

Tuning Support Vector Machines for Robust Phoneme Classification with Acoustic Waveforms

Jibran Yousafzai¹, Zoran Cvetković¹, Peter Sollich²

¹Department of Electronic Engineering and ²Department of Mathematics,
King's College London, WC2R 2LS, UK

[jibran.yousafzai, zoran.cvetkovic, peter.sollich] @kcl.ac.uk

Abstract

This work focuses on the robustness of phoneme classification to additive noise in the acoustic waveform domain using support vector machines (SVMs). We address the issue of designing kernels for acoustic waveforms which imitate the state-of-the-art representations such as PLP and MFCC and are tuned to the physical properties of speech. For comparison, classification results in the PLP representation domain with cepstral mean-and-variance normalization (CMVN) using standard kernels are also reported. It is shown that our custom-designed kernels achieve better classification performance at high noise. Finally, we combine the PLP and acoustic waveform representations to attain better classification than either of the individual representations over the entire range of noise levels tested, from quiet condition up to -18dB SNR.

Index Terms: Kernels, Phoneme classification, Robustness, Support vector machines

1. Introduction

Automatic speech recognition (ASR) systems lack the level of robustness inherent to human speech recognition (HSR). This has a detrimental effect when these systems are operated in adverse acoustical environments, while humans can still recognize isolated speech units above the level of chance at -18dB SNR, and significantly above it at -9dB SNR [1]. No ASR system achieves performance close to that of human auditory system under such severe noise. While language and context modelling are essential for reducing many errors in speech recognition, accurate recognition of phonemes and the related problem of classification of isolated phonetic units is a major step towards achieving robust recognition of continuous speech.

State-of-the-art ASR front-ends are mostly some variant of Mel-frequency cepstral coefficients (MFCC) or Perceptual Linear Prediction (PLP) [2]. These representations are derived from the short term magnitude spectra followed by non-linear transformations to model the processing of the human auditory system. They remove variations from speech signals that are considered unnecessary for recognition while preserving the information content. This allows for more accurate modelling when the data is limited. However it is not known whether, by reducing the dimension significantly, one also discards some of the information that makes speech such a robust message representation. To make these state-of-the-art representations of speech less sensitive noise, several methods have been proposed to reduce explicitly the effect of noise on spectral representations [3] in order to approach the optimal performance which is achieved when training and testing conditions are matched [4].

The alternative approach investigated in this paper is the use of high-dimensional acoustic waveform representations for

robust classification in the presence of additive noise. PLP/MFCC are designed in a way that removes non-lexical invariances (sign, time alignment); however, for classification in the acoustic waveform domain these invariances need to be taken into account by incorporating them in the kernel. By doing this, in combination with straightforward noise adaptation in the kernel, classification performance can be made rather robust to noise [5]. In a key refinement, we try to capture the idea of time derivatives of cepstral features, which measure the rate of change of features and hence represent the dynamics of a speech signal. This is done by embedding variations in signal across a phoneme into the kernel. Our experiments demonstrate the effectiveness of the kernels tuned for acoustic waveforms under adverse conditions. For comparison, classification results in the PLP representation domain using standard SVM kernels are also reported. While this study is focused on phoneme classification for comparison of the acoustic waveform and PLP representations of speech, we believe the results also have implications for the construction of ASR systems as SVMs can be extended to handle the task of continuous speech recognition as detailed in [6, 7].

The SVM approach to classification of phonemes using error-correcting output codes (ECOC) [8] is reviewed briefly in Section 2. Kernel design for the classification task in the acoustic waveform domain is described in Section 3, where we also address the issues regarding the noise adaptation for both the PLP and acoustic waveform features. The classification results in both representation domains are reported in Section 4. It is shown that the acoustic waveforms perform better in severe noisy conditions. We also use the method of [5] to combine the PLP and acoustic waveform representations for improved accuracy. Finally, Section 5 has conclusions and an outlook towards future work.

2. Classification using SVMs

In this paper, we use support vector machines [9] to estimate non-linear decision surfaces separating two classes of data. A kernel-based decision function which classifies an input vector \mathbf{x} is expressed as

$$h(\mathbf{x}) = \sum_i \alpha_i y_i \langle \varphi(\mathbf{x}), \varphi(\mathbf{x}_i) \rangle + b = \sum_i \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b \quad (1)$$

where φ is a non-linear function that maps data points to high-dimensional feature vectors while $\mathbf{x}_i, y_i = \pm 1$ and α_i , respectively, are the i -th training sample, its class label and its Lagrange multiplier. $K(\mathbf{x}, \mathbf{x}_i)$ is a kernel function that satisfies Mercer's theorem and b is the classifier bias determined by the training algorithm. Two commonly used kernels are the polynomial kernel, $K_p(\mathbf{x}, \mathbf{x}_i) = (1 + \langle \mathbf{x}, \mathbf{x}_i \rangle)^\Theta$, and radial basis

function (RBF) kernel, $K_r(\mathbf{x}, \mathbf{x}_i) = e^{-\Gamma\|\mathbf{x}-\mathbf{x}_i\|^2}$.

To obtain a multiclass classifier, binary SVM classifiers are combined via ECOC [8] methods. A standard approach is to use $K(K-1)/2$ pairwise classifiers, each trained to distinguish two of the K classes. For a test point \mathbf{x} , we then predict the class k for which $d_k(\mathbf{x}) = \sum_{l=1, l \neq k}^K \xi(h_{kl}(\mathbf{x}))$ is minimized, where ξ is some loss function and $h_{kl}(\mathbf{x})$ is the output of the classifier trained to distinguish classes k and l , with sign chosen so that a positive sign indicates class k . We compared a number of loss functions $\xi(h)$; the hinge loss $\xi(h) = \max(1-h, 0)$ performed best and is used throughout this paper.

3. Kernels for Acoustic Waveforms

A crucial aspect of SVMs is the choice of the kernel function for classification of non-separable data. The use of appropriate kernels that express prior knowledge about the physical properties of the data can improve the performance significantly. In [5], we embedded some key properties of speech in standard SVM kernels. Here, we try to capture the dynamics of speech over a finer timescale in a manner similar to the time derivatives of the cepstral coefficients. This is done by evaluating the kernel over T non-overlapping subsegments of data points as

$$K_{n,s}(\mathbf{x}, \mathbf{x}_i) = \frac{\sum_{u,v=-n}^n \sum_{t=1}^T K_l(\mathbf{x}_t, \mathbf{x}_{i,t}) K_e(\mathbf{x}_t^{u\Delta}, \mathbf{x}_{i,t}^{v\Delta})}{(2n+1)^2}, \quad (2)$$

where

$$K_l(\mathbf{x}_t, \mathbf{x}_{i,t}) = e^{-\frac{(\log\|\mathbf{x}_t\|^2 - \log\|\mathbf{x}_{i,t}\|^2)^2}{2a^2}}, \quad (3)$$

K_e is the *even-polynomial* kernel described in [5], Δ is the shift increment, $[-n\Delta, n\Delta]$ is the shift range, \mathbf{x}_t and $\mathbf{x}_{i,t}$ are the t^{th} subsegments of the test waveform \mathbf{x} and the i^{th} training sample, \mathbf{x}_i respectively. $\mathbf{x}_t^{u\Delta}$ is a subsegment of the same length as the original subsegment \mathbf{x}_t^0 but extracted from a position shifted by $u\Delta$ samples. This kernel is sensitive to correlation between the individual subsegments of the phonemes and gives information about the phoneme energy over a finer resolution, which can help to distinguish phoneme classes with different energy profiles as shown in Figure 1. Since PLP, MFCC and other state-of-the-art representations are based on short-time magnitude spectra and contain information about the energy, using similar custom-designed kernels for classification in the PLP domain will not have any advantage over the standard (polynomial or RBF) kernels.

In order to improve the robustness, both PLP and acoustic waveform features are adapted to noise. Since the noise variance, σ^2 can be estimated during pause intervals (non-speech activity) between speech signals, we assume that its value is known. In the case of acoustic waveforms, both the features and the kernel are adapted to noise. First, the test data is transformed using DCT and each frequency component is scaled by $1/\sqrt{1+q\sigma_r^2}$ where q is the number of frequency components and σ_r^2 is the noise variance of the r^{th} frequency component. The data is transformed back to time domain using IDCT and normalized to $\sqrt{1+\sigma^2}$ whereas the training data is set to have a unit norm for computation of the inner product in the polynomial kernel. This is done to keep the norm of the speech roughly independent of the noise. It should be noted that the frequency scaling would have no effect on the structure of the noisy test waveform in the presence of white

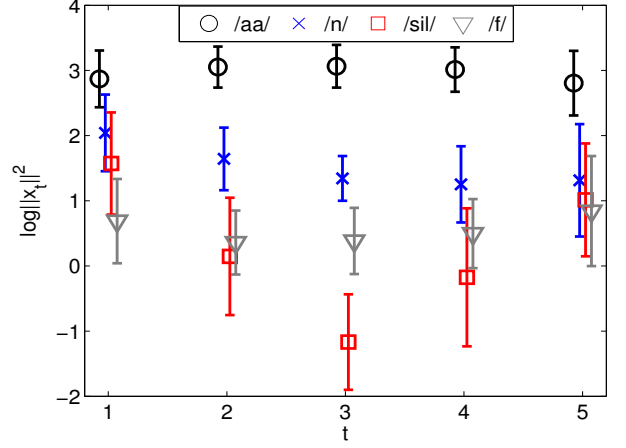


Figure 1: Time profile (mean \pm standard deviation) of log energies of $T = 5$ subsegments (20ms each) of phoneme classes /aa/, /n/, /sil/ and /f/.

Gaussian noise. Explicitly, for a test waveform \mathbf{x} and training waveform \mathbf{x}_i , let $\tilde{\mathbf{x}} = \hat{\mathbf{x}}\sqrt{1+\sigma^2}/\|\hat{\mathbf{x}}\|$ and $\tilde{\mathbf{x}}_i = \mathbf{x}_i/\|\mathbf{x}_i\|$ where $\hat{\mathbf{x}}$ is the waveform obtained from the frequency scaling of \mathbf{x} . Then the baseline polynomial kernel for the normalized waveforms is $\tilde{K}_p(\mathbf{x}, \mathbf{x}_i) = (1 + \langle \tilde{\mathbf{x}}, \tilde{\mathbf{x}}_i \rangle)^\Theta$, and $K_e(\mathbf{x}, \mathbf{x}_i) = \tilde{K}_p(\mathbf{x}, \mathbf{x}_i) + \tilde{K}_p(\mathbf{x}, -\mathbf{x}_i)$.

Under the assumption that speech and noise are uncorrelated, a similar adaptation as in the polynomial kernel can be used for K_l . By subtracting the estimated noise variance (σ^2) from the energy of the noisy subsegments \mathbf{x}_t , energies of the clean subsegments can be approximated. We, therefore, use this adapted energy of the subsegments in evaluating (3), giving

$$K_l(\mathbf{x}_t, \mathbf{x}_{i,t}) = e^{-\frac{(\log\|\mathbf{x}_t\|^2 - \sigma^2 - \log\|\mathbf{x}_{i,t}\|^2)^2}{2a^2}}. \quad (4)$$

As training for acoustic waveforms is performed in quiet conditions, noise adaption of the training data \mathbf{x}_i is not required. The absolute value of the subtracted energy is used to catch the cases when the overlap of speech and noise is negative enough to overwhelm the energy of clean speech. There are two important issues to be addressed when using (2) in the presence of noise: (a) *Normalization of the subsegments* - the use of (2) requires normalization of the clean subsegments to unit norm and of the noisy ones to $\sqrt{1+\sigma^2}$. For short subsegments there can however be wide variation in local SNR in spite of the fixed global SNR, and so this normalization may not be in accordance with the local SNR. (b) *Orthogonality* - using short (lower dimensional) subsegments makes fluctuations away from our assumed orthogonality of speech and noise more likely. To address these issues, we also consider a kernel $K_{n,u}$ given by

$$K_{n,u}(\mathbf{x}, \mathbf{x}_i) = \left(\sum_{t=1}^T K_l(\mathbf{x}_t, \mathbf{x}_{i,t}) \right) \frac{\sum_{u,v=-n}^n K_e(\mathbf{x}_t^{u\Delta}, \mathbf{x}_{i,t}^{v\Delta})}{(2n+1)^2}. \quad (5)$$

Here the time-correlation part of the kernel is left unsegmented, while K_l is still evaluated over T subsegments of the phonemes. It is expected that $K_{n,s}$ gives significantly better performance than $K_{n,u}$ in less noisy conditions because of its sensitivity to the correlation of the individual subsegments. On the other

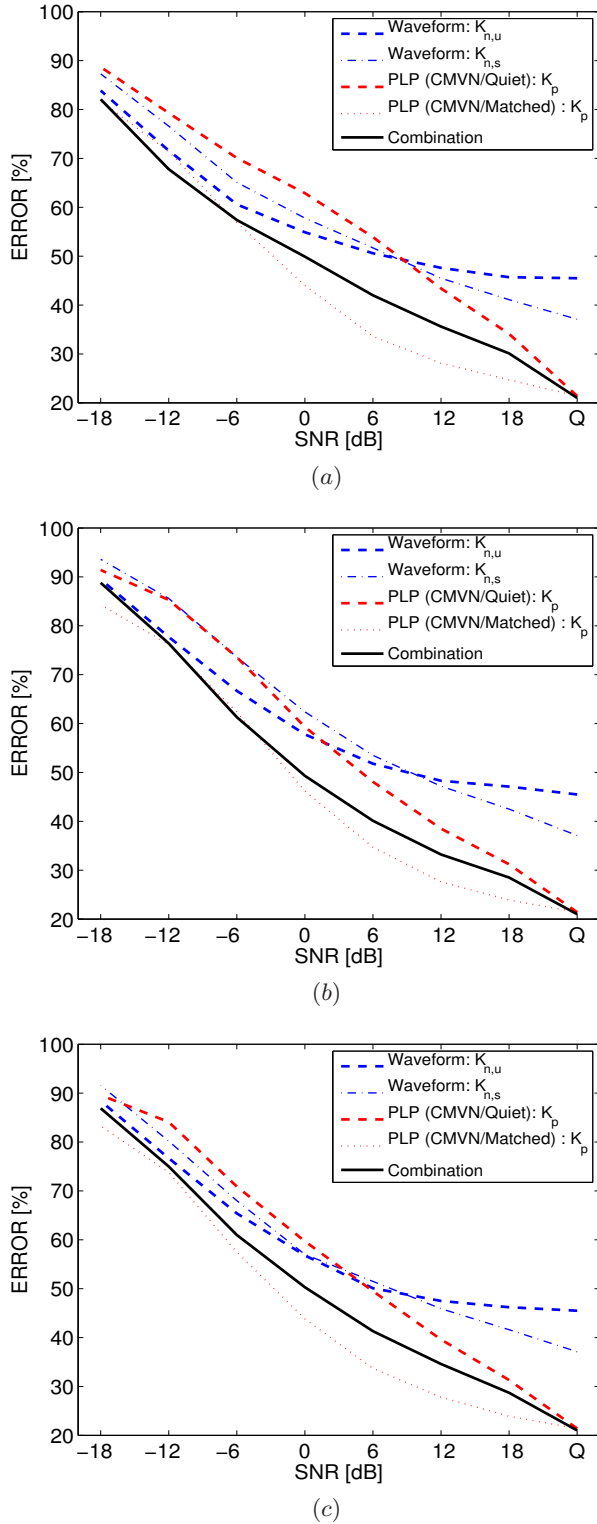


Figure 2: Classification in the presence of (a) white Gaussian noise (b) speech-weighted noise [10] (c) pink noise from NOISEX-92: PLP with CMVN using K_p trained in quiet and matched conditions, waveforms using $K_{n,s}$ and $K_{n,u}$ kernels and the combination of waveforms ($K_{n,u}$) and PLP (CMVN) trained in quiet conditions.

hand, $K_{n,s}$ performs worse than $K_{n,u}$ at high noise due to the two limitations discussed above. It should be noted that the classification performance with kernel $K_{n,s}$ in quiet conditions can be improved even more by increasing the number of subsegments, T .

For the PLP representations, the features are standardized, i.e. scaled and shifted to have zero mean and unit variance per sentence. The optimal performance with PLP is obtained under matched training and test conditions [4]. However, this is an impractical target which could be achieved only if one had access to a large set of classifiers trained for different noise types and levels. Therefore, in order to have a fair comparison of PLP with acoustic waveforms, we use classifiers trained in quiet conditions, with feature vectors of the test data adapted to noise using cepstral mean and variance normalization (CMVN) [3], a noise compensation technique that modifies the cepstral coefficients in order to minimize the mismatch between the training and test data. Here, cepstral features are standardized on each noisy test sentence [3].

In the next section, we show the performance of these kernels for the phoneme classification task in the acoustic waveform domain. Classification in the PLP domain using standard SVM kernels is used as a benchmark for comparison with acoustic waveforms.

4. Results

Experiments are performed on the 'si' and 'sx' sentences of TIMIT database [11]. The training set consists of 3696 sentences from 168 different speakers. The core set is used for testing which consists of 192 sentences from 24 different speakers not included in the training set. We remove the glottal stops /q/ from the labels and fold certain allophones into their corresponding phonemes using the standard Kai-Fu Lee clustering [12], resulting in a total of 48 phoneme classes. Among these classes, there are 7 groups for which the contribution of within-group confusions towards multiclass error is not counted [12].

Regarding the SVM classifiers for acoustic waveform representation, results are reported for $K_{n,s}$ and $K_{n,u}$ with \tilde{K}_p as a baseline kernel. For the PLP representation, comparable performance is obtained with K_p and K_r so we show results for the former. Fixed hyperparameter values are used throughout for training binary SVMs: the degree of K_p , $\Theta = 6$ and the penalty parameter $C = 1$.

For the acoustic waveform representation, phoneme segments are extracted from the TIMIT sentences by applying a 100 ms rectangular window at the center of each phoneme waveform (of variable length), which at 16 kHz sampling frequency gives fixed length vectors in \mathbb{R}^{1600} . In the evaluation of (2) and (5), we use a shift increment of $\Delta = 100$ samples (≈ 6 ms) over a shift range ± 100 (so that $n = 1$), giving three shifted segments of length 1400 samples each. Each of these segments is broken into $T = 5$ subsegments of equal length. The value of a is set to 0.5 for both kernels. For the PLP representation, we convert each waveform into a sequence of 13 dimensional feature vectors, their time derivatives and second order derivatives which are combined into a sequence of 39 dimensional feature vectors. Then, the 9 frames (with frame duration of 25ms and a frame rate of 100 frames/sec) closest to the center of a particular phoneme are concatenated to give a representation in \mathbb{R}^{351} .

In this study, we focus on investigating robustness in the presence of additive white Gaussian noise, pink noise from NOISEX-92 database and speech-weighted noise [10]. To test the classification performance of PLP and acoustic waveforms

in noise, each sentence is normalized to unit energy per sample and then a noise sequence with variance σ^2 (per sample) is added to the entire sentence. It should be noted that SNR at the sentence level is fixed but SNR at the level of individual phonemes will vary widely.

Classification results using SVMs for the PLP and acoustic waveform representations in the presence of additive white Gaussian noise are shown in Figure 2(a). For acoustic waveforms, classification results with kernels $K_{n,s}$ and $K_{n,u}$ are presented whereas K_p is used for classification of PLP features. One observes that a PLP classifier trained on clean data gives very good performance when tested on clean data i.e. 21% error. (We achieve slightly better performance than [13] due to different cepstral representations.) But at 0dB SNR, we get an error of 63% even with CMVN. This can now be contrasted with the results of acoustic waveform classifiers. Classification with kernel $K_{n,u}$ exhibits a more robust behavior to noise and achieves improvements over PLP for noise levels above a crossover point between 12dB and 6dB SNR. The largest improvement over PLP, of 10% is achieved by $K_{n,u}$ at -6dB SNR. It is interesting to observe that the acoustic waveform classifier performs better in high noise compared to the PLP classifier even though the time derivatives and second order derivatives of the PLP features use significantly more information about the signal from the frames of the adjacent phonemes.

In a comparison of kernels for the acoustic waveforms, $K_{n,s}$ achieves an 8% improvement in quiet conditions over $K_{n,u}$ because $K_{n,s}$ is sensitive to correlation between the individual phoneme subsegments. However, $K_{n,u}$ performs better in high noise e.g. $K_{n,u}$ achieves a 4% average improvement over $K_{n,s}$ between 6dB and -18dB SNRs. This is due to the limitations of $K_{n,s}$ in high noise as discussed in Section 3. Similar conclusions can be drawn for classification in the presence of speech-weighted noise and pink noise as shown in Figure 2(b)-(c).

In our previous work [5], we established that the convex combination of the PLP and waveform classifiers attains better classification performance than either of the individual representations. As waveform classifiers with kernel $K_{n,u}$ achieve significantly better results in high noise, therefore we consider its combination with the decision values of the PLP classifiers trained in quiet conditions. In Figure 2, we compare the classification performance in the PLP and acoustic waveform domains with the combined classifier for different noise types. One observes that the combined classifier often performs better or at least as well as the individual classifiers. It should be noted that the combined classifier is not simply a hard-switch between the two representations and a genuine improvement in performance is achieved by this combination. Although the combined classifier does not achieve the impractical target of PLP classifier trained and tested in matched conditions for $\text{SNR} > -6\text{dB}$ as shown in Figure 2, the gain in classification accuracy is significant compared to a standalone PLP classifier with CMVN. For instance in presence of white Gaussian noise (Figure 2(a)), the combined classifier achieves an average of 11% less error than the PLP classifier trained on clean data with CMVN for $-12\text{dB} \leq \text{SNR} \leq 12\text{dB}$.

5. Conclusions

The robustness of phoneme classification to additive noise in the PLP and acoustic waveform domains was investigated using SVMs. We observe that embedding invariances and variations in the signal energy across a phoneme into the kernel can sig-

nificantly improve the classification performance. While PLP representation allows very accurate classification of phonemes especially for clean data, its performance suffers severe degradation at high noise. On the other hand, the high-dimensional acoustic waveform representation, although not as accurate as PLP classification on clean data, is more robust in severe noise. Finally, the convex combination of classifiers achieves performance that is consistently better than both individual domains across the entire range of SNRs. Our preliminary experiments on the use of multi-layered, multi-category SVMs [14] have shown to improve the performance even more in high noise. This work can be steered in a number of different directions. An important and necessary extension of this work would be to investigate the robustness of these representations to linear filtering. It would also be interesting to extend our work to handle continuous speech recognition tasks using SVM/HMM hybrids [6, 7].

6. References

- [1] G. Miller and P. Nicely, "An Analysis of Perceptual Confusions among some English Consonants," *J. of the Acous. Soc. of America*, vol. 27, no. 2, pp. 338–352, 1955.
- [2] H. Hermansky, "PLP Analysis of Speech," *J. of the Acous. Soc. of America*, vol. 87, no. 4, pp. 1738–1752, 1990.
- [3] O. Viikki and K. Laurila, "Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition," *Speech Comm.*, vol. 25, pp. 133–147, 1998.
- [4] M. Gales and S. Young, "Robust Continuous Speech Recognition using Parallel Model Combination," *IEEE Trans. on Speech and Audio Proc.*, pp. 352–359, 1996.
- [5] J. Yousafzai, Z. Cvetković, and P. Sollich, "Custom-Designed SVM Kernels for Improved Robustness of Phoneme Classification," *EUSIPCO*, 2009.
- [6] N. Smith and M. Gales, "Speech Recognition using SVMs," in *Advances in Neural Information Processing Systems 14*. 2002, pp. 1197–1204, MIT Press.
- [7] S. E. Krger, M. Schaffner, M. Katz, E. Andelic, and A. Wendemuth, "Speech recognition with Support Vector Machines in a Hybrid System," in *In Proc. EuroSpeech, 2005*, 2005, pp. 993–996.
- [8] T. Dietterich and G. Bakiri, "Solving Multiclass Learning Problems via Error-Correcting Output Codes," *J. of AI Research*, vol. 2, pp. 263–286, 1995.
- [9] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.
- [10] S. Phatak and J. B. Allen, "Consonant and Vowel Confusions in Speech-weighted Noise," *J. of the Acous. Soc. of America*, vol. 121, Apr. 2007.
- [11] W. Fisher, G. Doddington, and K. Goudie-Marshall, "DARPA Speech Recognition Research Database," *DARPA Sp. Recogn. Workshop*, pp. 93–99, 1986.
- [12] K. F. Lee and H. W. Hon, "Speaker-Independent Phone Recognition Using Hidden Markov Models," *IEEE Trans. Ac. Speech Sig. Proc.*, vol. 37, no. 11, 1989.
- [13] P. Clarkson and P. Moreno, "On the Use of Support Vector Machines for Phonetic Classification," *ICASSP*, 1999.
- [14] K. Crammer and Y. Singer, "On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines," *J. Mach. Learn. Res.*, vol. 2, pp. 265–292, 2002.