



SVM Based Feature Extraction in Speech Synthesis

Peter Cahill, Jan Macek, Julie Carson-Berndsen

School of Computer Science and Informatics, University College Dublin, Dublin, Ireland

peter.cahill@ucd.ie, jan.macek@ucd.ie, julie.berndsen@ucd.ie

Abstract

Annotations of speech recordings are a fundamental part of any unit selection speech synthesiser. However, obtaining flawless annotations is an almost impossible task. Manual techniques can achieve the most accurate annotations, provided that enough time is available to analyse every phone individually. Automatic annotation techniques are a lot faster than manual, doing the task in a much more reasonable time frame, but such annotations contain a considerable amount of error. In this paper a technique is introduced that can quite accurately ensure a degree of articulatory-acoustic similarity between annotated units. The synthesiser will encourage the use of units that have been identified to have appropriate articulatory-acoustic parameters, but will not limit the domain of the speech database. This helps to identify where joins can be performed best and also identifies which annotations should be avoided at the phone level.

1. Introduction

Unit selection speech synthesis can produce natural sounding intelligible speech. This is particularly true if the domain of the speech being synthesised is the same as the domain of the recorded speech database being used. It is when out-of-domain words or phrases are used that the quality and naturalness reduces significantly. The reduction of quality is often due to relying on small speech units to construct the target speech. Modern synthesis algorithms are capable of constructing human sounding words from small speech units, however the quality of constructing natural sounding units from smaller units is dependent on the available speech annotations.

As the primary dependency of any unit selection speech synthesiser is the speech database in use, having accurate annotations of the speech data is crucial. Annotations of speech data for synthesis are commonly available in a phone or diphone format. The accuracy of such annotations is variable, and very dependent on the how the annotations were created. Experiments on different annotation techniques show that there will always be quite a significant degree of error [1]. The minimal degree of error is obtained when a detailed manual correction is performed on automatically segmented data. In [1] manual correction to obtain minimal error is described to take approximately 2 minutes for a first pass and 30 minutes for the second pass of correction per utterance. Annotating a full speech database this way (which may be from 1000-10000 utterances) is very resource intensive. It is also quite likely that when doing such a tedious annotation that after a few utterances the segmentors will loose focus and start to annotate less accurately. It is clear from [1] that no matter which annotation technique is used, there will always be a significant amount of error introduced into the synthesis system.

In this paper we introduce an automated language independent technique that identifies inaccurately annotated phones so

that the synthesiser can avoid them at a later stage. Parts of this concept have been previously described in earlier work [2]. In this paper, the concept is developed further and a comparison is performed with the technique described in [2].

The remainder of this paper is structured as follows: Section 2 describes the concepts of how the use of an articulatory-acoustic analysis may assist in the selection of more suitable units for synthesis. Section 3 describes some potential applications of this work as well as the primary application, unit selection speech synthesis. Section 4 contains an analysis of how voices built using the techniques described in this paper compare with our previous work on this topic. Section 5 concludes and describes future work.

2. Hypothesis

Annotations of the speech data in unit selection speech synthesis are commonly in a phonological format, consisting of a unit label (perhaps phoneme or diphone) as well as the temporal endpoints for that particular unit. The unit labels are often estimated automatically by using a grapheme to phoneme or word to phoneme technique on the orthographic transcriptions of the speech database. Such annotations work quite well and are the foundation of many modern synthesis systems. The problem with using this type of annotation is that the unit label is estimated from the orthographic transcriptions. For this reason, if two speech databases were recorded using the same orthographic transcriptions, all that would differ in the resulting annotations would be the temporal endpoints, but it is unlikely that the two speakers would have articulated every basic unit in the database identically.

The use of an independent, second level of annotations is presented in this paper. The second level of annotations is a phonetic analysis of the audio data in the speech database. This form of annotation is intended to be used in a cross validation technique with the more common phonological form of annotation. The aim of this is to improve the consistency of the annotations as well as to automatically identify misaligned, mislabelled or mispronounced units in the speech database.

The phonetic analysis is an analysis of the actual speech data in the database. This is done to give an alternative perspective on the speech data, as the phonological annotations are derived from the orthographic transcriptions only. Some phonetic data has been used previously in speech synthesis, although it was typically used at the joining of units rather than the annotation stage. In cases where phonetic analysis has been used for annotation previously, it was mostly using spectral parameters such as Mel-frequency Cepstral Coefficients (MFCCs), F0 and power in techniques such as [3]. This work focuses on using articulatory-acoustic features rather than the spectral parameters.

As each articulatory-acoustic feature extractor is essentially

checking for a group of spectral properties, the set of feature extractors used will ensure the presence or absence of a number of acoustic properties. When the feature extractors are used, the decision as to whether or not a phone should be marked as preferred is essentially a check on how acoustically similar the phone is to other units with the same phonological label. Units that contain the appropriate acoustic properties are marked as preferred units. During diphone synthesis, the point at which the diphone boundary lies is the same as the point where the articulatory-acoustic feature comparison is performed. This technique results in the end of a diphone having the same articulatory-acoustic features present as the start of the following diphone. Therefore, many of the potentially bad joins are removed before the distance measure or join cost between units needs to be evaluated.

Situations where the feature extractors used are not 100% accurate still perform well, as long as the feature extractor is performing any acoustic analysis. This concept can be extended further than just feature extraction as well, any form of acoustic or spectral analysis can be quite useful in this technique.

This technique has also proved useful for optimisation of unit selection speech synthesis as it can be used to reduce the number of potential units used in the Viterbi search, and can result in a much faster synthesis depending on how many of the phones have been identified as preferred. Some informal tests showed a speed increase of approximately 300%, at the same time as an improvement in output quality. The speed increase is due solely to the use of preferred phones reducing the range of units in the Viterbi search.

2.1. Articulatory-Acoustic Feature Extraction

2.1.1. Articulatory-Acoustic Features

The articulatory-acoustic features were introduced in the speech recognition community as an alternative extension to the acoustic analysis of a speech signal. They help to improve robustness of speech recognition systems used in various uncontrolled environments where performance of traditional speech recognition systems degrades rapidly [4].

Articulatory-acoustic features are thought to be a good compromise in the description of a speech signal, offering a more detailed description of the acoustic signal than phonemes, yet still providing a linguistically interpretable symbolic annotation. Acoustic correlates of features are described in [5].

Machine learning techniques were used to detect the presence or absence of articulatory-acoustic features.

2.1.2. Support Vector Machines

Support Vector Machines (SVMs) learn separating hyperplanes to classify instances in the feature space that are mapped from the input space of the classified data. The mapping from input space to feature space is performed with the application of a kernel on the feature space. The dimension of the feature space is typically much higher than that of the original input space. [6] provides a thorough mathematical background.

The motivation for SVMs comes from the pattern recognition community with mathematical properties of linear classifiers and from the statistical learning theory community with the structural risk minimisation properties of SVMs [7, 8].

For the training of the SVM feature extraction models the TIMIT corpus of read speech was used. 52 values were extracted for every frame of the speech signal, these values were used as inputs for the SVM classifiers. From each frame 12

MFCCs were extracted together with first and second order differences, frequencies of formants (F1-F5) with first order differences, bandwidths of detected formants, and fundamental frequency. The length of the speech signal frames was set to 25 ms and step between two adjacent frames to 10 ms. The original speech signal was sampled at 16 kHz. The distributions of classes vary significantly for different types of features. While the distribution of classes is almost equal (the case of the *vocalic* feature) for half of the articulatory features, in the rest of the cases the positive classes are rare in the data. This has a strong influence on the recall of the positive classes while the overall accuracy remains high.

The training of the SVMs was performed only on one dialect region of the TIMIT corpus (namely dialect region no. 3) mainly for the reasons of time complexity of the training. It has been shown in [9] that the SVMs with second-order polynomial kernels give best performance for the task of articulatory feature recognition and this setting was used throughout the experiments reported in this article.

2.1.3. Performance of SVMs in context

In previous work it has been shown that SVMs outperform other classifiers at the task of articulatory feature recognition [9]. Namely, it gives superior results over hidden Markov models (HMMs), a fact that is analysed in more detail below.

The main distinction between the two approaches, that of SVMs and that of HMMs, is that the former treats the stream of speech signal frames as independent frames and the latter treats them as adjacently dependent. The dependency of adjacent frames is employed in current state-of-the-art speech recognition systems and it is exploited on the phone level. At this level the HMMs model a much larger set of events in the speech signal as opposed to the binary set of feature presence/absence in the case of recognition of articulatory features. This limits the possibility of constructing reliable articulatory feature models with HMMs as the dependency between adjacent frames can not be utilised in the case of binary classification.

In the approach taken with SVMs, only information from the processed speech frame is used. As this might be limiting in some tasks, in the case of articulatory feature recognition it is more detrimental to the performance when the model tries to capture non-existent dependencies in the data as in the HMM based approach. The performances of both of the methods are compared in Section 4.

2.2. Phonetic annotations

Previous work [2] investigated the use of articulatory-acoustic feature extraction in speech synthesis. In this article the idea has been developed further. We investigate the use of SVM based feature extraction as well as the HMM based feature extraction previously used.

Features are extracted on all audio data in the speech database. The features present at the diphone joining point of each phone are examined. For each phone label, the features present are compared with all other phones with the same label. The aim is to identify most common set of features present and absent at the diphone joining point for each phone. The comparison of the most common set of features present is explicit.

When the phones that have the most common features present at their diphone joining point are identified, they are then considered to be preferred phones. This is done so that at a later stage, when the synthesiser is performing the synthesis, it can try to use as many preferred phones as possible.

Furthermore, the concept of using both SVM and HMM for a speech synthesis database is introduced. Features are extracted using both SVM and HMM models for a speech database. The types of features being identified are the same for both the SVM extractors and the HMM extractors. The SVM and HMM models are used independently to improve the robustness of the technique.

The preferred phones identified by the HMMs are organised into sets, resulting in the set of all preferred phones of label p being identified as A_p . Similarly the set of preferred phones identified by the SVM based feature extractors is identified as B_p . The aim is to find the set C_p of phones that have been identified as preferred by both A_p and B_p , such that

$$C_p = A_p \cap B_p \quad (1)$$

where C_p is the intersection of A_p and B_p . The elements of C_p are then given the highest phone priority possible.

The use of the set C_p is essentially a cross validation on the preferred phones. This is performed in an attempt to identify the most acoustically similar units. The use of both SVMs and HMMs is an attempt to maximise consistency by verifying the preferred phones.

We expect that the set of units C_p is best used in larger speech databases, as in a relatively small speech database too many units may be pruned from the speech database. Although it will also be significantly dependent on how many articulatory-acoustic feature types are being extracted. Identifying the thresholds of optimal database size for this technique has yet to be identified. Current work suggests it varies significantly between speech databases; furthermore, the phone set in use will also have considerable influence on this.

3. Applications

The technique described was designed for use in a unit selection diphone synthesiser. The synthesiser uses a Viterbi search on all possible sequences of appropriate diphones to identify the ideal path of units. The synthesiser uses the Mahalanobis distance measure to estimate joint cost between the two vectors (\vec{x}, \vec{y}) , such that:

$$D(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T \Sigma^{-1} (\vec{x} - \vec{y})} \quad (2)$$

where Σ is the covariance matrix, and \vec{x} and \vec{y} are the vectors at the diphone joining point of different phones. The vectors contain Mel-frequency cepstral coefficients (MFCCs), F_0 and power. For further information on using this distance measure see [10, 11].

The comparison of the phonetic annotations, which are the articulatory-acoustic features in the case of this study, is always done at the diphone boundaries of phones. This ensures that when two units are being considered for a join by the Mahalanobis distance measure, both units have previously been classified as acoustically similar units at that exact join point.

Voice data for the synthesiser was compiled into a single file. Building of the voice was a fully automatic process, where the orthographic transcriptions and the speech recordings were input, all necessary processing was performed on the speech data and the result was output as a single voice file. Annotations in the voice data are stored in Unicode, using IPA characters when labelling phones.

The speech annotations were used in a phonological hierarchy, quite similar to the phonological structure matching technique described in [12].

It is quite likely that this technique could also be useful for other applications than intended. Any system that uses a database of speech recordings may find the phonetic annotations useful. We also expect that this technique may be useful for measuring how good or consistent the phonological annotations for a set of speech recordings are.

4. Testing and Results

The speech database used for testing was the full ATR Blizzard Challenge 2007 speech data. The speech recordings had phonological hierarchies automatically generated using C4.5 decision trees ([13]) trained from the CMUDICT dictionary. The CMUDICT phone set was used with the phoneme labels translated into their IPA equivalents.

Three voices were built from the data. For all of the voices built, all of the data except the extracted features was identical. The first voice, V_{hmm} was the voice using the features extracted by the HMMs to identify the preferred phones. The second voice, V_{svm} was the voice data using the features extracted by the SVM models to identify the preferred phones. The remaining voice, V_{hybrid} was the same voice data as V_{hmm} and V_{svm} , but the set of preferred phones was set to be the intersection of the preferred phones of V_{hmm} and V_{svm} .

The set of feature extractors used for both SVMs and HMMs was: *anterior, consonantal, nasal, vocalic, and voiced*. This set of feature extractors was chosen as tests currently indicate that these are our best performing models. A set of five feature extractors seemed reasonable as in the case of this work they are being used to identify articulatory-acoustic coherent sounds, rather than to identify the phone labels exactly. The respective accuracies of the classifiers on the frame-level for each of the used features are presented in Table 1.

Feature	HMM Accuracy (in %)	SVM Accuracy (in %)
anterior	74.82	91.01
consonantal	82.65	89.10
nasal	89.65	97.92
vocalic	82.16	93.12
voiced	84.93	93.59

Table 1: Performance of classifiers on articulatory-acoustic features.

Of the speech data used, every word aligned in 6559 of the utterances. There was an average of 8756 phones for each phoneme label. This was dispersed over a wide range, where vowels would typically have the most phones - the most for any phone label was 33863 for the [ɪ] phone. The phone with the lowest occurrence was the [ʒ] phone with 159 occurrences.

In the test results, all sets of phones with an A_p label refers to a set of preferred phones identified by the V_{hmm} voice, the B_p label refers to a set of preferred phones identified by the V_{svm} voice, and C_p is the intersecting set as described in Section 2.2. Tables on the data compare the sets that the phones are in. The first column shows the percentage of A_p that is in the C_p set, the second column shows the same data for B_p , and the third column shows a percentage of the quantity of phones in the C_p set to the amount of phones in the full set for that phone label. The data can also be interpreted that the percentages in the first column show the percentage of preferred phone agreement from the HMMs with the SVMs, and the second column shows

the percentage of preferred phone agreement for the SVMs with the HMMs. In most cases most of the SVM preferred phones are a subset of the HMM preferred phones, but not *visa versa*.

Phone	$A_p \cap C_p$ %	$B_p \cap C_p$ %	$C_p \cap Full_p$ %
b	46.01	33.43	16.08
d	55.45	49.43	22.73
g	43.59	42.81	18.09
k	39.02	91.08	31.83
p	31.48	83.50	26.72
t	41.91	70.50	29.75
j	46.01	81.64	31.89

Table 2: *Phones with stops.*

Table 2 illustrates the performance of the feature extraction technique for phones with stops. Phones with stops have the lowest amount of agreement between the two techniques used, resulting in the smallest C_p sets, where the [b] phone has only 16.08% of phones of all phones labelled [b] in the C_b set. This is the lowest percentage of any phone in the C_p sets. The fact that phones with stops have the smallest C_p sets is expected. This is due to the dynamic nature of stops, where the transition leading to occlusion is quite variable. As the focus of this application of features is only on the diphone joining points of phones, perhaps the stops would get larger C_p sets if the diphone points of speech with the presence of stops allowed for temporal variance. Since the current technique only looks at a single point in each phone the acoustics of stop sounds are only measured at a single point, not allowing for the acoustic variance that occurs at stops, resulting in the low amount of articulatory acoustic consistency indicated by the size of the C_p sets. It is also possible that the forced alignment technique (or perhaps the acoustic model used) does not work very well with stops. If this was the case it would result in an increased amount of variance between the diphone boundaries. In cases such as this the synthesiser will not prioritise preferred phones when the percentage of phones in a C_p set is this low, resulting in it encouraging a minimal amount of joins to occur at a stop.

Phone	$A_p \cap C_p$ %	$B_p \cap C_p$ %	$C_p \cap Full_p$ %
tʃ	79.67	99.09	78.08
f	36.39	84.51	31.53
h	36.81	48.45	17.35
ç	59.66	95.43	56.07
s	79.16	99.39	78.35
ʃ	92.56	99.46	91.68
θ	47.22	95.80	42.75
v	68.90	60.76	39.24
z	65.56	94.94	62.53
ʒ	76.43	98.36	75.47

Table 3: *Phones with frication.*

Table 3 shows the results for fricative sounds. In almost every case the results in Table 3 are higher than the results in Table 2. It is clear that the intersection set, C_p has much more phone coverage than in the case of the stops. The C set for the [ʃ] phone had the highest percentage of any C_p set. Of the data in Table 3, the lowest set C scores were for C_f and C_h . The score for both of these is reasonable as the pronunciation of a [f] or [h] sound in real speech is very dependent on the

following phoneme. The [h] phone scores far lower than any other fricative phone, it obviously has some significantly different property than the other fricative sounds, hence its relatively low C_p score. We expect this to be due to it being glottal, which is the unique factor that distinguishes the [h] phone from the other voiceless fricatives described in the table. Additionally, glottal sounds are underrepresented in the training data and are typically of low energy which makes the task of distinguishing them more difficult.

Phone	$A_p \cap C_p$ %	$B_p \cap C_p$ %	$C_p \cap Full_p$ %
b	46.01	33.43	16.08
h	36.81	48.45	17.35
g	43.59	42.81	18.09
ɪ	29.43	55.40	19.90

Table 4: *Lowest four scoring phones in terms of C_p .*

Table 4 shows the four lowest scoring phones in respect of C_p . Two of these phones also occurred in Table 2, and one of them was in Table 3. In cases where the synthesiser comes across phones that have such a low score for the C_p set, it will use the full set of that phone rather than C_p . The [ɪ] phone got the lowest score for a vowel. This is as expected in this case, and is due to the dictionary using the [ɪ] phone too frequently when other vowel sounds would be more appropriate. This is also the reason for the [ɪ] phone to be the most common in the speech data.

Although the use of the C_p sets of phones instead of the full set of phones prunes the database considerably, in the case of a speech database similar in size to the one used in this article it still leaves a significant portion of data for synthesis. Of the 8 hour corpus used, the average size of a C_p set was 38.33% of the full set of phones, resulting in 3 hours and 4 minutes of the speech data remaining in the C_p sets. This is still a very large amount of data for the synthesiser to use in comparison with the commonly used 1 hour ARCTIC corpus as used in [14]. If a threshold is used for the synthesiser to use the full set of phones in the case of the percentage of C_p in the full set being below the defined threshold, the duration of the preferred data would increase significantly.

The patterns in the performances of the classifiers can be summarised in three types of behaviour. In the first type, the outputs from the SVMs were a clear subset of the outputs from the HMMs, regardless of the performance of the HMMs. This could be attributed to generally high recall with lower precision of the HMMs and nearly even precision and recall of the SVMs in the task of recognition of articulatory-acoustic features. This type of observation counts for 16 cases of the total of 36 phones used, where the presence of preferred phones calculated from the SVMs is higher than 90% of the set of phones preferred by the HMMs.

The second type of behaviour observes a low percentage of preferred phones being the result of agreement between the two phone classification approaches. A percentage is considered to be 'low' when it comes below 20%. In this case either the agreement of assignments to the class preferred is low between the two classifiers or the agreement is high for one of the classifiers but the percentage of phones preferred by the other of all phones is low.

The third type of behaviour shows a higher agreement of the HMMs with the assignments for the preferred phones given by the SVMs. The level of agreement is much lower in this case

for HMMs than it is in the first type of behaviour for SVMs (<65% vs. >90%).

From a general perspective, the SVM based phone selection is more conservative than the HMM based one as it results in more non-preferred phones in 30 of the 36 cases.

When using smaller speech databases is it reasonable to reduce the set of features being used to increase the sizes of the C_p sets. Even using one or two feature extractors will still help to prune out many acoustic mismatches that would have otherwise resulted in a bad join, even when using the Mahalanobis distance measure.

5. Conclusion

A technique was introduced that identifies the most acoustically similar units in the speech database. The acoustically similar units are intended for use in diphone synthesis, as the synthesiser is aware of the acoustic properties of the start and end points of units. This allows for the synthesiser to ensure a high degree of acoustic consistency during joins, as well as the Mahalanobis distance measure is still used to measure join cost. In experiments to date this technique has only been used at diphone joining points and if the synthesiser is synthesising a word that does exist in the voice data as a complete word, it will select the word regardless of the acoustic properties at the diphone points in the unit, resulting in the domain of the speech data not being affected.

This paper develops further previous work on this topic [2], and now the technique is more robust by using both SVM and HMM models to analyse the acoustic properties of the speech data. A cross validation technique does result in identifying the set of the most acoustically similar units to be used during synthesis.

An analysis of how the HMM and SVM models performed is described. Experiments were done on the 8 hour ATR Blizzard Challenge 2007 speech data. The use of the technique described identified the most acoustically similar units, approximately 3 hours of the speech data. In many cases the set of preferred phones identified by the SVM models used would be almost a complete subset of the set of preferred phones identified by the HMM models used. Results show that phones containing stops had the least amount of acoustic consistency at their diphone boundaries, and phones containing frication had the most acoustic consistency. Vowel sounds had a quite variable amount of consistency, but was always between that of the stops and the fricatives. It is expected that the variation of the vowels' consistency to be due to the dictionary used to train the grapheme to phoneme technique used, and the pronunciation of vowels being more irregular than consonants in real speech.

In future work we intend to further develop this concept. Current results are very encouraging, and the net result is a fully automatic, language independent technique to obtain an additional, reliable perspective on the content of the voice data. The most acoustically similar 3 hours of speech data in the test data was identified, and the system is aware of which phones have the most acoustic irregularities—allowing the synthesiser to weigh joins at such phones to encourage more joins at phones that are more acoustically consistent.

6. Acknowledgements

This work was funded by the Irish Research Council for Science, Engineering and Technology (IRCSET), IBM, and SFI. The experiments were done using the ATR Blizzard Challenge

2007 speech database. This material is based upon works supported by the Science Foundation Ireland for the support under Grant No. 02/IN1/I100. The opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland. The authors express their thanks to ATR for access to their speech database. The authors also express thanks to Supphanat Kanokphara for training some of the feature extractors.

7. References

- [1] J. Kominek, C. Bennett, and A. Black, "Evaluating and Correcting Phoneme Segmentation for Unit Selection Synthesis," *Proceedings of Eurospeech*, pp. 313–316, 2003.
- [2] P. Cahill, D. Aioanei, and J. Carson-Berndsen, "Articulatory Acoustic Feature Applications in Speech Synthesis," *submitted to Interspeech 2007*, 2007.
- [3] A. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," *Proc. Eurospeech*, vol. 2, pp. 601–604, 1997.
- [4] J. Carson-Berndsen, *Time Map Phonology: Finite State Models and Event Logics in Speech Recognition*. Kluwer, 1998.
- [5] K. Stevens, *Acoustic Phonetics*. Cambridge, MA, USA: MIT Press, 1999.
- [6] B. Schölkopf and A. J. Smola, *Learning with Kernels*. Cambridge, MA, USA: MIT Press, 2002.
- [7] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer Verlag, 1995.
- [8] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.
- [9] J. Macek and J. Carson-Berndsen, "Articulatory manner features recognition with linear and polynomial kernels," in *Fifth Slovenian and First International Language Technologies Conference*, (Ljubljana, Slovenia), Oct. 2006.
- [10] A. Gray Jr and J. Markel, "Distance measures for speech processing," *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing]*, *IEEE Transactions on*, vol. 24, no. 5, pp. 380–391, 1976.
- [11] R. Donovan, "A new distance measure for costing spectral discontinuities in concatenative speech synthesizers," *The 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, 2001.
- [12] P. Taylor, "Concept-to-speech synthesis by phonological structure matching," *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, vol. 358, no. 1769, pp. 1403–1417, 2000.
- [13] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81–106, 1986.
- [14] A. Black and K. Tokuda, "The Blizzard Challenge–2005: Evaluating corpus-based speech synthesis on common datasets," *Proc. of Interspeech 2005*, pp. 77–80, 2005.