The effect of speech rate variation on acoustic phone stability in Afrikaans speech recognition

Jacob A.C. Badenhorst and Marelie Davel

North-West University / Human Language Technologies Research Group
Meraka Institute, Pretoria, South Africa
jbadenhorst@csir.co.za, mdavel@csir.co.za

Abstract

We analyse the effect of speech rate variation on Afrikaans phone stability from an acoustic perspective. Specifically we introduce two techniques for the acoustic analysis of speech rate variation, apply these techniques to an Afrikaans speech recognition corpus containing extensive speech variation, and demonstrate how these techniques can be used to better understand the performance of a speech recognition system trained on such data.

1. Introduction

Typically there are various sources of variation present in the acoustic training data that are used when building an automatic speech recognition (ASR) system. Sources of variation include dialect differences, speaker differences, channel effects such as bandwidth and background noise, speaking style and vocabulary used. One such source of variation that has received some attention in literature is speech rate variation [1]. In this paper we investigate the effect of speech rate variation on Afrikaans phone realisation. Specifically we aim to answer the following questions:

1. How stable are the different Afrikaans phones with regard to speech rate variation?
2. Which techniques can assist us in analysing such phone variability?
3. Can these techniques assist us in understanding and explaining ASR performance?

While literature exists that addresses the effect of speech rate variability on phone duration [2], we are specifically interested in the effect of speech rate variability on phone acoustics.

The remainder of this paper is structured as follows: In section 2 we provide some general background with regard to the analysis techniques used. Section 3 contains a description of our experimental setup, while section 4 contains a discussion of the specific analysis techniques used and the results obtained. Section 5 summarises the main conclusions of the study and describes possible future work.

2. Background

Two of the tools used extensively in our analysis are (1) the Bhattacharyya distance [3], a distance measure defined between two Gaussian distributions, and (2) speaker space correlation matrices, a measure of the amount of cross-speaker correlation across phones.

2.1. The Bhattacharyya distance

Many texts on statistical pattern recognition cover the Bhattacharyya distance. This is a useful measure to compare how far two Gaussian distributions are apart. If we define the following notations [4]:

$$\begin{aligned} \omega_i &: \text{class } i = 1, 2 \\ M_i &: \text{mean vector of class } \omega_i \\ \Sigma_i &: \text{covariance matrix of class } \omega_i \end{aligned}$$

then the Bhattacharyya distance, D_{bhat}, is defined as:

$$\begin{aligned} D_{bhat} = \frac{1}{8}(M_2 - M_1)^T \Big[\frac{\Sigma_1 + \Sigma_2}{2}\Big]^{-1}(M_2 - M_1) \\ + \frac{1}{2} ln \frac{\left|\frac{\Sigma_1+\Sigma_2}{2}\right|}{\sqrt{|\Sigma_1||\Sigma_2|}} \end{aligned} \quad (1)$$

The equation consists of two parts that give separability due to the difference between class means and the difference between class covariance matrices respectively [4].

2.2. Speaker space correlation matrices

Speaker space correlation matrices provide a measure of the amount of cross-speaker correlation across phones. The eigenvector decomposition of the speaker space correlation matrix has been shown to be useful in normalising speech (performing speaker normalisation) prior to speech recognition system training [5].

Speaker space correlation matrices as defined in [5] are constructed by extracting feature vectors (such as Mel-Frequency Cepstral Coefficient (MFCC) observations or formant frequencies) according to phone identities, and generating a single mean value per feature per speaker per phone. For every speaker an *m x d*-dimensional vector is constructed for the *m* observed phones and *d*-dimensional feature vector.

The vectors of phones being investigated are then concatenated in the same sequence for each speaker, which results in a matrix of speaker vectors. If the correlation values are given by:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \quad (2)$$

$$\text{where } Cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)] \quad (3)$$

and σ_X indicates the standard deviation of X and μ_X the mean of X, then by calculating the correlation matrix of the matrix of speaker vectors we end up with a speaker space correlation matrix.

The means of the feature vector of a specific context (phone) varies with a certain pattern between different speakers.

Speaker space correlation matrices thus describe the similarity of these variations across contexts.

3. Experimental setup

We use two acoustic analysis techniques and the results of a set of speech recognition experiments to perform our analysis. In this section we briefly describe the data set and speech features used, as well as the basic construction of the speech recognition system.

3.1. Data set

The data set utilised during this study was custom-developed by North-West University. This corpus contains transcribed read speech with samples of both slow and fast speech for each speaker. The set of Afrikaans data that we analyse contains 127 speakers and consists of two and a half hours of 16 bit data with a sample rate of 22 kHz. It is not the full corpus which contains 130 speakers and consists of three and a half hours of data. Due to transcription errors, pronunciation errors and two low quality speakers one hour of data were not used. Durations for the slow and fast data sets are approximately one and a half hours and one hour respectively. This is a limited resource as far as speech recognition is concerned.

3.2. Speech features

Mel-Frequency Cepstral Coefficients (MFCCs) are used as speech features during the analysis. These coefficients are calculated from the spectrum of the speech signal. From the Discrete Fourier Transform (DFT) of the signal, a filter-bank containing triangular filters is used to compute the average of the spectrum around the center frequency of each filter. These filters have increasing bandwidth and are spaced along the mel-scale. MFCCs are then defined as the discrete cosine transform of the log filter-bank amplitudes [6].

We use only the first 13 MFCC coefficients and include delta and acceleration coefficients in some of the analyses. The delta and acceleration coefficients are calculated by taking the first and second derivatives of MFCC 0 - 12 (0 being a representation of total spectral energy) [7], resulting in a total of 39 coefficients used. We use a frame size of 25ms and an overlap of 15ms, resulting in one observation every 10ms.

3.3. Speech recognition system

We build a Hidden Markov Model (HMM)-based speech recognition system, using the toolkit HTK [7], Gaussian mixtures to model observation probabilities, a 3-state left-to-right HMM per phone. For different purposes we build a monophone system, a single-mixture triphone based system and a 17-mixture triphone-based system.

Different speech recognition systems are constructed during the analysis: a system built using only a portion of the slow speech (S), only a portion of the fast speech (F) or a portion of all available data (A). The various train and test data sets are randomly selected to consist of 90% and 10% of the above data sets. In the case of slow data, the training set used for ASR experiments were further limited to be of the same duration as the fast data training set.

4. Analysis and results

In this section we describe the different analysis techniques used: (1) an initial duration analysis of the data set in order to ensure that significant speech rate differences do indeed exist, (2) acoustic analysis of the data according to distance measures, (3) acoustic analysis of the data according to speaker space correlation, and (4) an analysis of speech recognition performance.

4.1. Duration analysis

In order to do a first analysis of the durations of each phone, we perform a forced alignment using the best available speech recognition system (A) and the known transcriptions. For each phone we calculate the average duration across all speakers for both the fast and slow speech respectively.

Table 1 lists the difference in the average phone durations between the slow and fast data sets for the phones considered. The actual phone duration times are derived from the average amount of MFCC observations allocated to the particular phone for that dataset. Given the fact that one MFCC observation has a length of 10 ms, it can be seen that the average differences between phones of the fast and slow data range between 2 and 20 ms.

Phone	Duration difference (10ms)
a:	2.035
e:	1.971
s	1.529
x	1.322
f	1.305
p	1.259
ey	1.240
o:	1.145
k	1.020
ao	0.991
t	0.926
ah	0.917
eh	0.892
m	0.888
v	0.805
n	0.768
b	0.723
iy	0.670
h	0.629
l	0.586
r	0.578
ax	0.488
d	0.255

Table 1: *Average phone duration difference between fast and slow data*

As can be expected, the difference in phone length is also visible through the transition probabilities of the HMMs. For example, for a monophone such as /a:/ that has a lengthened duration for slow speed rates, the probability of remaining in a given state is 0.59, 0.80 and 0.61 for states 1, 2 and 3 respectively when trained on fast data, and 0.65, 0.84 and 0.68 for states 1, 2 and 3 respectively, when trained on slow data. In this way the HMM provides an automatic compensation mechanism for the difference in duration between the two models.

4.2. Distance measures between acoustic models

A logical place to look for variations that occur in data is to look at the acoustic models themselves. The single-mixture HMMs we use all consist of three emitting states, each containing one Gaussian distribution to model the acoustics of that state. Using the Bhattacharyya distance as distance measure to compare the distributions between similar states of models, enables us to determine how similar two models trained on different data sets are. As the data set is fairly limited, we constrain this analysis to single-mixture monophone models.

4.2.1. Defining confidence intervals

When comparing two acoustic models of different data sets, it is important to consider the amount of training data that resulted in the estimation of the model. The Bhattacharyya distance between two acoustic models trained on different data sets results in part from the fact that the distributions themselves are different for poorly estimated and well estimated models. This effect will decrease as more data is used and will disappear when sufficient data results in accurate density estimation.

In order to correctly interpret comparative results we use the following process to define confidence intervals for our results: Firstly, the total amount of phone occurrences for every phone found in the data is measured. From these values we select a phone that occurs very frequently (over 10 000 occurrences in the training data) in order to ensure an accurate estimation of the acoustic model for that particular phone. We also ensure that the Bhattacharyya distance between two sufficiently estimated models of different datasets (such as F and S) for this phone is small compared to the other phones found in the data. This is a good indication that other variations in the data does not have a big effect on the chosen phone. The Afrikaans phone we choose is the short vowel /ax/ as found in the Afrikaans pronunciation of 'kind' (/k ax n t/, 'child' in English).

We then construct training subsets from the full training set containing an increasing number of occurrences of the phone in question. For each training subset we estimate a separate model. Models are thus estimated for phone counts of /ax/ from 10 up to 10 000 occurrences. For each of these models the Bhattacharyya distances with a well estimated model are determined.

Confidence intervals can be described as the accuracy of the estimation of a model at a particular phone occurrence as a function of the total deviation (worst case scenario) observed. By taking the distance value of the most poorly estimated model of our selected phone /ax/ as the maximum variance, each of the other distance values can be expressed as a percentage of the maximum deviation, resulting in a the confidence intervals listed in Table 2.

4.2.2. Grouping of phones based on Bhattacharyya distance

Not all Afrikaans phones were sufficiently covered by the limited amount of training data. It was thus unavoidable that some acoustic models were not estimated well. By comparing the distance between models found in Table 3 with the confidence interval values of Table 2, it is easy to see that the phones cannot be classified when model estimations is performed on less than 850 occurrences. As a result all phones occurring less than 850 times for the test being performed were omitted.

We are left with 23 phones that can further be divided in two sets. One for phone occurrences below and above 2000 respectively. By doing this we end up with two groups of phones that can be classified according to their acoustic distances within a confidence interval of 2% and 0.59% of the expected maximum deviation respectively.

Mean Bhattacharyya	Percentage of max	Occurrences
1.793	100.00	10
0.660	36.80	50
0.443	24.73	100
0.169	9.44	200
0.094	5.25	500
0.027	1.51	700
0.037	2.05	850
0.025	1.40	1000
0.011	0.59	2000
0.002	0.13	5000
0.000	0.00	10000

Table 2: *Bhattacharyya distance between /ax/ model for all slow data and partially estimated slow model with selected counts*

This grouping being applied to the phones found in Table 3, the Bhattacharyya distance values can now be used to compare the acoustics of two speech recognition models. The values represent the average (across the three emitting states) of the Bhattacharyya distances between the models for the same phone trained on a selected fast and slow dataset of same duration 3.3.

Phone	AVG Bhat distance	Occurrence
iy	0.068	2025
l	0.061	2185
n	0.060	3895
k	0.059	2047
d	0.058	2592
s	0.054	3562
t	0.053	3989
ax	0.044	6162
ah	0.044	2326
r	0.031	3660
eh	0.141	1397
ao	0.137	1565
e:	0.102	1024
ey	0.085	1040
v	0.071	1183
a:	0.064	1567
m	0.057	1569
f	0.054	1625
h	0.045	1313
x	0.037	1744

Table 3: *Average Bhattacharyya distance between models for fast and slow data*

It is interesting to note that the size of the acoustic difference observed does not correlate strongly with the observed changes in duration (as listed in Table 1), but rather seem to correlate fairly well with the place of articulation.

Overall sounds that require the most effort to pronounce correctly are affected most. By this effort we refer to sounds that are pronounced with a raised tongue, rounding or widening and phones requiring transitions such as diphthongs [8]. Table 4 summarizes these difficulties for all the vowels in Table 3.

Phone	Rounding	Tongue raised	Change
eh	Wide	Medium	No
ao	Rounded	Medium	No
e:	Wide	High - Medium	Yes
ey	Neutral	Medium - High	Yes
iy	Wide	High	No
a:	Neutral	Low	No
o:	Rounded	Low	No
ax	Neutral	Medium	No
ah	Neutral	Low	No

Table 4: *Pronunciation difficulties of vowels ordered according to difficulty (starting with the most difficult)*

4.3. Speaker space correlation

Next, we analyse the data according to speaker space correlation. We only use the first 13 coefficients to construct speaker space correlation matrices. For each sound file the MFCC observations (13 dimensional vectors) are extracted and grouped according to phone pronunciations using our optimal speech recognition system (A) and forced alignment.

We now create a new speaker space correlation matrix structure consisting of four phone averages per speaker instead of the normal one. This is done by taking the same fast and slow data sets used to train the acoustic models and obtaining separate means for the fast and slow data respectively. In addition, each data set is split into two random sets of equal size, resulting in two means for each slow feature vector as well as two means for each fast feature vector (per phone). The mean vectors of these four representatives of the same phone (two fast and two slow) are then concatenated for every speaker. We calculate the correlation matrix of the matrix of speaker vectors (according to section 2.2) as shown for the phone /ax/ in Figure 1. The phone labels /ax/,/ax2/,/ax_F/,/ax2_F/ indicate the order that the mean vectors of each phone have been added for the slow, second slow, fast and second fast phones respectively.

From the correlation matrix it is possible to extract sub-matrices for every combination of the two fast and slow phones. Each of the diagonal elements of these sub-matrices then represent the correlation between the means of the same MFCC coefficient across speaker space for the particular combination. These diagonal elements are referred to as 'strip correlations' [5].

4.3.1. Evaluating differences in correlation

Comparison of the strip correlations yield interesting results. When we group the values of the strips for phone combinations slow-slow (SS), fast-fast (FF) and slow-fast (SF), Table 5 clearly shows stronger correlations for the SS and FF combinations than for the SF combination. This is the case for almost all of the selected set of 23 phones. This result indicates that there is stronger cross-speaker correlation across the same speech rate, than between different speech rates.

Table 6 shows the same experiment, but now two phones at a time are compared across speaker space. The comparative values for three short vowels and two very similar consonants is shown. As is expected, it can be seen that the values are significantly lower than the values where the same phone is compared with itself. Note that the difference caused by speech rate variation is still apparent in the SF column.

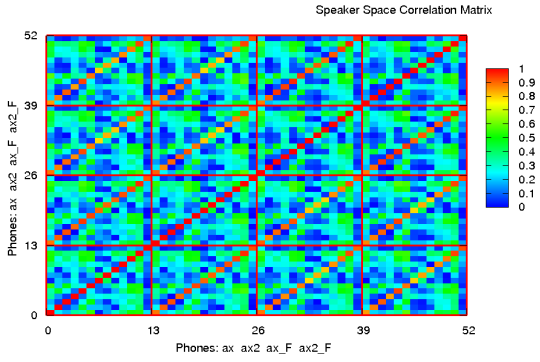

Figure 1: *Speaker space correlation matrix comparing four phones*

4.4. Implications for speech recognition

We are interested in determining whether the analysis described in the preceding section sheds light on the confusion results observed. In order to analyse speech recognition performance, we analyse the phone confusion matrices. Such a matrix consists of a row for each phone occurring in the true transcriptions and a column for each model used during recognition. Each matrix entry contains a count that corresponds to the number of times that a particular phone was recognized by a particular model. When a transcribed phone is recognized as being closer to a different model than intended, a phone confusion (recognition error) occurs. From this it is possible to classify the phones according to how well they are recognized. When the testing data has a different speech rate than the training data, the recognition accuracies of phones may be affected. While confusion matrices are reported on according to monophones, actual recognition (when generating the confusion matrices) utilises the most accurate triphone-based system.

4.4.1. Defining confidence intervals

As before, it is important to determine when results are significant given the issue of data scarcity. We define a confidence interval according to the expected variation of the mean of n measurements (σ_n) where:

$$\sigma_n(phone) = \frac{\sigma}{\sqrt{n}} \simeq \sqrt{\frac{p(1-p)}{n}} \text{where, } p = \frac{correct}{n} \quad (4)$$

and σ is the standard deviation of the measurement.

4.4.2. Evaluating ASR performance

We first evaluate overall performance on the three systems built when recognising either slow or fast data. Phone recognition results are listed in table 7 when recognising a 5000-word vocabulary using a flat word model (no statistical language model used to guide recognition).

It is clear that overall results improve with additional training data, and that slow and fast testing data do not provide better accuracy on systems trained only on slow or fast training data.

We are specifically interested in the difference in confusion when slow data is recognised on the best performing system A_S

Phone	SS	FF	SF
iy	0.820	0.815	0.776
l	0.697	0.696	0.685
n	0.904	0.903	0.863
k	0.616	0.615	0.587
d	0.732	0.744	0.690
s	0.889	0.888	0.816
t	0.777	0.758	0.704
ax	0.918	0.927	0.873
ah	0.863	0.837	0.820
r	0.811	0.817	0.757
eh	0.796	0.803	0.765
ao	0.749	0.761	0.737
e:	0.822	0.813	0.794
ey	0.823	0.801	0.785
p	0.440	0.396	0.440
v	0.649	0.602	0.596
a:	0.850	0.803	0.798
b	0.560	0.614	0.581
o:	0.735	0.734	0.729
m	0.797	0.841	0.803
f	0.804	0.829	0.732
h	0.548	0.621	0.595
x	0.741	0.706	0.670

Table 5: *Strip correlations between SS FF and SF of the same phone*

Phone	SS	FF	SF
ah eh	0.587	0.604	0.558
ao eh	0.605	0.602	0.569
ax ah	0.780	0.782	0.734
ax ao	0.670	0.673	0.625
ax eh	0.698	0.721	0.670
d t	0.627	0.636	0.558
t s	0.647	0.641	0.580

Table 6: *Strip correlations between SS FF and SF of different phones*

versus when slow data is compared on the most mismatched system F_S and similarly the difference in confusion when fast data is recognised on the best performing system A_F versus when fast data is compared on the most mismatched system S_F. Here X_Y indicate training on system X and testing with data from Y. We then calculate these differences as: $D_1 = A_F - S_F$ and $D_2 = A_S - F_S$ where A_F, A_S, S_F and F_S are the confusion matrices of that particular testcase, expressed in percentages.

Confidence intervals were calculated for all the phones of every confusion matrix. Only phones that have smaller confidence intervals than accuracy differences can be considered.

When evaluating the results for the phones in Tables 8 and 9 it is clear that less phone comparison results are significant (distances greater than confidence intervals) for the D_2 case. This is due to the fact that less phones can be realized within the same time duration for the slow data testset. As can be inferred from Equation 4, when dividing by a smaller n, the confidence intervals get bigger.

As it turns out the distances between the same phones of the A_F and S_F test case are greater than the distances between the same phones for the A_S and F_S test case. Looking at Table 7, we see that this results correlate well, as the phone recognition results for the S_F case are indeed worse than the F_S case.

Training	Testset S	Testset F
S	69.63%	67.58%
F	71.86%	65.70%
A	76.76%	74.65%

Table 7: *Phone recognition results for combinations of speech recognition datasets*

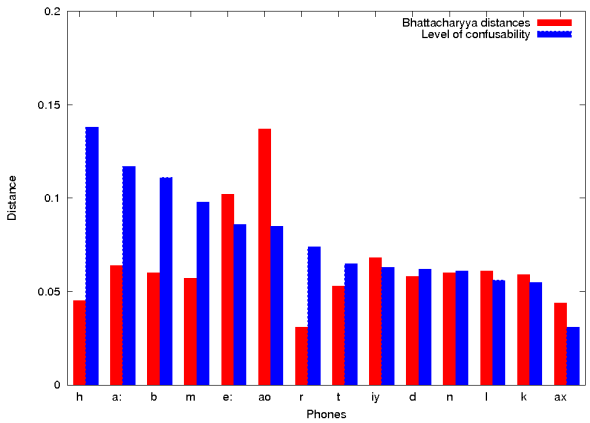

Figure 2: *Bar chart of the Bhattacharyya distances and the confusability of the phones selected*

Figure 2 shows the Bhattacharyya distances found in Table 3 compared to the level of confusability of the phones for the phone recognition results of the D_1 case as in Table 8. Normalization of the Bhattacharyya distances can be done to compensate for the estimation differences of the models due to the amount of phone occurances found in the datasets. A common operating point for every phone, of say 2000 phone occurances, can then be estimated for both the (S) and the (F) datasets. When this is done it is clear that the goodness of model estimation still plays the larger role in the phone recognition results for the investigated datasets.

5. Conclusion

This paper produces initial results with regard to speech rate variation in Afrikaans. While more conclusive results with regard to speech rate variability would require analysis on a larger corpus, we note the following:

1. The correlation between phone duration change and acoustic change (due to speech rate variation) is weak.
2. We found that phone acoustics are affected differently for each particular phone in question.
3. There is an interesting relationship between the amount of acoustic change (between slow and fast realization of the same phone) and the difficulty of pronouncing the phone using measures such as place of articulation and rounding.
4. The Bhattacharyya distance between single mixture monophones of slow and fast speech provides some indication of the expected phone confusebility as measured by a speech recognition system.

Phone	% Correct	Difference	Conf 1	Conf 2
ay	57.14	0.429	0.187	0.132
eu	66.67	0.333	0.111	0.111
u	70.49	0.164	0.058	0.064
ng	87.65	0.123	0.037	0.048
h	58.47	0.093	0.032	0.033
m	84.25	0.075	0.021	0.025
ah	86.65	0.071	0.017	0.021
t	86.76	0.066	0.013	0.015
k	91.80	0.050	0.014	0.017
ao	84.27	0.049	0.022	0.024
a:	87.50	0.045	0.020	0.023
l	88.95	0.043	0.016	0.019
ax	85.19	0.042	0.011	0.012
iy	78.64	0.041	0.020	0.021
r	92.21	0.036	0.011	0.013
n	86.95	0.035	0.013	0.014

Table 9: *Difference D_2 of A_S and F_S and % of correct recognitions measured on the best model*

Phone	% Correct	Difference	Conf 1	Conf 2
oei	83.33	0.667	0.152	0.152
g	83.33	0.583	0.152	0.217
u:	84.62	0.385	0.100	0.138
y	66.67	0.333	0.136	0.136
ow	58.54	0.293	0.077	0.071
ui	83.34	0.267	0.068	0.090
h	55.97	0.138	0.039	0.039
a:	82.46	0.117	0.029	0.035
b	86.87	0.111	0.034	0.043
m	81.52	0.098	0.029	0.033
e:	82.76	0.086	0.035	0.041
ao	82.42	0.085	0.030	0.034
r	84.36	0.074	0.018	0.021
t	86.08	0.065	0.017	0.019
iy	73.52	0.063	0.028	0.030
d	70.92	0.062	0.026	0.027
n	86.21	0.061	0.016	0.019
l	83.76	0.056	0.024	0.027
k	86.18	0.055	0.023	0.027
ax	83.12	0.031	0.014	0.015

Table 8: *Difference D_1 of A_F and S_F and % of correct recognitions measured on the best model*

5. Speaker space correlation matrix analysis provides a perspective on the effect of speech rate change relative to the acoustic differences that normally exist between closely related phones.

Further work includes refining and extending the techniques introduced here to analyse other forms of variation in speech, including cross-language acoustic variation.

6. References

[1] B. Wrede, *Modelling the Effects of Speech Rate Variation for Automatic Speech Recognition*, Ph.D. thesis, University of Bielefeld, 2002.

[2] J.G. Carbonell and J. Siekmann, *Speaker Classification II*, Springer Berlin / Heidelberg, 2007.

[3] A.R. Webb, *Statistical Pattern Recognition Second Edition*, Butterworth Heinemann, 2002.

[4] B. Mak and E. Barnard, "Phone clustering using the Bhattacharyya distance," in *Proceedings of ICSLP*, Oct 1996, vol. 4, pp. 2005–2008.

[5] Zhihong Hu, *Understanding and adapting to speaker variability using correlation-based principal component analysis*, Ph.D. thesis, Oregon Graduate Institute of Science and Technology, 1999.

[6] O. Grebenskaya, "Speaker Clustering in Speech Recognition," M.S. thesis, University of Joensuu, 2005.

[7] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Veltchev, and P. Woodland, *The HTK Book*, Cambridge University Engineering Department, http://htk.eng.cam.ac.uk/, 2005.

[8] Sami Lemmetty, "Review of Speech Synthesis Technology," M.S. thesis, Helsinki University of Technology, 1999.