, both FER (Frame Error Rate) and WER (Word Error Rate) will be analyzed, but here we find it useful to illustrate the effects of balancing the data in terms of the *entropy* of outputs of the MLPs:

$$h_t = -\sum_{i=1}^{N_q} p\left(q_i|x_t\right) log_2 p\left(q_i|x_t\right) \tag{3}$$

where h_t is a measure of the *difficulty* of a classification decision based on the outputs of the MLP and N_q is the total number of emitting states of the system. Therefore, high entropy values indicate that taking a decision is going to be difficult while low values signify it will be easy to make it (not necessarily implying the right class will be chosen). Entropy was the metric employed in the filter approach in [41] and in this section we illustrate that, though indirectly, the balancing solution presented in this paper also contributes to reduce the entropy of the minority classes.

Fig. 3 represents the average entropies of the outputs of the MLP for the samples associated with each state of the HMM. Not all phonemes improve with the balancing of the training data, but most of them do (/B/, /N/, /d/, /i/, /j/, /k/, /r/, /tS/, /u/, and /w/). A second group only increases entropy slightly or in some of their states (/T/, /a/, /n/, /o/, and /s/) and a third one becomes the most negatively affected by the balancing (/e/, /t/, and /sil/). Overall, the most harmed phonemes are the most overrepresented in the unbalanced database.

The case of /sil/ deserves special attention since its middle state was performing exceedingly well for the full unbalanced database. The fundamental difference of this state is that it is trained with short pauses (between words) as opposed to the ones in the extremes that are exclusively trained with long pauses (beginnings and ends of utterances). Contrary to their denominations short pauses are longer than long ones (see Fig. 2), but this is an effect of the presentation of the speech waves in our database given that the initial and final silences in the utterances have been suppressed as we indicated in section III-B. This makes the central state of /sil/ to be the most overrepresented class. Hence the classifier trained with the unbalanced database finds it more profitable (in terms of accuracy) to erroneously assign the /sil/ label to the minority classes rather than miss a truthful one. This effect is clearly rectified using the balanced sets and can explain the favorable results we have obtained with these configurations as will be

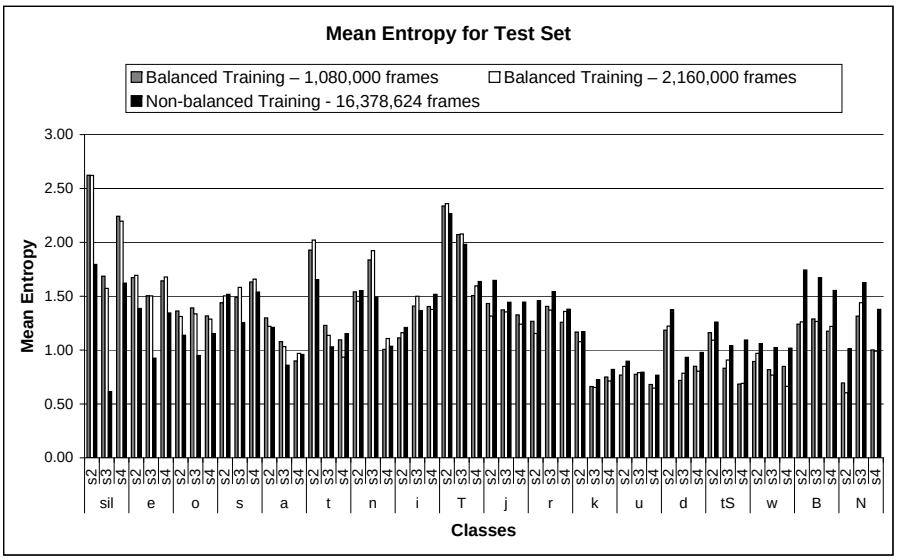

Fig. 3. *Average entropies per state for a connected digits test set when the MLP has been trained with approximately a million of balanced samples (gray series), two million (white) and the full unbalanced trainset (black) of figures 1 and 2.*

shown in section IV. We would like to highlight that, though this is an extreme and clear case, the same principles apply to the rest of majority classes.

IV. Experiments and results

A. Database

1) Description: The database employed in our experiments is the well-known SpeechDat Spanish database [43]. This large vocabulary (more than 24,000 words) continuous speech recognition database comprises recordings from 4,000 Spanish speakers recorded at 8 kHz over the PSTN using an E-1 interface, in a noiseless office environment.

The database is partitioned into three main sets: training set, development or validation set, and test set, representing 80%, 8%, and 12% of the database, respectively.

The original database is preprocessed to eliminate the *silence* samples placed at the beginning and end of the sentences, using the time marks available for this purpose in the database label files. As a result of this preprocess, the *training* set contains approximately 50 hours of speech from 3,146 speakers (71,046 utterances). The callers speak 40 items with varied content comprising isolated and connected digits, natural numbers, spellings, city and company names, common application words, phonetically rich sentences, etc. Most items are read and some of them are spontaneously spoken.

The *development* set contains 7,436 utterances from 350 different speakers (5 hours of voice after preprocessing) with the same varied content as the training data set. We use this dataset to select both the back-propagation coefficient (μ) in the MLP and the word insertion log probability for the Viterbi decoder.

Finally, the *test* set (connected digits task) consists of 2,122 utterances and 19,855 digits (5 hours of post-processed speech) from 499 different speakers. Thus, the number of recognized phones is restricted to 18 (we have dropped the remaining phones from our training data set). As shown in Fig. 4, the number of discarded samples (on the right of the vertical dotted line) only represents an 8.8% of the original training set. Once more, we clearly observe the skewness of the data.

As we already mentioned in section III-C, we have prepared alternative balanced data sets to reduce the training computational burden. The new training data sets are built from the original one (non-balanced) by selecting phone samples randomly so that each class is equally represented. Table I summarizes the distribution of data into these sets.

2) Database contamination: We have tested our systems in clean conditions and in the presence of additive noise. For that purpose, we have used two different types of noises (*white* and *babble*) extracted from the NOISEX-92 database [54]. These noises have been added to the clean speech signals at four different signal-to-noise ratios (SNRs), namely 12 dB, 9 dB, 6 dB, and 3 dB. Only the testing subset has been corrupted in the way previously stated, whereas the acoustic models have been estimated or trained using only clean speech.

B. Baseline system

1) Feature extraction: In our experiments, we have used a classical parameterization based on 12 MFCCs (Mel-Frequency Cepstral Coefficients) plus energy, and their first and second derivatives. Thus, the resulting feature vectors have 39 components. These MFCCs are computed every 10 ms using a temporal window of 25 ms.

In this work, we have considered a per-utterance normalization of the cepstral coefficients, more appropriate in the case of noisy environments where training and testing conditions do not match. Besides, in the case of MLP/HMM hybrid systems, this normalization is necessary to ensure the convergence of ANNs [28]. Thus, every parameter is normalized in mean and variance according to the following expression:

$$\hat{x}_t^{(i)} = \frac{x_t^{(i)} - \mu^{(i)}}{\sigma^{(i)}}, \tag{4}$$

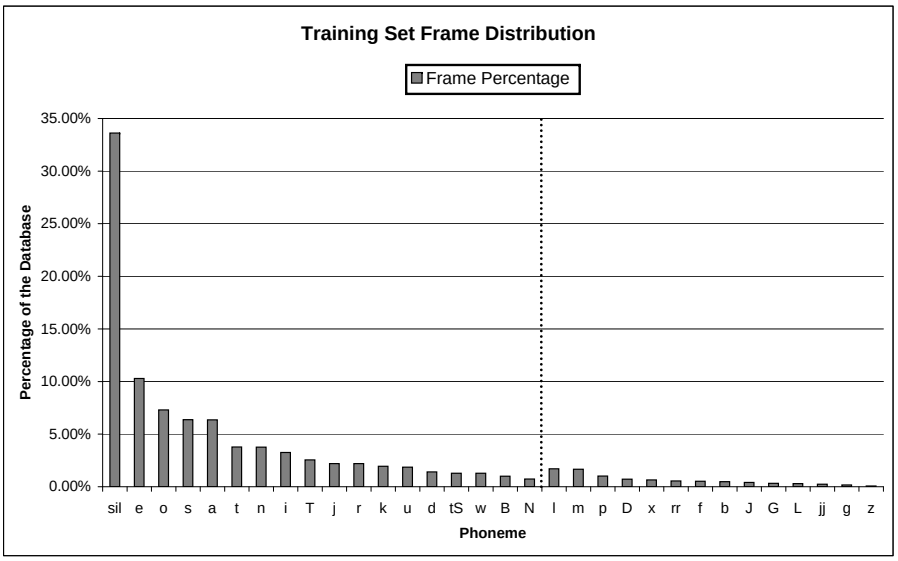

Fig. 4. *Phoneme distribution in original unbalanced Training Set. The vertical dotted line separates the phonemes present in the Digits Test Task (to the left) we have employed from the absent.*

TABLE I
DATASET PARTITION: THE THREE EXPERIMENTS (NB -NON-BALANCED-, B1 -BALANCED 1- AND B2 -BALANCED 2-) CARRIED OUT IN THE PAPER DIFFER IN THE TRAINING PORTION OF THE AVAILABLE DATASET USED. DEVELOPMENT AND TEST SETS ARE THE SAME FOR ALL OF THEM.

Exp.	**Train**		**Development**		**Test**	
	# frames	**Distribution**	**# frames**	**Distribution**	**# frames**	**Distribution**
NB	$16,378,624$	Non-Balanced	$1,682,065$	Non-Balanced	$1,656,102$	Non-Balanced
B1	$1,080,000$	Balanced				
B2	$2,160,000$	Balanced				

where $x_t^{(i)}$ represents the i^{th} component of the feature vector corresponding to frame t, and $\mu^{(i)}$ and $\sigma^{(i)}$ are the estimated mean and standard deviation from the whole utterance, respectively, for the i^{th} component.

2) HMM alignment system: A simpliflied left-to-right HMM-based recognition system, based on that described in [55], is employed to produce a forced alignment necessary to obtain the labels for the MLP, as SpeechDat is not phonetically labeled. We have used such a basic baseline ASR system for the sake of simplicity. Nonetheless, it could include more sophisticated techniques, with minimal impact on the overall conclusions of this work.

Each of the 18 context-independent phone models consists of 3 active states (plus initial and final non-emitting states) where emission probabilities are modeled by a mixture of 32 Gaussians. From this system, we are interested in obtaining the state-level segmentation of the training set, i.e., we label each frame with one of the possible 54 states. To avoid the potential appearance of empty states, the HMM topology does not allow to obviate any of the states except in the */sil/* model whose central state is designed to model short pauses and allows a jump from the first emitting state to the last one and viceversa.

The Word Error Rate (WER) obtained for this baseline HMM recognition system in clean conditions is 2.41%, this value being higher than the WER obtained with the MLP/HMM baseline system considered here, as we will show in the following sections.

3) Baseline MLP/HMM system: Our hybrid MLP/HMM baseline system employs an MLP to estimate the HMM state emission probabilities that will be used by a Viterbi decoder to obtain the transcribed word sequence. Fig. 5 shows the block diagram of our MLP/HMM system.

We have trained MLPs with balanced and non-balanced training sets specified in Table I. In all the cases, the MLP has a single hidden layer with 1,800 units. The input MLP dimension depends on the input context window considered (see section IV-E), and the 54 outputs provide *a posteriori* probabilities for each of the states of our system.

The different MLPs were trained using a relative entropy criterion, and the back-propagation factor, μ, was empirically found for every network by using the development set described in Table I. This set has also been employed to select the adequate word insertion log probability for each experiment with different training and test conditions, as we have found the different balancing alternatives tested very sensitive to this value [29].

C. Scaled likelihoods and a posteriori probabilities

In the hybrid approaches, the *a posteriori* probabilities obtained as the outputs of the MLP substitute the emission probabilities which are modeled by GMMs in classical recognition systems. As we stated in section II-B these probabilities must be transformed into *(scaled) likelihoods* to comply with the theoretical framework. Nevertheless, this scaling does not

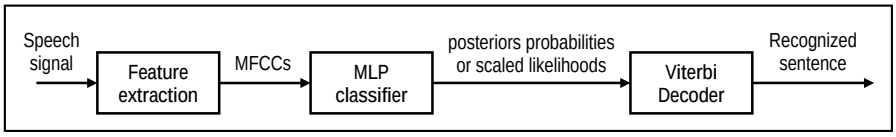

Fig. 5. *Block diagram for a generic hybrid MLP/HMM system*

always lead to the best performance, which is mostly attributed to mismatches between the *a priori* probabilities of the training and test datasets.

Given that the balancing of the data we are proposing heavily modifies the training data structure and that for our task, training and test sets have markedly different lexical and syntactic structures (see section IV-A), it is of paramount importance to investigate the effects of scaling on this system. For this purpose, we have applied several scalings to the MLP outputs with the aim of establishing the most suitable class prior. In particular, we have compared the following *a prioris*: none (*No-scaling*), *a priori* probabilities estimated from the training set (*P_Trn*), from the development set (*P_Dev*) and from the test set (*P_Tst*) (the last one to verify that the data from the development set was capable of providing a suitable estimation of the test structure, though of no practical utility).

Figs. 6-8 show Word Error Rates (WER) for these experiments and the three balancing conditions described in Table I in clean and noisy conditions. For clarity reasons, the figures only show the results for $SNR = 6\ dB$ for both *white* and *babble* noises. For all the other SNRs evaluated (3, 9, and 12 dB), the conclusions are identical. Figs. 6-8 only show the results when the MLP input feature vectors are not augmented with previous or future vectors (i.e. one frame long analysis window). The same analysis with different context influence (analysis windows of 3, 5, 7, and 9 frames long) has been done, but it has been omitted for the sake of clarity, though the same conclusions also hold (see section IV-E for more information about the context influence in WER).

The main conclusion we can draw is that when the training data are strictly balanced (experiments B1 and B2) the best results are obtained with no scaling. In these cases the *a prioris* of the balanced training sets are identical and thus the outputs of the MLP are directly the *likelihoods* we are looking for.

On the contrary, when we do not balance the training set (experiment labeled NB), we must normalize the MLP outputs by the *a priori* estimated from the training data. This fact is true across all the noisy conditions we have evaluated except for clean conditions where no-scaling beats any normalization. Anyhow, the difference appreciated with the best normalization (*a prioris* from the training set) in clean conditions is small while the preference for this normalization is clear for all the noises and SNRs evaluated. This leads us to choose the trainingset-based one as the reference normalization for NB.

It is worth mentioning that the differences between the normalization by *a prioris* derived from the development and test sets are usually non-significant (except from some balanced experiments with babble noise for which the clear winner is, nonetheless, the *no scaling* option).

As a conclusion, we can postulate that scaled likelihoods should always be estimated using the prior probabilities from the training data, which implies that, when these data are balanced, there is no need for scaling at all.

D. The benefits of balancing training data

Once we have obtained the best scaling for each experiment (no-scaling for B1 and B2 experiments, and training *a prioris* for NB), we proceed to analyze the benefits of balancing training data.

Fig. 9 presents the effects of balancing data in both Frame and Word Error Rates. While FER decreases with the addition of more training data, WER exhibits the opposite effect. Frame accuracy is not an appropriate metric of the modeling ability of the ANN over the whole unbalanced data set due to the skewness of the training database that makes the system biased toward the majority classes. Nevertheless, we must keep in mind that, as the MLP training stage is done independenty of the HMM, Frame Error Rate is the only available performance metric for designing and evaluating the MLP.

Taking into account the WER, which is our ultimate goal, we can conclude that using a balanced training subset of approximately 6.5% of the training data does not damage the performance of the system in clean conditions. Moreover, in noisy environments, this balanced data selection is beneficial being even more advantageous as SNR decreases.

E. Including context in balanced sets

The beneficial influence of the context in hybrid systems is well attested [2], [17]. However, its inclusion incurs an increase of computational demands owing to the increment in the MLP input dimensions. Usually, a context length of nine frames is employed when the acoustic units considered are monophones. This is justified by empirical measures of the phones' mean length.

In this paper, however, the acoustic units employed by the MLP are the states of the segmenting HMM in which case the mean length is approximately three frames. Therefore, we find it significant to analyze the evolution of the performance of our hybrid MLP/HMM system with this contextual information looking for some influence of the reduction of the training data set.

Figs. 10-12 show these results. Again, for the noisy conditions, we have chosen $SNR = 6\ dB$ as a representative case for the sake of brevity, as the conclusions hold for the rest of the $SNRs$.

As an overall conclusion, we can observe that as we include bigger contexts the performance differences between the three experiments (B1, B2 and NB) decrease. In general, though in

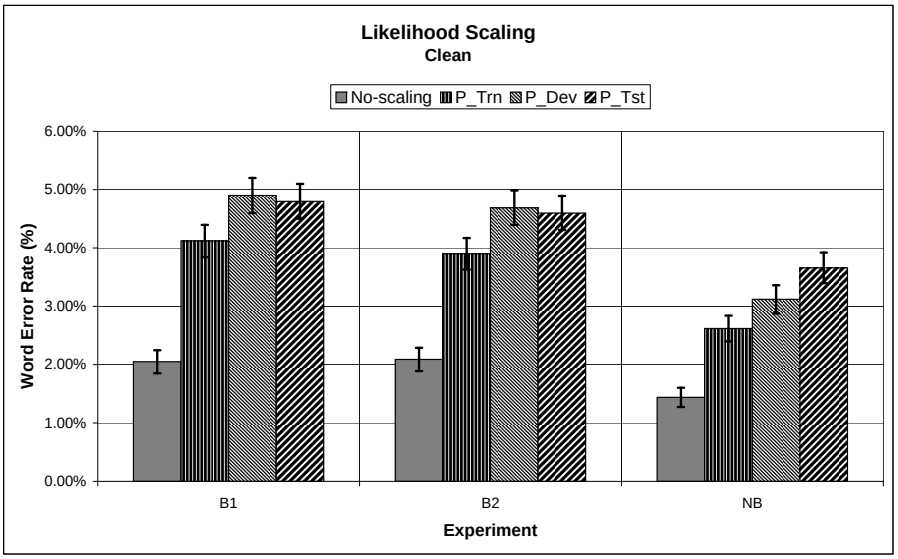

Fig. 6. *Likelihood scaling in a Clean Environment with a context window length of 1 frame (i.e. no context). Vertical segments in the middle of the bars represent 95% Confidence Intervals (CI)*

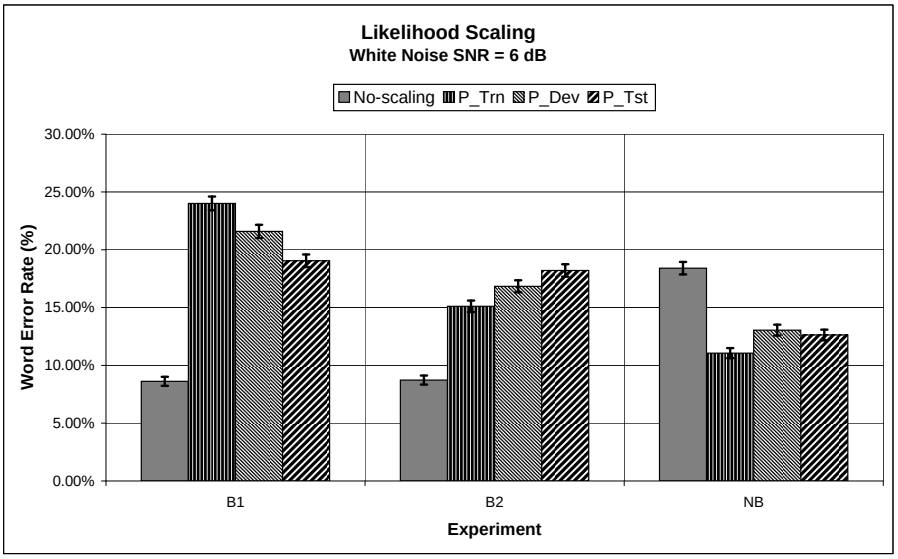

Fig. 7. *Likelihood scaling in a White Noise Environment with SNR = 6 dB with a context window length of 1 frame (i.e. no context). Vertical segments in the middle of the bars represent 95% Confidence Intervals (CI)*

some cases the differences are not significant, the best choice is to use B2.

Once more, we can observe a difference between the evolution of the balanced and non-balanced experiments now with the context length; for the balanced one, the best choice is a three-frame wide context window as we cannot observe any significant improvement from a bigger one. For the non-balanced experiment, the curves are steeper from 1 to 5 context lengths, the last being the best choice. We can hypothesize that the skewness of the non-balanced set that, as pointed out in section III, makes the MLP biased toward better modeling of the longest phones, finds the use of wider context windows helpful as they are more appropriate for those phones. When that skewness is corrected, a smaller window is preferred.

Thus, we can conclude that the benefits of using a reduced training set are twofold; fewer samples for training and a smaller input dimension. Moreover, if we compute the number of free parameters to estimate in the MLP, we find that the choice of a reduced acoustic unit together with a balanced training set implies an important reduction of the context, compensating for the increase in the output dimension of the MLP. In particular, for the state acoustic unit chosen in this paper, the total number of free MLP parameters to estimate for the balanced experiments is half those needed for monophones and a standard context window of 9 frames, and for the non-balanced, the proportion is approximately, two-thirds.

It is worth mentioning that the experiments presented in section IV-C were also performed with all the context windows presented in the present section and the conclusions extracted were exactly the same as those obtained in that section: scaled

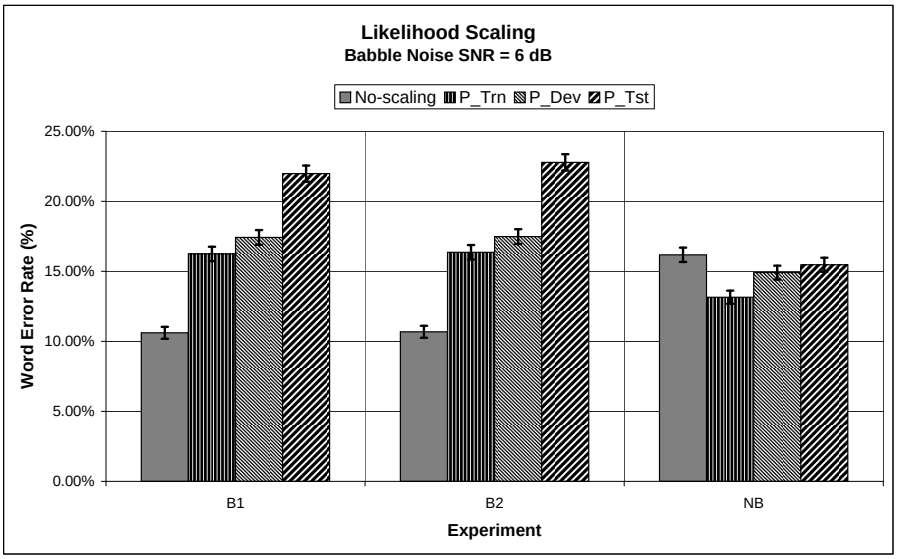

Fig. 8. *Likelihood scaling in a Babble Noise Environment with SNR = 6 dB with a context window length of 1 frame (i.e. no context). Vertical segments in the middle of the bars represent 95% Confidence Intervals (CI)*

likelihoods employed in NB experiment must be obtained by normalizing the MLP outputs (*a posteriori* probabilities) with the *a priori* probabilities estimated from the training set while, for experiments B1 and B2 there is no need for normalization.

Finally, with regard to the computational burden, we must point out that the best choice, B2 with a context length of 3 frames, is 18 times faster than NB and only 1.5 times slower than B1 experiment.

V. Conclusions and further work

The ANN/HMM hybrid systems presented in this work are inspired by Bourlard and Morgan [18] and have been found to compare favorably with a classical HMM-based system. In this paper, we have investigated the reduction of the computational burden associated with them by reducing the size of the training data set. Specifically, we have found that:

- Balancing the data presented to the ANN to train with the same number of samples per class (acoustic unit) is a good and simple choice, obtaining similar or even better performances than those obtained for the whole database. This is attributed to a more balanced modeling of the different classes by the MLP, suppressing the bias toward better modelling of the most populated.
- In the previous situation there is no need to obtain scaled likelihoods to introduce the outputs of the ANN into the hybrid system. *A posteriori* probabilities give the best results, as the *a priori* probabilities of the balanced training set are identical for every class.
- Besides, shorter context windows provide comparable results reducing the dimensionality of the input feature vectors, which has an impact on the computational requirements. For the acoustic unit considered in this paper (state of phoneme), the optimal context window is three frames wide with balanced data sets and five with non-balanced ones.
- The previous conclusions become even more evident and remarkable in noisy environments.

An immediate line of future research is the application of the previous conclusions to more computationally demanding models like SVMs. In this line of research, several peculiarities of the SVM must be taken into account. In addition, the application of more elaborate means of selecting the training data is also a challenge always bearing in mind, however, that the data selection method needs to remain very simple for the final solution to be advantageous in terms of computational requirements.

Acknowledgments

This work is partially supported by the regional grant (Comunidad Autónoma de Madrid - UC3M) CCG06-UC3M/TIC-0812 and a project funded by the Spanish Ministry of Science and Innovation (TEC 2008-06382).

References

[1] E. Trentin and M. Gori, "A Survey of Hybrid ANN/HMM Models for Automatic Speech Recognition," *Neurocomputing*, vol. 37, no. 1-4, pp. 91–126, April 2001.

[2] N. Morgan, B. Chen, Q. Zhu, and A. Stolcke, "Scaling up: learning large-scale recognition methods from small-scale recognition tasks," in *Special Workshop in Maui (SWIM)*, Maui, USA, 2004.

[3] N. Morgan, Q. Zhu, A. Stolcke, K. Sonmez, S. Sivadas, T. Shinozaki, M. Ostendorf, P. Jain, H. Hermansky, D. Ellis, G. Doddington, B. Chen, Ö. Çetin, H. Bourlard, and M. Athineos, "Pushing the envelope-aside," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 81–88, September 2005.

[4] Q. Zhu, A. Stolcke, B. Chen, and N. Morgan, "Using MLP features in SRI's conversational speech recognition system," in *Proceedings of the 9th International Conference on Speech Communication and Technology (Interspeech'2005)*, Lisboa, Portugal, September 2005.

[5] R. Solera-Ureña, D. Martín-Iglesias, A. Gallardo-Antolín, C. Peláez-Moreno, and F. Díaz-de-María, "Robust ASR using support vector machines," *Speech Communication*, vol. 49, no. 4, pp. 253–267, April 2007.

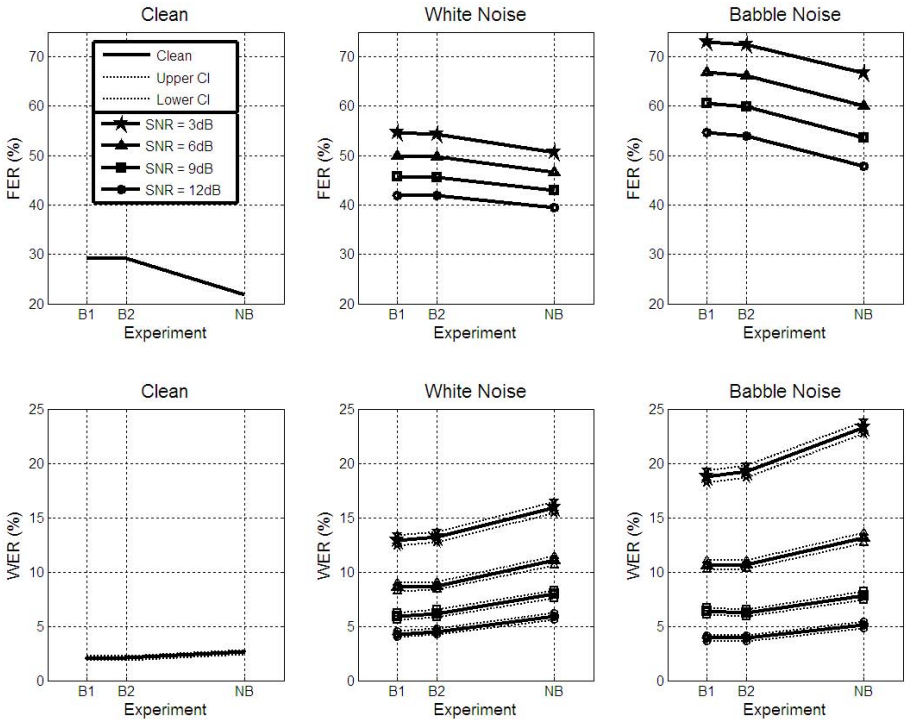

Fig. 9. *Effects of balancing training data in Frame and Word Error Rates. Legend in top-left figure applies to all of them and dotted lines represent 95% Confidence Intervals (CI).*

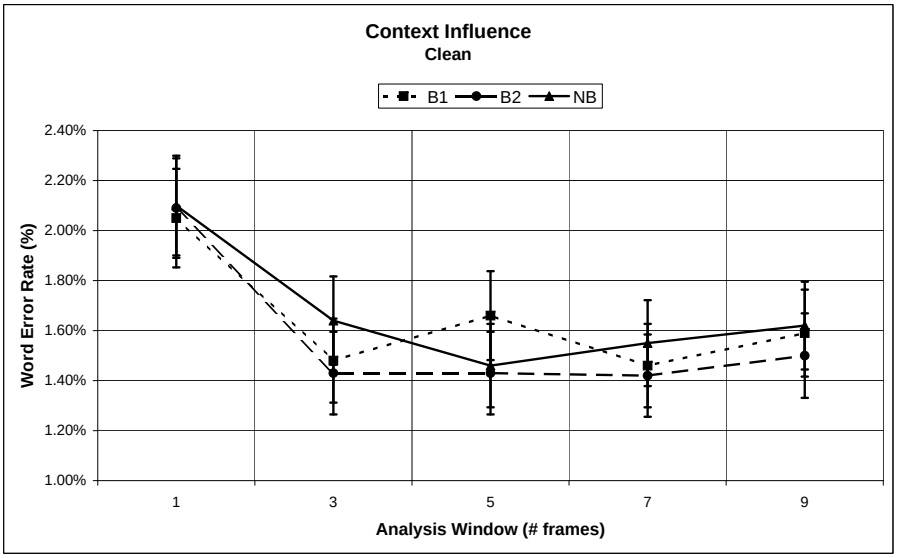

Fig. 10. *Evolution of the WER with the Context Length in a Clean Environment. Vertical segments represent 95% Confidence Intervals (CI)*

[6] R. Solera-Ureña, J. Padrell-Sendra, D. Martín-Iglesias, A. Gallardo-Antolín, C. Peláez-Moreno, and F. Díaz-de-María, *Progress in Non-linear Speech Processing*, ser. Lecture Notes in Computer Science. Berlin/Heidelberg, Germany: Springer-Verlag, May 2007, vol. 4391, ch. SVMs for Automatic Speech Recognition: A Survey, pp. 190–216.

[7] J. Padrell-Sendra, D. Martín-Iglesias, and F. Díaz-de-María, "Support vector machines for continuous speech recognition," in *Proceedings of the 14th European Signal Processing Conference*, Florence, Italy, September 2006.

[8] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 39, no. 2, pp. 539–550, April 2009.

[9] G. Weiss, B. Zadrozny, and M. Saar-Tsechansky, "Guest editorial: special issue on utility-based data mining," *Data Mining and Knowledge Discovery*, vol. 17, no. 2, pp. 129–135, October 2008.

[10] V. Garcia, J. Sánchez, and R. Mollineda, *Progress in Pattern Recognition, Image Analysis and Applications*, ser. Lecture Notes in Computer Science (LNCS). Berlin/Heidelberg, Germany: Springer-Verlag, 2007, vol. 4756, ch. An Empirical Study of the Behavior of Classifiers on Imbalanced and Overlapped Data Sets, pp. 397–406.

[11] L. Toth and A. Kocsor, *Artificial Neural Networks: Biological Inspirations - ICANN 2005*, ser. Lecture Notes in Computer Science (LNCS). Berlin/Heidelberg, Germany: Springer-Verlag, 2005, vol. 3696, ch. Training HMM/ANN Hybrid Speech Recognizers by Probabilistic Sampling, pp. 597–603.

[12] T. Kamm and G. Meyer, "Automatic selection of transcribed training material," in *IEEE Workshop on Automatic Speech Recognition and*

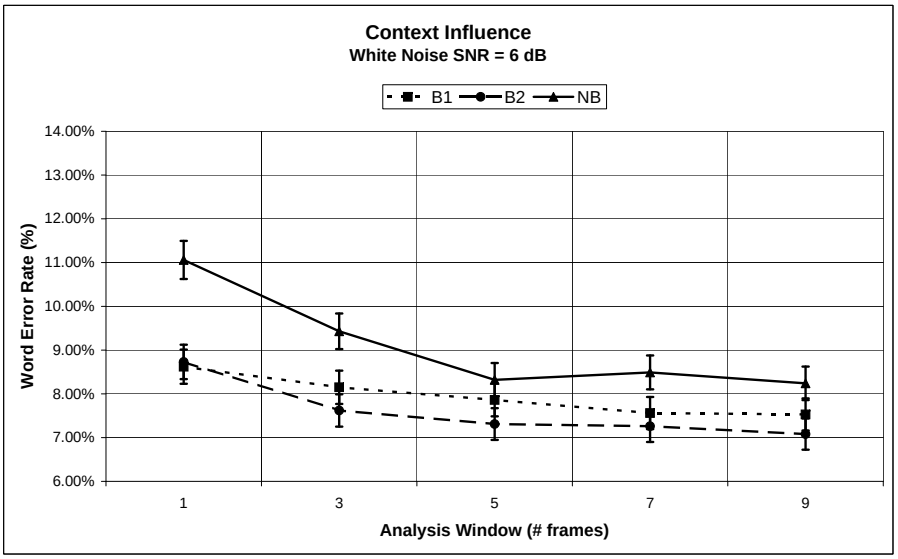

Fig. 11. *Evolution of the WER with the Context Length in a White Noise Environment with SNR = 6 dB. Vertical segments represent 95% Confidence Intervals (CI)*

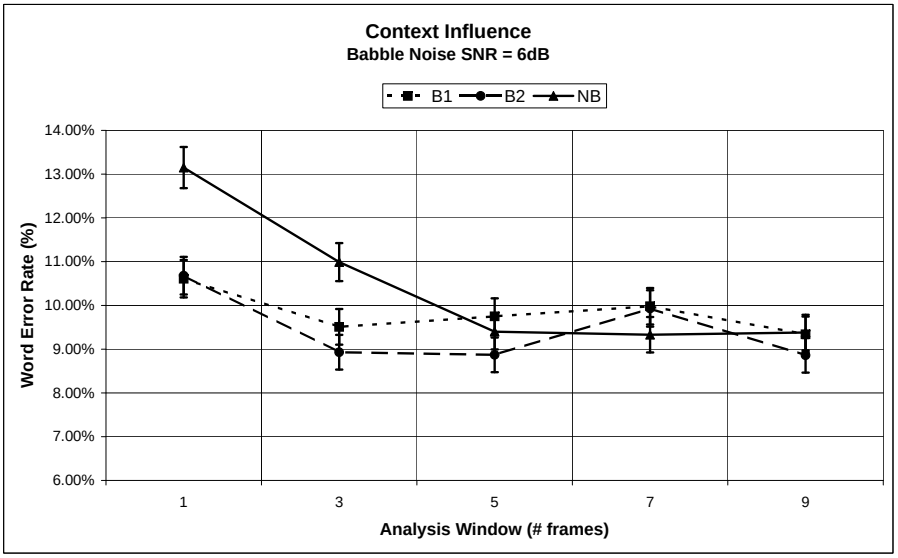

Fig. 12. *Evolution of the WER with the Context Length in a Babble Noise Environment with SNR = 6 dB. Vertical segments represent 95% Confidence Intervals (CI)*

Understanding 2001 (ASRU'01), Trento, Italy, December 2001, pp. 417–420.

[13] ——, "Selective sampling of training data for speech recognition," in *Proceedings of the 2nd international conference on Human Language Technology Research*. San Diego, California (USA): Morgan Kaufmann Publishers Inc., March 2002, pp. 20–24.

[14] T. Kamm, "Active learning for acoustic speech recognition modeling," Ph.D. dissertation, John Hopkins University, Baltimore, Maryland (USA), January 2004. [Online]. Available: http://www.clsp.jhu.edu/people/tkamm/papers/kamm_thesis.pdf

[15] Y. Wu, R. Zhang, and A. Rudnicky, "Data selection for speech recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'07)*, Kyoto, Japan, December 2007, pp. 562–565.

[16] A. Nagroski, L. Boves, and H. Steeneken, "In search of optimal data selection for training of automatic speech recognition systems," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'03)*, Virgin Islands, USA, November-December 2003, pp. 67–72.

[17] B. Chen, Ö. Çetin, G. Doddington, N. Morgan, M. Ostendorf, T. Shinozaki, and Q. Zhu, "A CTS task for meaningful fast-turnaround experiments," in *Proceedings of Rich Transcription Fall Workshop*, Palisades, New York (USA), November 2004.

[18] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: a Hybrid Approach*. Norwell, MA (USA): Kluwer Academic Publishers, 1994.

[19] N. Morgan and H. Bourlard, "Continuous speech recognition: An introduction to the hybrid HMM/connectionist approach," *IEEE Signal Processing Magazine*, vol. 12, no. 3, pp. 24–42, May 1995.

[20] P. Pujol, S. Pol, C. Nadeu, A. Hagen, and H. Bourlard, "Comparison and combination of features in a hybrid HMM/MLP and a HMM/GMM speech recognition system," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 14–22, January 2005.

[21] L. Tóth and A. Kocsor, "A segment-based interpretation of HMM/ANN hybrids," *Computer Speech & Language*, vol. 21, no. 3, pp. 562–578, July 2007.

[22] R. Gemello, F. Mana, S. Scanzio, P. Laface, and R. D. Mori, "Linear hidden transformations for adaptation of hybrid ANN/HMM models," *Speech Communication*, vol. 49, no. 10-11, pp. 827–835, October 2007.

[23] J. Stadermann and G. Rigoll, "Hybrid NN/HMM acoustic modeling techniques for distributed speech recognition," *Speech Communication*, vol. 48, no. 8, pp. 1037–1046, August 2006.

[24] J. Frankel and S. King, "A hybrid ANN/DBN approach to articulatory feature recognition," in *Proceedings of the 9th International Conference on Speech Communication and Technology (Interspeech'2005)*, Lisboa, Portugal, September 2005.

[25] H. Bourlard and N. Morgan, *Adaptive Processing of Sequences and Data Structures*, ser. Lecture Notes in Computer Science (LNCS). Berlin/Heidelberg, Germany: Springer-Verlag, 1998, vol. 1387, ch. Hybrid HMM/ANN Systems for Speech Recognition: Overview and New Research Directions, pp. 389–417.

[26] J. Neto, C. Martins, and L. Almeida, "An incremental speaker-adaptation technique for hybrid hmm-mlp recognizer," in *The Fourth International Conference on Spoken Language Processing (ICSLP'96)*, Philadelphia, USA, 1996, pp. 1289–1292.

[27] ——, "Speaker-adaptation in a hybrid hmm-mlp recognizer," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'96)*, vol. 6, Washington, DC, USA, May 1996, pp. 3382–3385.

[28] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Upper Saddle River, New Jersey, (USA): Prentice Hall, 1998.

[29] S. Renals, N. Morgan, H. Bourlard, M. Cohen, and H. Franco, "Connectionist probability estimators in HMM speech recognition," *Speech and Audio Processing*, vol. 2, no. 1, pp. 161–174, Jan 1994.

[30] A. Hagen, "Robust speech recognition based on multi-stream processing," PhD Thesis, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, December 2001. [Online]. Available: http://infoscience.epfl.ch/search.py?recid=32973

[31] C. Raymond and G. Riccardfi, "Learning with noisy supervision for spoken language understanding," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'08)*, Las Vegas, USA, March-April 2008, pp. 4989–4992.

[32] I. Jars and F. Panaget, "Improving spoken language understanding with information retrieval and active learning methods," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'08)*, Las Vegas, USA, March-April 2008, pp. 5001–5004.

[33] A. Bondu, V. Lemaire, and B. Poulain, *Advances in Data Mining: Theoretical Aspects and Applications*, ser. Lecture Notes in Artificial Intelligence (LNAI). Berlin/Heidelberg, Germany: Springer-Verlag, 2007, vol. 4597, ch. Active Learning Strategies: A Case Study for Detection of Emotions in Speech, pp. 597–603.

[34] D. Farris, C. White, and S. Khudanpur, "Sample selection for automatic language identification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'08)*, Las Vegas, USA, March-April 2008, pp. 4225–4228.

[35] S. Vijayakumar and H. Ogawa, "Improving generalization ability through active learning," *IEICE Transactions on Information and Systems*, vol. E82-D, no. 2, pp. 480–487, February 1999.

[36] G. Tur, D. Hakkani-Tür, and R. Schapire, "Combining active and semi-supervised learning for spoken language understanding," *Speech Communication*, vol. 45, no. 2, pp. 171–186, February 2005.

[37] G. Riccardi and D. Hakkani-Tür, "Active learning: theory and applications to automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 504–511, July 2005.

[38] R. Zhang and A. Rudnicky, "A new data selection approach for semi-supervised acoustic modeling," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'06)*, vol. 1, Toulouse, France, May 2006, pp. 421–424.

[39] S.-H. Liu, F.-H. Chu, S.-H. Lin, H.-S. Lee, and B. Chen, "Training data selection for improving discriminative training of acoustic models," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'07)*, Kyoto, Japan, December 2007, pp. 284–289.

[40] H. Jiang, F. Soong, and C.-H. Lee, "A dynamic in-search data selection method with its applications to acoustic modeling and utterance verification," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 945–955, September 2005.

[41] C. Peláez-Moreno, Q. Zhu, B. Chen, and N. Morgan, "Automatic data selection for MLP-based feature extraction for ASR," in *Proceedings of the 9th International Conference on Speech Communication and Technology (Interspeech'2005)*, Lisboa, Portugal, September 2005, pp. 229–232.

[42] N. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: special issue on learning from imbalanced data sets," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 1–6, June 2004.

[43] A. Moreno, "Speechdat spanish database for fixed telephone network," Universitat Politecnica de Catalunya, Tech. Rep., 1997.

[44] G. Batista, R. Prati, and M. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, June 2004.

[45] T. Jo and N. Japkowicz, "Class imbalances versus small disjuncts," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 40–49, June 2004.

[46] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: a review," *GESTS International Transactions on Computer Science and Engineering*, vol. 30, no. 1, pp. 25–36, 2006.

[47] A. Joshi, "Applying the wrapper approach for auto discovery of under-sampling and over-sampling percentages on skewed datasets," PhD Thesis, University of South Florida, Tampa, Florida (USA), November 2004. [Online]. Available: http://purl.fcla.edu/fcla/etd/SFE0000491

[48] K. Veropoulos, C. Campbell, and N. Cristianini, "Controlling the sensitivity of support vector machines," in *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI99)*, Stockholm, Sweden, August 1999, pp. 55–60.

[49] N. Chawla, D. Cieslak, L. Hall, and A. Joshi, "Automatically countering imbalance and its empirical relationship to cost," *Data Mining and Knowledge Discovery*, vol. 17, no. 2, pp. 225–252, October 2008.

[50] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," in *Proceedings of the Fourteenth International Conference on Machine Learning*. Morgan Kaufmann, 1997, pp. 179–186.

[51] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, February 2002.

[52] A. Estabrooks, T. Jo, and N. Japkowicz, "A multiple resampling method for learning from imbalanced data sets," *Computational Intelligence*, vol. 20, no. 1, pp. 18–36, February 2004.

[53] Z.-H. Zhou and X.-Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 1, pp. 63–77, January 2006.

[54] A. Varga, H. Steenneken, M. Tolimson, and D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," Defence Evaluation and Research Agency (DERA), Speech Research Unit, Malvern, United Kingdom, Tech. Rep., 1992.

[55] F. Johansen, N. Warakagoda, B. Lindberg, G. Lehtinen, Z. Ka-Ci-Ce, K. Elenius, and G. Salvi, "The COST 249 SpeechDat multilingual reference recogniser," in *COST 249 MCM, Technical Annex*, 1999.