



# Robust F0 Modeling for Mandarin Speech Recognition in Noise

Sheng Qiang<sup>2</sup> Yao Qian<sup>1</sup> Frank K. Soong<sup>1</sup> Congfu Xu<sup>2</sup>

<sup>1</sup>Microsoft Research Asia, Beijing, China

<sup>2</sup>College of Computer Science and Technology, Zhejiang University

josephqiang@zju.edu.cn, {yaoqian, frankkps}@microsoft.com, xucongfu@zju.edu.cn

## ABSTRACT

The F0 contour plays an important role in recognizing spoken tonal languages like Mandarin Chinese. However, the discontinuity of F0 between voiced and unvoiced transition has traditionally been a bottleneck in creating a succinct statistical tone model for automatic speech recognition applications. By applying successfully the Multi-Space Distribution (MSD) to tone modeling, we recently reported a relative 24% reduction of tonal syllable errors on a Mandarin speech database. In this paper, we test MSD further in a noisy, continuous Mandarin digit recognition task, where eight types of noises are added to clean speech signals at five SNRs. The experimental results show that our MSD-based digit models can significantly improve the recognition performance in noise over a baseline system. Relative digit error rate reductions of 19.1% and 15.0% are obtained for noises seen and unseen in the training data, respectively. The improvements are also better than other reference systems where F0 information is incorporated.

**Index Terms:** Tone model, Mandarin speech recognition, MSD, Noisy digit recognition

## 1. INTRODUCTION

It is commonly believed that tone information plays an important role in recognizing spoken tonal languages like Mandarin Chinese for its lexical role. However, to construct a succinct tone model, which is critical for improving the performance of an automatic speech recognizer, is not a trivial task. The discontinuity in F0 contour between voiced and unvoiced transition is one reason why building a succinct F0 model is not so straightforward. Many ad hoc approaches like interpolating F0 in unvoiced segments to get around the problem have been proposed [1-4]. The interpolated F0 can be generated from a quadratic spline function [1], an exponential decay function towards the running F0 average [2], or a probability density function (pdf) with a large variance [3-4]. These approaches are instrumentally effective to incorporate F0 as extra information with other short-time spectral features frame synchronously. As a result, the concatenated spectral and pitch features are used as a frame synchronous feature in one-pass Viterbi decoding. However, the artificially interpolated F0 values do not reflect the actual tone and voicing or no-voicing information in the unvoiced region and the potential of pitch information for improving recognition performance can not be fully exploited. Furthermore, in terms of corresponding time window sizes, the spectral (segmental) feature is distinctive in a phonetic or phone segment while the pitch (supra-segmental) feature is embedded in a longer, say, word, phrase or sentence. By using a two-stream approach we can model spectral and pitch features more appropriately than a one-stream [5].

In our recent work [6], we adopted Multi-Space Distribution (MSD) [7], which uses two probability space approach, discrete probability density function (pdf) for unvoiced region and continuous pdf for modeling voiced pitch contour in tone

modeling. The tone features and spectral features are further separated into two streams and stream-dependent models are built (clustered) in separated decision trees. Experimental results of tonal syllable recognition on a Mandarin Chinese database show: 1) A 24% relative performance improvement of tonal syllable recognition error reduction is obtained; 2) Our approach outperforms conventional methods by 15% in relative reduction of tonal syllable recognition errors.

In this paper, we further test the MSD approach in a noisy, continuous Mandarin digit database to evaluate its robustness against noise. The rest of the paper is organized as follows. The approach of MSD-HMM for Mandarin Chinese tone modeling is first reviewed in section 2 and its application in noise is investigated in section 3. The experimental setups, results, comparison among different methods and error analysis are shown in section 4. In section 5, we give our conclusion.

## 2. MSD-HMM FOR TONE MODELING

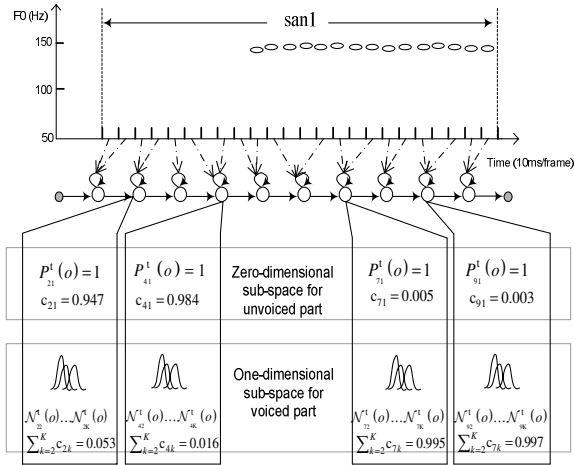
Multi-Space Distribution (MSD) was first proposed by Tokuda *et. al.*[8] to model the piece-wise continuous F0 trajectory stochastically for HMM-based speech synthesis. It assumes that the observation space  $\Omega$  of an event is made up of  $G$  sub-spaces. Each sub-space  $\Omega_g$  has its prior probability  $p(\Omega_g)$  and all its priors summed up to one,  $\sum_{g=1}^G p(\Omega_g) = 1$ . An observation vector,  $o$ , in each sub-space is randomly distributed according to an underlying pdf,  $p_g(o)$ . The dimensionality of the observation vector can be different from one sub-space to the other. The observation probability of  $o$  is defined by

$$b(o) = \sum_{g \in S(o)} p(\Omega_g) p_g(o) \quad (1)$$

where  $S(o)$  is the index set of the sub-spaces that observation  $o$  belongs to. It is determined by the extracted features at each time instant of observation. A mixture of  $K$  Gaussians can be seen as a special case of MSD, i.e.,  $K$  subspaces of MSD with the same dimensionality and a Gaussian distribution in each sub-space. The mixture weight associated with the  $k$ th Gaussian component  $c_k$  can be regarded as the prior probability of the  $k$ th sub-space  $c_k = p(\Omega_k)$ .

F0, the fundamental frequency, is the most relevant feature used in recognizing tone languages. But F0, a continuous variable, only exists in the voiced region of speech. In unvoiced segments where no harmonic structure exists in the signal, a discrete variable is adequate to characterize this un-voicing property. Fig.1 shows the F0 contour of a Mandarin digit 3 “san1” (the numerical label denotes its tone type: tone 1, a high flat tone), and the corresponding MSD model. The Mandarin digit 3 is a syllable consisting of an unvoiced consonants “s”, a vowel “a” and a voiced consonant “n”. F0 is only observed in the vowel nucleus “a” and nasal coda “n” but not in onset “s”. The conventional modeling can only characterize a feature as either continuous or discrete. The discontinuity of F0 between voiced and unvoiced segments makes tone modeling difficult.

MSD is an effective model to characterize the piece-wise continuous F0 contour without resorting to heuristic assumptions. In the voiced region, F0 is regarded as sequential one-dimensional observations generated from several one-dimensional Gaussian sub-spaces, while in the unvoiced region, F0 is treated as an indicator-like, discrete symbol. We use Gaussian mixtures, the most commonly used form in speech recognition systems, for characterizing the output distributions. The corresponding MSD assumes that the output pdf of the zero-dimensional, unvoiced sub-space is a Kronecker delta function and the one-dimensional, sub-space of the voiced sub-space has a Gaussian mixture distribution.



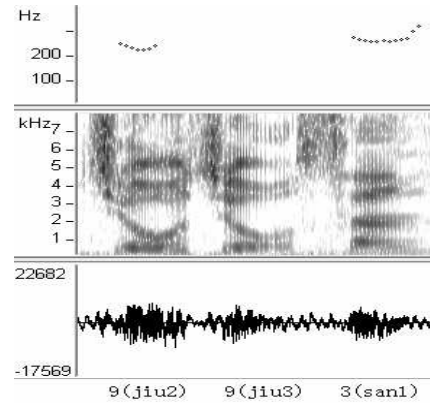
**Fig. 1** F0 contour of Mandarin digit “3 (san1)” and a schematic MSD-based, tone HMM

Fig.1 also gives a schematic representation of MSD-based tone HMM. At the beginning unvoiced part, the consonant onset “s”, the mixture weight which represents an unvoiced sub-space is close to one, while the weight summation of the Gaussian mixture components corresponding to the voiced sub-spaces is close to zero. At the syllable end, for the voiced, nasal coda “n”, the opposite is true. MSD tone modeling does not need any contour interpolation preprocessing of F0 trajectories. It models the original F0 feature in a probabilistic manner and no hard decision errors are made.

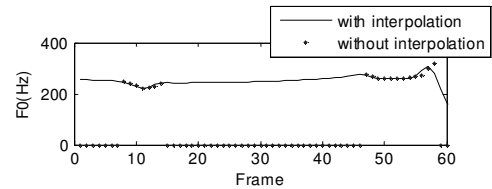
### 3. MSD-HMM TONE MODELING IN NOISE

Besides making a succinct tone model like MSD to describe the piece-wise continuous F0 trajectory of speech signal, it is probably even more challenging when pitch estimates and voiced/unvoiced decisions become more difficult to make in low SNRs. Fig 2 shows F0 contour, spectrogram, speech waveform of a digit sequence “9(jiu2)9(jiu3)3(san1)” corrupted by street noise under 5dB SNR. Due to noise contamination, F0 contour of the second digit “9” (jiu3) can not be successfully extracted and the incorrect raw F0 values or erroneous interpolated F0 based upon the mis-tracked pitch can have negative impact on recognition performance. However, MSD-based HMM, designed for modeling piecewise continuous F0 contour stochastically, is robust to noisy F0 feature in the recognition process. Since voiced and unvoiced observations are evaluated with either a continuous Gaussian mixture or discrete probabilities, misdetection of pitch can have negative but not disastrous effects on the likelihood computation. For example, no F0 in the 2nd digit “9” is evaluated as a stochastic event with a lower probability in MSD. If MSD model is trained

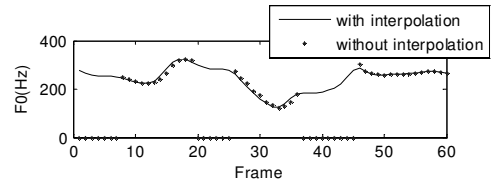
with both clean and noisy data, it will be more robust to pitch extraction errors because they are part of the training set.



**Fig. 2** F0 contour, spectrogram, waveform of a digit sequence “9 9 3” in street noise at 5 dB SNR.



**Fig.3.** F0 contours of a digit sequence “9 9 3” contaminated by street noise at 5 dB SNR with/without F0 interpolation



**Fig.4.** F0 contours of a digit sequence “9 9 3” (clean) with/without F0 interpolation.

A popular tone feature preprocessing employs a continuation algorithm [2] to interpolate the missing F0 values in unvoiced regions. The pitch is interpolated by running an exponential decay function towards the running average, plus a random noise. The target value of exponential decay function is usually set at the first F0 value in the next voiced segment. The entire F0 contour after interpolation is then smoothed through a low-pass filter. Figs 3 and 4 show the F0 contours of a digit sequence “993” in clean and noisy conditions, respectively, with/without F0 interpolations. The interpolated F0 values depend upon both the preceding and succeeding F0 values. Consequently, the interpolated F0 contour may deviate significantly from the true F0 values (if they are indeed in a voiced region but missed by the pitch tracking algorithm) or become artificial values in truly unvoiced regions. Furthermore, interpolating F0 values in a long unvoiced region can become difficult to implement for real-time applications, e.g. getting the target value for interpolation without a long look-ahead.

## 4. EXPERIMENTAL RESULTS AND ANALYSIS

### 4.1. Experimental Setup

The recognition experiments are performed on a noisy speech database of connected Chinese digits (CNDigits), consisting of

8,000 digit strings for training and 39,480 digit strings for testing. Training set consists of clean (1,600 sentences) and four different kinds of noise: waiting room, street, bus and lounge, four subsets for each type of noise, each subset contains 400 sentences from 120 female speakers and 200 male speakers, at specified SNRs, from 5 to 20dB, at a step of 5dB. Testing set consists of four noises (waiting room, street, bus, and lounge in the training set (**Matched noise**) and four additional noises unseen in the training set: platform, shop, outside and exit (**Mismatched noise**), five subsets for each type of noise, each subset contains 987 sentences from 56 female speakers and 102 male speakers, at specified SNRs, from 0 to 20dB, at a step of 5dB.

The acoustic features contain spectral features: 39-dimensional MFCC, including 12-dimensional cepstral coefficients, log energy and their first and second order derivatives; and pitch features: a 5-dimensional vector, consisting of log F0, its first and second order derivatives, and pitch duration and long-span pitch [8]. A whole-word HMM was trained for each of the ten Chinese digits (from “0” to “9” as “ling2”, “yi1 or yao1”, “er4”, “san1”, “si4”, “wu3”, “liu4”, “qi1”, “ba1” and “jiu3”). Each model consists of 10 left-to-right states without skipping. Each state output pdf is a mixture of 3 diagonal covariance Gaussians. Free word loop (i.e., no language model) is employed in the decoding.

## 4.2. Experimental Results

The configurations for experiments are listed as follows.

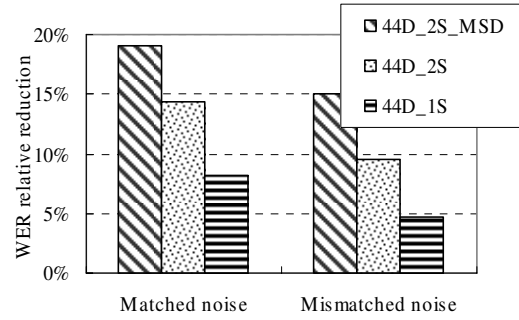
- 1) 39D\_1S: 39D MFCC, one stream
- 2) 44D\_1S: 39D MFCC, 5D pitch, F0 interpolation, one stream
- 3) 44D\_2S: 39D MFCC, 5D pitch, F0 interpolation, two streams
- 4) 44D\_2S\_MSD: 39D MFCC, 5D pitch, MSD, two streams

The recognition performance of our baseline system (39D\_1S) in different noises at various SNRs is shown in Table 1. It shows that recognition performance degrades with decreasing SNRs. The baseline system achieves an average 4.1% word error rate (WER) in clean condition.

**Table.1.** Recognition performance (WER) of 39D\_1S (baseline) for CNDigits.

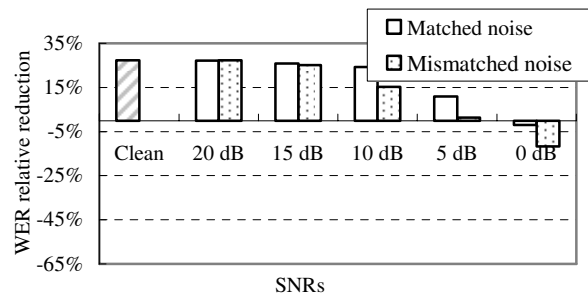
	Matched noise (%)				Mismatched noise (%)			
	Waiting Room	Street	Bus	Lounge	Platform	Shop	Outside	Exit
20 dB	3.9	4.5	3.7	4.2	3.6	4.5	3.6	4.0
15 dB	5.0	4.5	3.7	4.6	4.1	5.4	4.2	4.5
10 dB	8.8	5.4	3.9	5.7	5.1	8.6	5.2	5.4
5 dB	17.5	6.8	4.6	9.2	8.7	16.9	7.5	7.6
0 dB	32.6	11.1	6.2	19.2	17.9	34.5	15.2	14.3

Fig.5 shows average recognition performance of 44D\_2S\_MSD, 44D\_2S and 44D\_1S in relative WER reductions, compared with 39D\_1S. Among all three configurations, 44D\_2S\_MSD achieves the best performance. It yields 19.1% and 15.0% relative WER reduction averaged over all SNRs for matched and mismatched noises, respectively. 44D\_1S can also improve recognition performance over the baseline but in a much more limited way, compared with other two configurations. Separating tone and spectral features into two streams can further improve the recognition performance in noise. It avoids the problem in one stream where the output likelihood is dominated by spectral features due to their much larger dimensionality of than that of tone features.

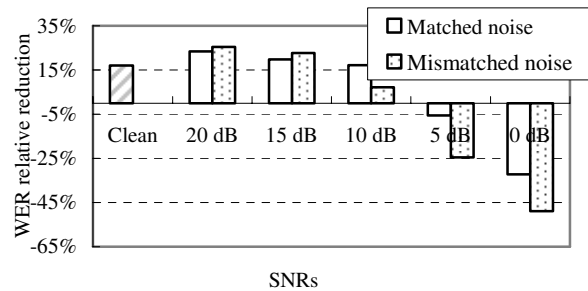


**Fig. 5.** The average recognition performance of 44D\_2S\_MSD, 44D\_2S and 44D\_1S for CNDigits in WER relative reduction, comparing with 39D\_1S.

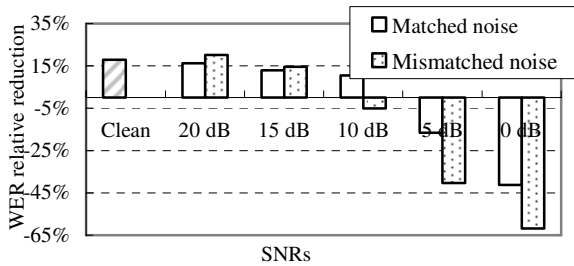
The breakdown of recognition performance of 4D\_2S\_MSD, 44D\_2S and 44D\_1S in clean and different SNRs from 20 down to 0dB noisy conditions is illustrated in Figs 6, 7 and 8. Fig. 6 shows that MSD-based tone modeling can significantly improve noisy Chinese digit recognition performance at SNRs from 20 to 10 dB. The maximum improvement of 27.4% in relative WER reduction is obtained at 20dB SNR, averaged over all mismatched noise conditions. Fig. 6 also shows that the performance improvements at SNRs 20 and 15dB are almost the same as that of clean speech. However, at 0 dB SNR, which is not included in the training data, the recognition performance is worse than that of 39D\_1S. It may be due to the fact that at such a low SNR, pitch extraction module fails to track the F0 contour. The Figs 7 and 8 show recognition performance of F0 interpolation is much worse than the baseline at low SNRs, e.g. 5 and 0 dB. It indicates that the interpolation method suffers more recognition performance loss from deteriorated pitch estimates in those two SNRs.



**Fig.6.** Recognition performance of 44D\_2S\_MSD for CNDigits in the WER relative reduction comparing with 39D\_1S.



**Fig.7.** Recognition performance of 44D\_2S for CNDigits in the WER relative reduction comparing with 39D\_1S.



**Fig.8.** Recognition performance of 44D\_1S for CNDigits in the WER relative reduction comparing with 39D\_1S.

### 4.3. Result Analysis

The recognition error patterns, or the confusion matrices, generated by MSD and interpolation-based F0 modeling, in lounge noise at 5 dB SNR are compared in Tables 2 and 3, for 44D\_2S\_MSD and 44D\_2S, respectively. MSD can significantly reduce digit deletion errors from 6.9% to 4.5%. Majority of deletion errors are associated with the semi-vowel, low pitch digit “5(wu3)”. The digit “5” is deleted frequently due to the fact that it is not well separated from preceding or succeeding digits, i.e., without unvoiced consonants to “protect” them from being merged together with adjacent digits.

## 5. CONCLUSIONS

We use MSD to model F0 or no F0 contour stochastically without resorting to artificially interpolating F0 in unvoiced regions. Tested in a noisy, continuous digit database, a significant recognition performance improvement is obtained over the baseline system at 5-20 dB SNRs of different noises.

## 6. ACKNOWLEDGEMENTS

The authors are grateful to Prof. Keiichi Tokuda and Dr. Heiga Zen, Department of Computer Science, Nagoya Institute of Technology, Japan for providing us MSD training tool HTS on their website: <http://hts.ics.nitech.ac.jp/>.

**Table.2.** The confusion matrix of recognition result using 44D\_2S under 5 dB SNR, lounge noise.

	lingyao	yi	er	san	si	wu	liu	qi	ba	jiu	Del
ling	708	1	15	0	0	1	6	3	2	0	<b>55</b>
yao	0	66	0	1	0	0	2	0	0	2	0
yi	2	3	610	0	0	1	0	2	10	1	77
er	0	2	0	687	1	1	3	3	0	20	37
san	0	0	1	1	838	17	0	0	0	2	11
si	0	0	0	1	4	730	0	0	2	0	9
wu	1	0	2	1	0	4	539	0	2	0	<b>290</b>
liu	15	2	2	1	0	0	5	683	1	0	13
qi	1	0	6	0	0	7	0	0	740	0	6
ba	0	0	0	6	1	1	0	0	0	821	9
jiu	5	3	0	1	1	0	5	6	23	0	<b>709</b>
Ins	1	0	9	2	2	16	24	1	6	0	0

**Table.3.** The confusion matrix of recognition result using 44D\_2S\_MSD under 5 SNR lounge noise.

	lingyao	yi	er	san	si	wu	liu	qi	ba	jiu	Del
ling	730	0	15	0	0	0	8	7	1	0	<b>29</b>
yao	0	62	0	4	0	0	1	3	0	1	0
yi	3	2	627	0	0	4	0	2	7	0	61
er	0	2	0	695	1	1	1	2	0	15	37
san	0	0	1	0	846	17	0	0	0	2	4
si	0	0	1	0	4	736	0	0	2	0	3
wu	1	0	1	0	1	5	648	0	1	0	<b>181</b>
liu	13	2	1	1	0	0	6	692	1	0	11
qi	0	0	6	0	0	9	0	0	743	0	3
ba	0	0	0	4	1	0	0	0	0	820	13
jiu	3	1	0	0	1	2	2	7	19	0	<b>16</b>
Ins	4	0	8	3	2	22	22	0	3	0	0

## 7. REFERENCES

- [1] Hirst, D. and Espesser, R., “Automatic Modeling of Fundamental Frequency Using a Quadratic Spline Function”, *Travaux de l’Institut de Phonétique d’Aix* 15, 71-85, 1993.
- [2] Chen, C. J., Gopinath, R. A., Monkowski, M. D., Picheny, M. A., and Shen, K., “New Methods in Continuous Mandarin Speech Recognition”, In *Proc. Eurospeech 1997*, 1543-1546, 1997.
- [3] Chang, E., Zhou, J. L., Di, S., Huang, C., and Lee, K-F., “Large Vocabulary Mandarin Speech Recognition with Different Approach in Modeling Tones”, In *Proc. ICSLP 2000*, 983-986, 2000.
- [4] Freij, G. J. and Fallside, F., “Lexical Stress Recognition Using Hidden Markov Models”, in *Proc. ICASSP 1988*, 135-138, 1988.
- [5] Seide, F. and Wang, N. J. C., “Two-stream Modeling of Mandarin Tones”, In *Proc. ICSLP 2000*, 495-498, 2000.
- [6] Wang, H.L., Qian, Y., Soong, F.K., Zhou, J.-L., Han, J.Q.: A Multi-Space Distribution (MSD) Approach to Speech Recognition of Tonal Languages. In *Proc. ICSLP 2006*
- [7] Tokuda, K., Masuko, T., Miyazaki, N., Kobayashi, T.: Multi-space Probability Distribution HMM. *IEICE Trans. Inf. & Syst.*, E85-D(3), (2002) 455-464
- [8] Zhou, J. L., Tian, Y., Shi, Y., Huang, C., and Chang, E., “Tone Articulation Modeling for Mandarin Spontaneous Speech Recognition”, In *Proc. ICASSP 2004*, 997-1000, 2004.