CREATING SPEAKER INDEPENDENT HMM MODELS FOR RESTRICTED DATABASE USING STRAIGHT-TEMPO MORPHING

Alexandre Girardi, Kiyohiro Shikano, Satoshi Nakamura[1]

[1]Nara Institute of Science and Technology
Takayama-cho 8916-5, Ikoma-shi, Nara-ken 630-0101 Japan
E-mail: alex-g@is.aist-nara.ac.jp

ABSTRACT

In speaker independent speech recognition, one problem we often face is the insufficient database for training. A system trained only with several male and female databases will likely lack the information which is present in speakers with different pitch and vocal tract lengths. In an extreme case, children database, that is not easy to obtain, is a good example of how different pitch and spectral frequency stretch affects in a real speaker independent speech recognition system. In this paper, as an approach to solve the above problem, we study the effect of a combined change in the pitch and spectral frequency stretch of the original utterances in the database, in order to construct more robust HMM acoustic models. We study this effect by constructing morphed speaker databases which are converted from available male and female databases to target female and children voices respectively. Using the morphed database, we analyzed the level of improvement that can be obtained, in terms of recognition rate, compared with the real database. The recognition rate of the female voice recognized with male models improved from 60.2% to 87.3% with the morphed models. This result attests the effectiveness of the proposed method and the strong influence of the combined *pconv* (pitch conversion rate) and *fconv* (frequency conversion rate) have in the quality of the acoustic models. The result also attests the usefulness of the proposed algorithm for estimating *pconv* and *fconv*. In this paper, experiments with male and female data morphed towards 6 children voices are carried out. The isolated effect of *pconv* and *fconv* is also reported.

1. INTRODUCTION

One of the major problems in LVCSR (large vocabulary continuous speech recognition) systems is the insufficient amount of training data. This problem is even more sensitive when we talk about children data.

At the same time, recently a high-quality morphing algorithm named STRAIGHT-TEMPO [1] [full detailed description can be obtained in the draft paper for CASA-IJCAI, 1997 in http://www.hip.atr.co.jp/~kawahara/STRAIGHT.html], with excellent morphing qualities capable of even 600% manipulation of speech parameters such as pitch, vocal tract length and speaking rate, while keeping the human like naturalness of the voice.

We decided to use STRAIGHT-TEMPO to increase the database diversity, by adding the usually large amount of adult voice data morphed towards children voice data. We evaluate how STRAIGHT-TEMPO performs in a extreme case like morphing data towards children voice data.

STRAIGHT-TEMPO [1][2] is a combination of three basic tools: STRAIGHT, TEMPO and SPIKES. With STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrogram) we can construct a spectral time-frequency envelope of the signal, free of the periodicity effects in time-frequency analysis. This time-frequency envelope can be manipulated, for example in the time and frequency domains. As an additional input STRAIGHT uses the F0 estimation obtained with TEMPO (Time-domain Excitation extractor based on Minimum Perturbation Operator) algorithm. The fundamental frequency F0 can also be manipulated and together with the manipulated spectral time-frequency envelope obtained in STRAIGHT we can synthesize a morphed version of the original signal using SPIKES(Synthetic Phase Impulse for Keeping Equivalent Sound).

The following sections will describe the way we estimated the parameter necessary to morph the adult data to children voice using STRAIGHT in section 2. The second section describes the data base used in section 3. The third section describes the experimental results in section 5, and the final section comment the results obtained in section 6.

2. PITCH AND FREQUENCY CONVERSION RATE ESTIMATION

This section describes the *pconv* (pitch conversion rate) and *fconv* (frequency conversion rate) estimation algorithms used in this paper. *pconv* and *fconv* are the average rate of conversion for the fundamental frequency and spectral stretch between two speakers. We estimated *pconv* and *fconv* from the vowels parts of speech. The vowels were obtained by using HTK to get alignment of the utterances of both base and target speakers.

Some additional cares have been taken:

- just alignments taken from correct recognitions were used.
- aligned vowel samples have been energy normalized before processed

To present the expression we used to estimate *pconv* and *fconv* we will first introduce some notation.

Consider a pair of speakers (x, y), where the speaker x is the base speaker and the speaker y is the target speaker.

From both speakers we extract the recognized vowels parts, producing vectors of fundamental frequency $F0_s(t)$ and spectrum $|A_s(t,f)|$, where t is the time of the sample and f is the frequency in the spectrum for the speaker s. $F0_s(t)$ and $|A_s(t,f)|$ are obtained from STRAIGHT by applying $pconv = 1.0$ and $fconv = 1.0$.

Aligning these recognized vowels between both speakers x and y, with vowels of same kind we get pairs of ranges for the fundamental frequency

$$(F0_x(t_{xs}(p), t_{xe}(p)), F0_y(t_{ys}(p), t_{ye}(p)))$$

and for the spectrum

$$(A_x(t_{xs}, t_{xe}(p), f), A_y(t_{ys}(p), t_{ye}(p), f)$$

where p is the pair number with $p \in 1, .., P$ and P is the maximum number of pairs we can align within the available data. The first index of the time t refers to the speaker and the second to the start s and the end e of the alignment respectively.

Using this aligned pairs of vowels we can estimate $pconv$ and $fconv$, according to the following equations (1), (2), (3) and (4).

$$N_x(p) = t_{xe}(p) - t_{xs}(p) \quad (1)$$

$$N_y(p) = t_{ye}(p) - t_{ys}(p) \quad (2)$$

$$pconv = \frac{1}{P} \sum_{p=1..P} \frac{\frac{1}{N_y(p)} \sum_{t=t_{ys}(p)..t_{ye}(p)} F0_y(t)}{\frac{1}{N_x(p)} \sum_{t=t_{xs}(p)..t_{xe}(p)} F0_x(t)} \quad (3)$$

$$fconv = \frac{1}{P} \sum_{p=1..P} \frac{\frac{1}{N_x(p)} \sum_{t=t_{xs}(p)..xe(p)} \sum_f f \log |A_x(t,f)|}{\frac{1}{N_y(p)} \sum_{t=t_{ys}(p)..t_{ye}(p)} \sum_f f \log |A_y(t,f)|} \quad (4)$$

3. DATABASE

The voices of 6 children have been recorded at 48kHz and downsampled to 16kHz. The ages and gender of the children are shown in Table 1. The 210 words uttered by children are common use words for children in their age.

Table 1: Children database description

children	age	gender
MCH01	7	male
MCH02	6	male
MCH03	7	male
FCH01	6	female
FCH02	7	female
FCH03	7	female

The adult data we used was the MHT and FSU 5240 words and MHT, MAU, MMY, FKN, FSU and FYM 216 balanced words of ATR SetA [3].

4. PARAMETERIZATION

Table 2 describes the experiment conditions (parameterization used) for the following experiment as well as for the rest of the experiments in this paper.

Table 2: Parameterization

speakers	6 adults, 6 children
sample rate	16 kHz
frame shift/length	10 / 25 ms
emphasis	0.97
1 stream	(12MFCC) (12DMFCC DE)

All the recognition rates in the next section refers to tied mixture models. 55 phoneme models were used.

5. EXPERIMENTS

The first experiment aims to confirm the reliability of the estimation of $pconv$ (pitch conversion rate) and $fconv$ (frequency conversion rate). The second experiment will evaluate how this algorithm works when applied to children data.

5.1. PITCH AND FREQUENCY RATE ESTIMATION

To evaluate $pconv$ and $fconv$ estimation a male (MHT ATR SetA [3]) speaker was morphed towards a female voice with appropriate constant $pconv$ and $fconv$ parameters, generating a morphed female voice (FMHT).

The values of $pconv$ and $fconv$ which are used to convert MHT speaker data towards FSU speaker data, as well as their reestimated values, are shown in Table 3.

Table 3: Reestimated $pconv$ and $fconv$ are close to values used for conversion of MHT to FMHT

Parameter	Used	Estimated
pconv	2.22	2.17
fconv	1.25	1.27

The word recognition rate using MHT, FSU and the morphed FMHT data are shown in Table 4. Models are trained with the odd numbered words and tested against even numbered words of the ATR SetA database.

Table 4: Word recognition rate (%) FMHT is the MHT speaker morphed towards FSU speaker

Train	Test Speaker		
Model	MHT	FMHT	FSU
MHT	99.2	51.7	60.2
FMHT	49.9	98.5	87.3
FSU	64.3	84.5	98.7

From this result we conclude that the morphed data, approximated to the real female data, increase the recognition rate from 60.2% to 87.3%.

Also for an adult male voice morphed towards a female voice, $pconv$ and $fconv$ estimation diverge only 2%, attesting the robustness of the algorithm.

5.2. MORPHING ADULT DATA TOWARDS CHILDREN DATA

The next experiment aims to evaluate the efficiency of STRAIGHT morphing adult voice towards children voice. We used 6 adult speakers (3 male, 3 female ATR SetA) with 216 balanced words each. For comparison we used 6 children (3 male, 3 female), each children uttered 210 words containing all the Japanese phonemes.

Three types of acoustic models were created:

- the first was by using only children voices.
- the second by using only adult voices.
- the third by using only morphed data.

The estimated *pconv* and *fconv* between 6 adults and 6 children have been evaluated. Results can be seen in Table 5.

Table 5: *pconv* and *fconv* between speakers

from speaker	to speaker	pconv	fconv
MHT	MCH01	2.08	1.19
MAU	MCH02	2.22	1.28
MMY	MCH03	1.43	1.12
FKN	FCH01	1.02	1.29
FSU	FCH02	1.06	1.24
FYM	FCH03	0.91	1.12

Table 6: Word recognition rate (%) using "male adult model"[/"morphed model"]

child		adult		
		MHT	MAU	MMY
		100.0	98.2	98.2
MCH01	85.2	17.1/68.3	4.8	18.6
MCH02	59.5	1.4	1.0/39.2	2.4
MCH03	90.0	17.1	7.6	22.4/63.8
FCH01	73.3	4.3	2.4	4.8
FCH02	70.0	1.4	0.5	0.0
FCH03	73.3	8.6	5.2	14.8

Table 7: Word recognition rate (%) using "female adult model"[/"morphed model"]

child		adult		
		FKN	FSU	FYM
		95.4	96.3	93.5
MCH01	85.2	65.7	67.6	61.0
MCH02	59.5	39.5	32.4	45.2
MCH03	90.0	71.9	69.1	68.1
FCH01	73.3	49.1/49.1	43.8	51.4
FCH02	70.0	28.1	19.1/31.0	37.6
FCH03	73.3	35.2	41.0	41.4/33.3

Four experiments were them performed:

- First, children voice was recognized with the model generated from the remaining children voices (correspond to the first collum in Table 6 and Table 7).
- Second, adult voice was recognized with the model generated from the remaining adult voices (correspond to the first row in Table 6 and Table 7).
- Third, children voice was recognized with the model generated from adult voices (correspond to values that fill the central part of Table 6 and Table 7).
- Fourth, children voice was recognized with the model generated from adult voices morphed towards the children test data (correspond to the values after a backslash in Table 6 and Table 7).

All experiments are carried out within the same gender and age. Table 6 and Table 7 shows the recognition results. The first letter of the speaker name represents its gender, where male and female are represented by M and by F, respectively.

5.3. FIX PITCH AND FREQUENCY CONVERTION RATE

In oder to compare the estimation of *pconv* and *fconv* with the optimum values, we carried out recognition using models morphed in steps of 0.05 for each parameter, close to the estimated values. These recognition results are shown in Tables 8 and 9.

Table 8: Confirming *pconv* and *fconv* male estimations by carrying out recognition using near to optimum *pconv* and *fconv*. Recognition rate is expressed in word accuracy (%).

MHT to MCH01			
	fconv		
pconv	1.15	1.20	1.25
2.00	46.7	57.6	61.4
2.05	57.1	60.0	68.6
2.10	60.2	68.2	62.1

MAU to MCH02			
	fconv		
pconv	1.25	1.30	1.35
2.15	22.9	30.0	38.6
2.20	26.2	38.3	35.7
2.25	28.7	34.8	34.3

MMY to MCH03			
	fconv		
pconv	1.05	1.10	1.15
1.40	44.3	51.4	51.3
1.45	47.6	63.0	52.1
1.50	41.9	49.5	49.1

The fixed *pconv* and *fconv* experiments resulted in the highest recognition near the optimum values (see Table 5) obtained with the estimation algorithms (equations (1) and (2)). This attests the robustness of the proposed algorithms.

For a male data morphed towards children data, a higher increase of the recognition rate was achieved, while the female data presented almost no significant improvement. This shows that additional degrees of manipulation are necessary to morph adult data towards children data.

Table 9: Confirming *pconv* and *fconv* female estimations by carrying out recognition using near to optimum *pconv* and *fconv*. Recognition rate is expressed in word accuracy (%).

FKN to FCH01			
	fconv		
pconv	1.25	1.30	1.35
0.95	40.0	28.1	22.9
1.00	39.3	31.9	23.8
1.05	38.4	33.7	21.5

FSU to FCH02			
	fconv		
pconv	1.20	1.25	1.30
1.00	29.5	25.7	15.7
1.05	30.0	25.7	17.6
1.10	31.4	29.5	18.1

FYM to FCH03			
	fconv		
pconv	1.05	1.10	1.15
0.85	41.5	31.5	27.3
0.90	43.3	34.5	29.1
0.95	40.9	31.9	28.6

6. CONCLUSIONS AND FUTURE WORK

This paper presented an alternative way to increase the database for HMM acoustic model generation by using the high-quality STRAIGHT-TEMPO algorithm.

Morphing adult data towards adult data the algorithm increased the female voice recognition rate using models trained with male data from 60.2% to 87.3% with morphed data.

The algorithms proposed for pitch and frequency conversion rate estimation proved to be robust for adult data.

The increase in the word recognition rate for children data, when adult data is morphed towards children data, attests the usefulness of the proposed method for both male and female adult data.

Adult data is morphed towards children data increase the word recognition rate for children data, which attests the usefulness of the proposed method for both male and female adult data. These way large amounts of adult male and female data can be morphed to match children data, while each children only need to record small amounts of words each.

In the future we plan to investigate a non linear frequency conversion, as well as a more robust estimation of the frequency conversion rate, by adapting the frequency range used to follow the frequency conversion obtained.

ACKNOWLEDGMENT

This work is supported by CREST (Core Research for Evolutional Science and Technology), JAPAN.

7. REFERENCES

1. H. Kawahara. Speech representation and transformation using adaptive interpolation of weighted spectrum. Vocoder revised *IEEE int. Conf. Acoust., Speech and Signal Process.*, vol2, pages 1303-1306, Muenich, 1997.
2. H. Kawahara and de Cheveigne. Error free f0 extraction method and its evaluation. *Tech. Report of IEICE*, SP-96-96:9-18, 1997. (in Japanese).
3. H. Kuwabara, Y. Sagisaka, K. Takeda, and M. Abe, "Construction of ATR Japanese speech database as a research tool," Technical Report TR-I-0086, ATR, 1989. (in Japanese).