

Using an underspecified ASR system as an indicator for phonetic similarity

Mark Kane, Julie Mauclair, Julie Carson-Berndsen

School of Computer Science and Informatics

University College Dublin

Belfield, Dublin 4, Ireland

mark.kane@ucdconnect.ie, {julie.mauclair, julie.berndsen}@ucd.ie

Abstract

This paper presents a novel approach to the identification of phonetic similarity using properties observed during the speech recognition process. An experiment is presented whereby specific phones are removed during the training phase of a statistical speech recognition system so that the behaviour of the system can be analysed to see which alternative phone is selected. The domain of the analysis is restricted to specific contexts and the alternatively recognised (or *substituted*) phones are analysed with respect to a number of factors namely, the common phonetic properties, the phonetic neighbourhood and the frequency of occurrence in the complete corpus. The results indicate that a measure of phonetic similarity based on alternatively recognised observed properties can be predicted based on a combination of these factors and as such can serve as an important additional source of information for the purposes of pronunciation variation in speech recognition.

Keywords: speech recognition, markedness, phonetic similarity

1. Introduction

A key challenge in speech recognition is to construct acoustic models which correctly estimate a sub-word unit or phonetic class label within a specific time interval. The smallest posited linguistically distinctive unit that is typically modelled is the phoneme. However, phonemes that belong to the same acoustic-articulatory group (i.e. have similar acoustic or articulatory properties) are easily confused and thus statistical context-dependent phone models are often used as a basis for deciding which phoneme seems to be the best according to an acoustic probability and the probability of its occurrence with respect to a language model. Confusability of phonemes and the relationship to underlying phonetic properties of speech sounds remains an important area of research in order to address variability in the domain of speech recognition.

The research presented in this paper is motivated by research into patterns which emerge from mis-recognition of phonemes during the speech recognition process and the improvements which can be achieved when phonetic similarity is employed as an additional source of information in speech recognition. Experiments presented in (Halberstadt and Glass, 1997), demonstrated that almost 80% of all mis-recognised phonemes belong to the same phonetic group as the correct phoneme. In (Scanlon et al., 2007) this information is used in order to build *Broad Phonetic Groups* (BPG) that are defined according to a confusability matrix and a new phoneme classifier is proposed consisting of modular arrangements of experts, with one expert assigned to each BPG and focused on discriminating between phonemes within that BPG. The result in PER achieved by that system on the TIMIT corpus (Garofolo et al., 1993) is 26.4%. More recently these phonetic and

phonological features are used to exploit similarities in phoneme recognition (Mauclair et al., 2009) with encouraging results. Neural Networks have also been used to classify between phonetic groups (Thuan and Kubin, 2005) and (Ghiselli-Crippa and El-Jaroudi, 1991). In (Thuan and Kubin, 2005) Neural Networks are used to improve the Discrete Wavelet Transform-based phonetic classification algorithm.

This paper focuses on the identification of phonetic similarity using properties observed during the speech recognition process. This approach also uses the notion of mis-recognitions but in a very different way. Rather than construct a confusability matrix for the recognised output, the output of two statistical speech recognition systems are compared, one where all phonemes to be recognised were in the training data and one where individual phonemes are systematically removed from the training data (note that all phones occur in the testing data). This allows the identification of the *substituted* choice where a particular phoneme is not available. Where the number of substitutions will increase dramatically in the testing phase for the removed phone due to there being no ASR (hidden Markov) model built to represent that phone from the training data. The domain of the analysis is restricted to specific phone contexts and thus the term *phone* will be used in the remainder of the paper. The substituted phones thus identified are then analysed with respect to their phonetic properties as given by a phonetic feature classification based on the IPA chart (The-International-Phonetic-Alphabet, 2005). These properties provide a principled way to investigate phonetic similarity, underpinned by insights from experimental phonetics and phonological theory.

The remainder of paper is structured as follows. Section 2. presents the speech recognition system used in the experiment and section 3. details the experimentation carried out to identify the substituted phones using the TIMIT corpus. The results and analysis of the experiment are discussed in section 4. and 5., some conclusions are drawn

This material is based upon works supported by the Science Foundation Ireland under Grant No. 07/CE/I1142. 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.

and directions for future work are highlighted in section 6.

2. Speech Recognition System and Corpus

The HMM based speech recognition system used in this experiment is implemented with HTK (Young et al., 2009). The TIMIT speech corpus (Garofolo et al., 1993) is used for training and testing of the HMM models. For completeness, the more technical details of the speech recognition system and the corpus are presented in the next section in the traditional way.

2.1. Parameterisation of speech, description of HMM and corpus structure

The chosen form of parameterisation of a phone within an utterance is mel frequency cepstral coefficients (MFCCs), with their associated log energy and first and second order regression coefficients. First, each speech waveform is passed through a pre-emphasis filter to compensate for the -6 dB/Octave roll-off associated with speech. The waveform is then framed at a rate of 10 ms with a window size of 25 ms. To avoid truncation errors arising from going between the temporal and frequency domain, a Hamming window function is used. The windowed frame is then changed from the temporal domain to the frequency domain by a fast-Fourier-transform. In the frequency domain the signal, up to the Nyquist frequency, is split into 26 equal overlapping triangular windows and then mapped to the mel-scale and the log filter bank amplitudes are calculated. A cepstral mean normalisation is also carried out. Twelve MFCCs are then taken from the Discrete Cosine Transform of these amplitudes. The log energy is also added as the 13th parameter. Delta (first order regression) and Acceleration (second order regression) coefficients are also used. Therefore every frame is represented by 39 coefficients.

These MFCCs representing the phones are then used in the calculation of the HMM models. The HMMs are context-dependent triphone models that were initially calculated by cloning and re-estimating context-independent monophone models. The triphones states were tied using phonetic feature decision trees for clustering. Each model is comprised of 5 states where only the centre 3 states are emitting. The decoding process is implemented with a bigram language model and without any pruning. The models are finally expanded to 10 mixtures.

The TIMIT corpus consists of read speech spoken by 630 speakers of American English. This corpus was designed to contain three types of speech, phonetically-compact (SX), phonetically diverse (SI) and dialect (SA). The SA data is not used in this paper as its training and test categories overlap to a certain degree. The remainder of the data is split into two sets; training and core test. The training set consists of 1386 SI and 2310 SX utterances totalling 3696 utterances. The core test consists of 72 SI and 120 SX utterances. This gives a total of 192 utterances. There is no overlap between any of the sets used in this paper. Finally, the TIMIT phone set was reduced from 61 to 39 phones using the CMU/MIT mapping as described in (Lee and Hon, 1989) and (Robinson and Fallside, 1990).

The experiment detailed in the next section distinguishes between a baseline, which is the standard speech recognition system as described above, and an underspecified system for specific phones. The underspecified system differs from the baseline system only in that one specific phone is removed from the training set and thus during recognition the system will not have encountered any previous instances of this phone.

3. Experiment

In this paper, two experiments based on the underspecified system for specific phones is presented. These experiments involve the identification of the substituted alternatives when one phone is removed from the training set. Of the 39 phones, the phones belonging to the fricative class were prioritised as a starting point for experimentation ([f], [v], [th], [dh], [s], [z], [sh]). The phonetic make-up of each of the fricative phones (and other phones within this paper) were classified according to the IPA chart. Results have been calculated for all of the fricative class, but only the analysis of the phones [f] and [th] are presented in this paper.

As mentioned in the previous section, two types of speech recognition systems were evaluated: baseline and underspecified. The baseline system is trained on the full TIMIT training set, whereas an underspecified system for each phone was trained on the full TIMIT training set where all instances of that particular fricative phone were removed. Both systems were evaluated on the TIMIT core test set.

From the baseline system all instances of the correctly recognised fricative phones and all substitutions for the fricative phones were identified. A certain level of control was required in these experiments and thus only substitutions in which the left and right contexts of the substituted phone were correctly recognised were considered. This ensured that preceding errors had a limited affect on results. It is clear that for an underspecified system for a particular fricative phone, there can be no correctly recognised instances of that phone. The substitution phones act as possible alternatives and serve as a comparator.

In this underspecified system where a phone is removed, it is anticipated that the substituted alternatives should be determined with respect to common phonetic properties, phonetic neighbourhood and perhaps with respect to frequency of occurrence in the full corpus, as this is an indicator of the relative frequency of the sound in the language. Each of these factors are underpinned by experimental phonetics and phonological theory (feature theory, markedness and underspecification, in particular (Chomsky and Halle, 1968)) and, in combination, should serve as indicators for phonetic similarity. The results of the experiments and analysis of each of the substituted possibilities for [f] and [th] fricative phones with respect to these factors are presented in sections 4. and 5.

4. Results

Based on the experiments presented above, the results are presented as follows. Firstly, the results for the recognition of the fricative phones [f] and [th] with the baseline

system are presented. Secondly the results for the underspecified system for the fricative phones [f] and [th] are presented. Finally, the results for the underspecified system for these fricative phones are analysed in the context of phonetic similarity.

4.1. General information

The baseline system serves to highlight which substitutions are found for each of the fricative phones when they are not correctly recognised, given a correct left and right context (LR context). Table 1 shows the general ASR statistics for these fricative phones while table 2 shows the substituted alternatives of these phones. The LR context quantity (a subset of the substitution quantity) is the number of the fricative phones that were recognised incorrectly and were replaced by another phone, where both the phone to the left and right of the substituted phone were correctly recognised. In this table the following are also highlighted; the number of times a phone occurred, how many times it was recognised correctly, how many times it was not recognised correctly and was substituted, how many times it was not recognised but deleted altogether and how many times it was inserted. The sum of the recognised, substituted and deleted is equal to the occurrence of that phone within the test set. Tables 1 and 2 also contain the corresponding information for the underspecified system.

Qty	Baseline		Removed	
	f	th	f	th
occurrence	131	38	131	38
recognised	114 (87%)	15 (39.5%)	0 (0%)	0 (0%)
substituted	14 (10.7%)	20 (52.6%)	104 (79.4%)	29 (76.3%)
deleted	3 (2.3%)	3 (7.9%)	27 (20.6%)	9 (23.7%)
inserted	5	1	0	0
LR context	5	7	26	9

Table 1: General information of the fricative phone class in the baseline and underspecified (removed) system when evaluated against the test set.

Baseline			Removed				
f	th		f	th			
p	2	dh	3	th	10	t	4
s	1	t	2	t	5	dh	2
v	1	s	1	s	3	d	1
th	1	f	1	v	3	s	1
				p	2	f	1
				ih	2		
				sil	1		

Table 2: Substituted alternatives and quantities from the baseline and underspecified (removed) system

As to be expected, it can be seen from table 1 that the number of substitutions and LR contexts of a phone is greater in the underspecified system. This gives a broader range of similar phones as they are generated from larger portion of data. This extra data allows a more complete pattern of substitutions to be observed.

5. Analysis of underspecified system

5.1. Fricative phones [f] and [th]

From the baseline system, [f] was substituted 14 times and [th] was substituted 20 times. The number of [f] substitutions that have a correct LR context is 5 and the number of [th] substitutions that have a correct LR context is 7.

Due to the fact that in the underspecified system [f] and [th] are not recognised, the amount of substitutions increased. In this system [f] was substituted 104 times and [th] was substituted 29 times. The number of [f] substitutions that have a correct LR context is 26 and the number of [th] substitutions that have a correct LR context is 9. This extra data helps disambiguate the substitution data as there is more of it to help ascertain a pattern. In this section, the following three factors are considered: 1. *common phonetic properties*, 2. *phonetic neighbourhood* and 3. *frequency of occurrence*. Finally, the combinational effects of these factors are explained.

5.2. Common phonetic properties of [f] and [th]

Phones can also be characterised in terms of the canonical properties or features they possess. Indeed this is typically one of the metrics employed to measure phonetic similarity. While the properties clearly relate to the notion of phonetic neighbourhood, depending on the feature set used, they may offer additional granularity which allows features to be grouped as natural classes. Tables 3 and 4 provides the subset of relevant features for the consonant phones in question, where + means present and - means not present. Note that other subset groupings of these sounds may also be relevant for further experimentation.

Phone	fric plosive		labio-dental	dental	alveolar	voiced
f	+	-	+	-	-	-
th	+	-	-	+	-	-
t	-	+	-	-	+	-
s	+	-	-	-	+	-
v	+	-	+	-	-	+
p	-	+	+	-	-	-

Table 3: Subset of features describing phonetic properties of substituted phones for [f]

Phone	fric plosive		labio-dental	dental	alveolar	voiced
th	+	-	-	+	-	-
t	-	+	-	-	+	-
dh	+	-	-	+	-	+
d	-	+	-	-	+	+
s	+	-	-	-	+	-
f	+	-	+	-	-	-

Table 4: Subset of features describing phonetic properties of substituted phones for [th]

5.3. Phonetic neighbourhood

The notion of phonetic neighbourhood can be visualised as in figure 1 where the four planes of the cube indi-

cated in the figure represent the fricative manner of articulation, the plosive manner of articulation, the voiced and the unvoiced articulations, where the horizontal dimension indicates the place of articulation within these planes. For the initial experiments, relative positions of phones on this cube were used as a basis for the analysis. A more detailed explication of the role of phonetic neighbourhood may be possible, based on a numerical distance measure and a differentiation between the relative weightings of the dimensions within the cube. For example, it may be that remaining on the *place* of articulation axis represents closer proximity than moving to the *voicing* axis for some occasions; it would certainly appear reasonable that a change in the manner of articulation should involve a greater phonetic distance in comparison to a change in place of articulation. This is a topic for future work, however.

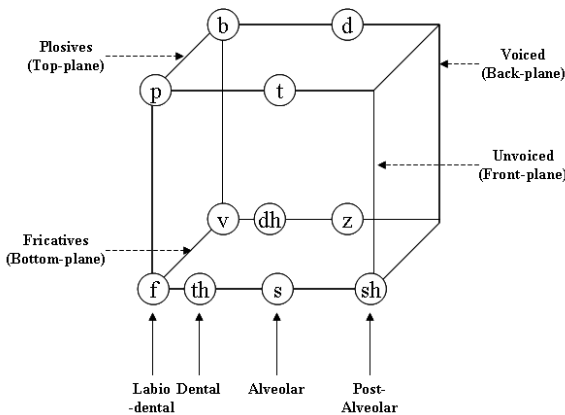


Figure 1: Cube representation of phonetic neighbourhood

In the following subsections, the phonetic neighbourhoods of the phones [f] and [th] are discussed in more detail in the context of the substituted phones in table 2.

5.3.1. Phonetic Neighbourhood of [f]

The position of phone [f] should be regarded as the starting point in the cube of figure 1. The consonants which appear as the substitutions for [f] in table 2 can be seen within the cube and their relationship is represented with respect to the four planes indicated in the figure and are as follows: unvoiced fricatives([th], [s]), voiced fricatives([v]) and unvoiced plosives([p], [t]).

At first glance, the LR context substitutes appear to be in rank order from table 2. Cautiously, due to the small amount of data, the ranking for these substitutes may be partitioned into two groups: first - [th]; second - [t], [s], [v], [p], [ih] and [sil] (silence). The phone [th] is the obvious first choice for substitution as its place of articulation is almost the same as that of phone [f]. They also share the same manner of articulation and are both unvoiced. The second group is more diverse with respect to the phonetic neighbourhood or proximity of [f]. [t] is of place alveolar, two places removed from the place of articulation of [f] and has a different manner of articulation also.

According to this depiction, moving one step in any dimension from [f] for example, phones [th], [p] and [v] would appear to have the closest phonetic neighbourhood

to the phone [f]. The phonetic neighbourhood may be measured in terms of relative distance between points on the cube. Phonetic neighbourhood alone does not appear to be a complete determining factor (albeit close) for the substituted alternative for phone [f] as the phones [t] and [s] seem to have as much priority as [v] and [p].

5.3.2. Phonetic Neighbourhood of [th]

The notion of phonetic neighbourhood for [th] is also accounted for in figure 1. In this case, the position of phone [th] should be taken as the starting point in the cube. Unlike the ranking consideration that was applied to the substitutions of phone [f], no ranking order is applied in the case of substitutions of [th] listed in table 2 as the numbers do not allow for a differentiation between the substitutions. It is not surprising that the substituted phone [t] is on the substitution list of phone [th] as [t] is only one place of articulation removed from [th]; however, it also has a different manner of articulation. Other principled reasons for the presence of [t] on this list become clearer when the frequency of occurrence is taken into account later in this paper. Therefore if phone [t] is a substitution possibility, its not surprising that its voiced counterpart [d] is also a substitution possibility. [dh] is also an obvious substitution as it is the voiced counterpart of [th]. The phone [s] also is just one place of articulation to the right of [th] with respect to the cube.

5.4. Frequency of occurrence

Another factor which is likely to be a determining factor for the substituted alternative is frequency of occurrence in the corpus as a whole. This factor relates to the notion of markedness in phonology which postulates that the most unmarked (or default) sound in a language is also likely to be the sound which is most common. The frequencies of occurrence of the phones in the full TIMIT corpus (excluding SA as described in section 2.1) together with their associated percentage of the corpus they cover are presented in table 5.

Phone	Frequency of phone in full TIMIT corpus	Percentage of phone in full corpus
t	8578	5.28%
s	8348	5.14%
d	5918	3.65%
p	4015	2.47%
dh	3272	2.02%
f	3126	1.93%
v	2704	1.67%
th	1004	0.62%
All 39 phones	162359	100%

Table 5: Frequencies of phones of [f] and [th] in the full speech corpus

These frequency of occurrences can also be visualised in the with respect to the phonetic neighbourhood cube as depicted in figure 2, where the larger the circle, the greater the frequency of occurrence.

The common phonetic properties, the phonetic neighbourhood and the frequency of occurrence in the corpus,

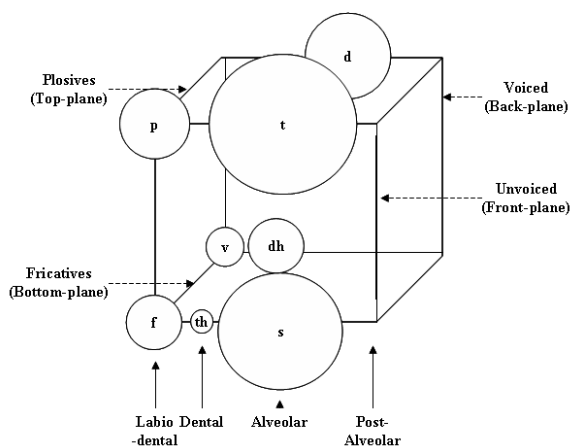


Figure 2: Representation of phonetic neighbourhood with respect to the most frequent consonant phones

together determine how the results of the underspecified systems for [f] and [th] should be interpreted. This is discussed in detail in the next section.

5.5. Effect of three determining factors

The three determining factors for substitution alternatives identified thus far allow an initial explanation for the pattern of the alternatives found for phone [f] and [th]. As can be seen from table 5, for the underspecified system for [f], the alveolar phones [t] and [s] are the most frequent substitutions. Similarly for the underspecified system for [th], the alveolar phones [t], [s] and [d] are the most frequent consonant substitutions. This high frequency of occurrence around the alveolar region has a strong effect on the substitution possibilities. The closer the removed phone is to this region, the greater the influence this region has on the substituting possibilities; these substitution possibilities align with the sounds which would typically be considered to have unmarked features (Chomsky and Halle, 1968), often also considered in the context of underspecification as default features.

5.5.1. Effect on [f] substitutions

- phone [th] is a strong obvious choice as the top substitution. The place of articulation of [f], *labio-dental*, is extremely close to the *dental* feature of [th]. All other phonetic properties of [f] are common to [th] as seen in table 3.
- phones [t] and [s] both with the property *alveolar* are also substitutions for [f]. Here the determining factor is the frequency of occurrence which takes precedence over phonetic neighbourhood.
- phones [v] and [p] are one position removed from [f] in terms of phonetic neighbourhood albeit it with respect to different planes of the cube (i.e. voicing as opposed to manner of articulation).

5.5.2. Effect on [th] substitutions

- phone [t] is just one place of articulation removed from [th] and has a high frequency of occurrence factor.

- phones [dh] and [d] (the voiced counterparts of [th] and [t]) are in close proximity in terms of phonetic neighbourhood but outweighed by the higher frequency of occurrence of [t]
- since the most frequently substituted phone for [f] is [th], it would seem reasonable to assume that the most frequently substituted phone for [th] should be [f]. However, the number of substitutions of phones [f] and [s] for [th] does not increase in the underspecified system for [th] although they have a relatively high frequency of occurrence; this indicates that perhaps an additional determining factor should be considered, namely the frequency of occurrence of the removed phone itself (i.e. [th] only constitutes a very small portion of the data).

It is apparent from the analysis of the effects on [f] and [th], that phonetic similarity is a combination of a number of determining factors which are balanced in a particular way. In some cases, the frequency factor outweighs the proximity in the phonetic neighbourhood and in other cases phonetic neighbourhood takes precedence.

6. Conclusion

This paper has presented a novel approach to the identification of phonetic similarity using properties observed during the speech recognition process. An experiment was presented whereby the phones [f] and [th] were separately removed during the training phase of a statistical speech recognition system so that the behaviour of the system could be analysed to see which alternative phones were selected. The substitution alternatives found by the underspecified system for the phone [f] and [th] were analysed with respect to three determining factors: common phonetic properties, phonetic neighbourhood and frequency of occurrence of the phone in the full corpus. The other fricative phones yielded similar results but were not presented in this paper.

The frequency of occurrence of a substituted phone was seen to have a strong effect on its ranking among the alternatives although phonetic neighbourhood and phonetic properties also played an important role. In the context of the results and analysis presented in section 4. and 5., phonetic neighbourhood was interpreted without resort to an exact measure of distance between the points on the cube. As mentioned there, a more detailed explication of the role of phonetic neighbourhood may be possible, based on a numerical distance measure and a differentiation between the relative weightings of the dimensions of the cube. For example, it may be that remaining on the *place* of articulation axis represents closer proximity than moving to the *voicing* axis for some occasions; it would certainly appear reasonable that a change in the manner of articulation would involve a greater phonetic distance in comparison to a change in place of articulation.

A factor which will be taken into account in future work is the extent to which the context (prosodic position and segmental context (Chomsky and Halle, 1968)) influences the substituted alternative; for example a substituted

phone may have emerged as a result of an influence of the preceding or following context/phone.

All of these points will be considered in the next phase of experimentation and the experiments will be extended to include other broad classes of phones with the aim of providing a principled methodology for the prediction of phonetic similarity for the purposes of speech recognition. A phonetic similarity measure can serve as an important additional source of information for the construction of acoustic models in statistical speech recognition, for enhancing the lexicon with appropriate phonetic variants and for the design of knowledge-based feature detection engines. In summary, future work envisages that these phonetic similarity measures will be used with the output of an ASR system so that if a confidence metric is low for some output, then the phonetic similarity measure will enable the system to offer an appropriate alternative.

7. References

- Chomsky, N. and M. Halle, 1968. The sound pattern of english. *Harper & Row*.
- Garofolo, J., L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, 1993. The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM.
- Ghiselli-Crippa, T. and A. El-Jaroudi, 1991. Voiced-unvoiced-silence classification of speech using neural nets. *IJCNN proceedings*:pp 851–856.
- Halberstadt, A. and J. Glass, 1997. Heterogeneous acoustic measurements for phonetic classification. *Eurospeech proceedings*:pp 401–404.
- Lee, K.-F. and H.-W. Hon, 1989. Speaker-independent phone recognition using hidden markov models. *IEEE Transactions on Acoustic, Speech, and Signal Processing*, Vol. 37, No. 11:pp 1641–1648.
- Mauclair, J., D. Aioanei, and J. Carson-Berndsen, 2009. Exploiting phonetic and phonological similarities as a first step for robust speech recognition. *EUSIPCO proceedings*.
- Robinson, T. and F. Fallside, 1990. Phoneme recognition from the timit database using recurrent error propagation networks. *Technical Report CUED/F-INFENG/TR.42, Cambridge University Engineering Department*.
- Scanlon, P., D. Ellis, and R. Reilly, 2007. Using broad phonetic group experts for improved speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3:pp 803–812.
- The-International-Phonetic-Alphabet, 2005. <http://www.langsci.ucl.ac.uk/ipa/>.
- Thuan, P. Van and G. Kubin, 2005. Dwt-based phonetic groups classification using neural networks. *ICASSP proceedings*:pp 401–404.
- Young, Steve, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying (Andrew) Liu, Gareth Moore and Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland, 2009. Hidden markov model toolkit (htk). <http://htk.eng.cam.ac.uk/>, Version 3.4.1.