

Yǒng Jǔ Fā Yīn: A Simple Mandarin Chinese Tone Recognizer

A Thesis presented

by

Jim Gilsinan IV

to

Computer Science

in partial fulfillment of the honors requirements

for the degree of

Bachelor of Arts

Harvard College

Cambridge, Massachusetts

April 2, 2001

Thesis Web Site:

<http://dunster1.student.harvard.edu/~gilsinan/thesis/>

Words came out of the womb of matter;
And whether a man dispassionately
Sees to the core of life
Or passionately
Sees the surface,
The core and the surface
Are essentially the same,
Words making them seem different
Only to express appearance.

— Lao Tzu

I. Introduction

The field of computer voice recognition has seen striking progress in recent years, due in part to the availability of cheaper and faster consumer microprocessors and memory subsystems, along with advances in digital signal processing and linguistic analysis methods. A majority of speech recognition efforts, however, have centered around English and, to a lesser extent, other European languages, as a result of the prominence of such languages in the technology and business sectors. But, whereas English is the most widely spoken *secondary* language, there is a language that boasts a native speaking population well over double the 322,000,000 for English—Mandarin Chinese, with some 885,000,000 native speakers (Grimes).

In addition to being the most popular first language in the world, there are other important features of Mandarin Chinese that make it an attractive target for speech recognition research. One of the most practical is that the non-phonetic, ideographic nature of Chinese writing does not lend itself at all well to traditional computer input frameworks, such as the keyboard. Current keystroke-based Chinese character entry systems are cumbersome, difficult to learn, and time-consuming, often taking five or more strokes to input a single character, with additional strokes often necessary to differentiate between homonyms. Whereas five or more strokes for a word in English orthography may not seem excessive, knowledge of the keystrokes for a particular character in Mandarin requires substantial training in a specific input

methodology, including extensive memorization of conventions and symbol locations, since these keystrokes may not have a phonetic basis (Fu, 1). Freed from the keyboard by voice recognition technology, even the highly complex “Traditional” Chinese ideograph system has the potential to be cleanly entered as fast as any phonetic writing system, such as the Roman or Cyrillic alphabets.

From the implementation perspective, though, there is one feature of Mandarin that stands out: the presence of lexical tone. Whereas in English we use changes in pitch to lend an additional connotation to a word or phrase, such as inquiry by raising the pitch of the speech over the course of an utterance (“This is a dog?”), or annoyance by dropping the pitch rapidly at the end of an utterance (“This is a dog!” — after requesting a cat), the pitch changes do not affect the intrinsic *lexical* meaning of any particular word: “dog!” and “dog?” both reference a canine animal, as does just “dog,” with no pitch change at all. In Mandarin, however, each syllable takes on a particular pitch inflection, a tone, as a lexical feature. The tone of a syllable is as important in Mandarin speech as the consonants and vowels of the syllable; changing the tone changes the fundamental meaning. There are four possible tones in Mandarin Chinese: high, rising, dipping, and falling, called tones one, two, three, and four respectively. There is also a “neutral” tone, sometimes thought of as a “fifth” tone in the system, and sometimes used as the tone of the second syllable in a two-syllable lexical cluster.

As an example of the way tones are used in Mandarin, one can examine the syllable “ma,” pronounced with each of the four tones. Said using the first (high) tone, “ma” means “mother”; with the second (rising) tone, “linen”; with the third (dipping) tone, “horse”; and with the fourth (falling) tone, “scold.” To hear the four tones of “ma,” listen to <http://www.research.ibm.com/beijing/projects/speech/fourtone.wav> (IBM Research) or

http://dunster1.student.harvard.edu/~gilsinan/thesis/jue_chen/ma.way. None of the most widely spoken Western languages incorporate lexical tone in speech, but it is an indispensable feature of Mandarin Chinese which requires certain additional analysis methods during the voice recognition process (Fu, 9).

Syllable	Meaning
mā (first tone)	mother
má (second tone)	linen
mǎ (third tone)	horse
mà (fourth tone)	scold

Table 1: Four Tones of “ma”

II. Method

In an effort to study the supplementary techniques required for accurate voice recognition of Mandarin Chinese due to the presence of lexical tone, we will concentrate on attempting to correctly identify the tone of arbitrary isolated syllables recorded by native Mandarin speakers. For simplicity, we will not attempt to deal with continuous speech, or multiple different-toned syllables, which may exhibit the phenomenon of “tone sandhi,” an application of the rules that govern combinatorial relationships between tones (Yeh, 1). In addition, we will study only the major four tones of Mandarin, excluding the neutral fifth tone, since it is used largely in multi-syllabic circumstances.

The corpus used in the application of the methods described below was compiled from two different sources. Recordings of native Mandarin speakers Zhongjue Chen and Tsiyun Cherry Fu, both of Harvard College, are a source of original test material. Over 100 speakers provided samples of the numbers one through ten on the Oregon Graduate Institute Multi Language Telephone Speech Corpus. The samples from Chen and Fu were used to test the efficacy of the method with same-speaker samples (tone templates were constructed from tone samples of the speaker who produced the test syllables), while the OGI corpus served as test material for speaker-independent analysis (tone templates were constructed from samples of a multitude of speakers, and test syllables were provided by a speaker independent of those involved in template production).

All sample waveform files were processed in Cool Edit Pro for normalization, noise-reduction, and file format conversion. The OGI corpus sample files were recorded at 8 KHz, in mono-channel 16-bit resolution, while the original test material from Chen and Fu was recorded at 44.1 KHz, in mono-channel 16-bit resolution. The resulting sound files were then analyzed

with the Speech Filing System for Win32 from the University College London Department of Phonetics and Linguistics to track pitch (fundamental frequency), with pitch measurements taken every 5 milliseconds. The results were exported to a standard ASCII file format using one integer pitch measure sample per line. All processed samples and pitch track analysis files are available on-line at <http://dunster1.student.harvard.edu/~gilsinan/thesis/>.

In Figure 1, we can see the four tones of Mandarin (marked 1, 2, 3, and 4) in both waveform and pitch-track view, generated with the Speech Filing System program:

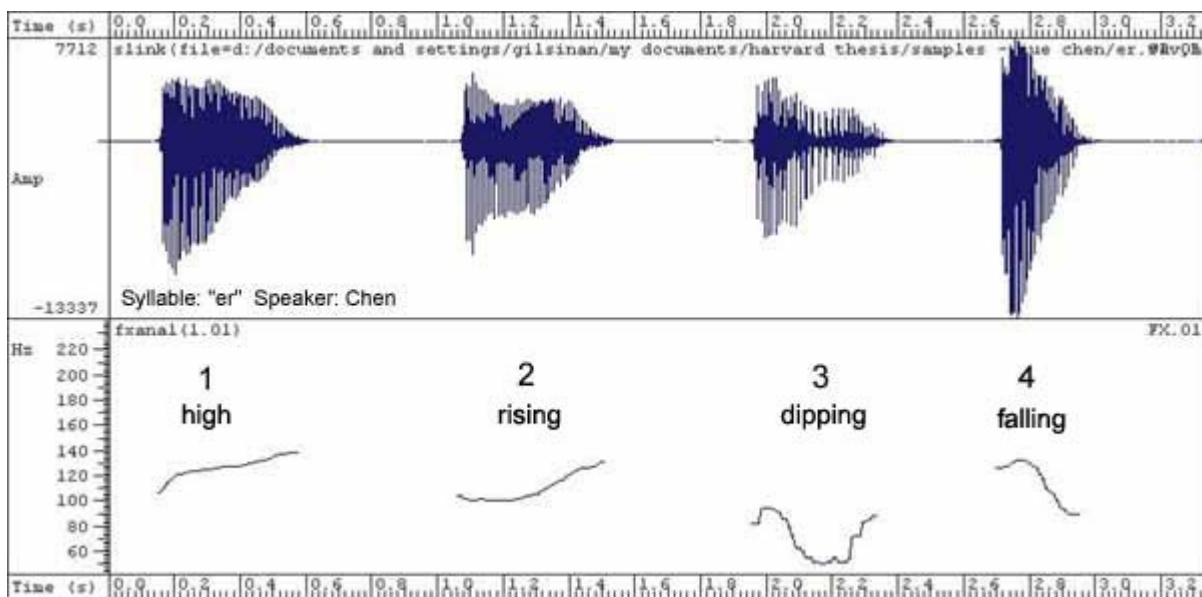


Figure 1: Four Tones of Mandarin, Waveform and Pitch-Track View

Note the clear correlation between the tone descriptions (high, rising, dipping, falling) and the shape of the pitch tracks.

Using the ASCII pitch track files, Yǒng Jǐu Fā Yīn, the tone recognizance program that forms the core of this thesis, analyses the shape of each syllable's track and correlates the result with similar analyses performed on a standard template track for each of the four tones. The template tracks were created from the corpus of sample files, by taking known instances of each

tone, normalizing their pitch tracks around their mean pitches, standardizing their duration with simple linear interpolation, and averaging the results.

The modular design of the main program allows for easy inclusion of arbitrary statistic analysis methods. For the purpose of this thesis, simple mean squared distance analysis (Press 657-661), best-fit linear regression analysis (Rice, 511-523; Press 661-666), and best-fit polynomial regression analysis (Press 681-688) have been included. Analysis begins by reading the syllable pitch tracks from the specified ASCII pitch track file, creating an array based on the sample point pitch measurements, and stacking each syllable's resulting pitch track array into a syllable array for easy access. The template pitch tracks are read in analogously.

Mandarin is a “contour” tone language: the lexical tone of a syllable is determined by the shape of the pitch track (the change in relative pitch across the syllable) and is not dependent on the absolute pitch position of the syllable. This contrasts with “register” tone languages, such as Nupe, in which the lexical tone of a syllable is wholly dependent on its absolute pitch (Katamba 188-192). To eliminate differences caused by speaker- or circumstance-specific variations in absolute pitch, we thus calculate the mean pitch of each syllable from its pitch track array, and subtract this number from each element to form a pitch track normalized in pitch space. As a final step before statistic analysis, we eliminate durational differences between utterances by linearly interpolating the test and template pitch tracks into a standard array length, `#defined` to be 200 (i.e. one second, since pitch readings are taken every 5 milliseconds), which should be sufficiently large to contain almost all syllabic utterances without loss of pitch data.

Using these standardized and normalized pitch arrays, we may then proceed with the command-line specified statistical analysis. Mean squared distance analysis in this context simply sums the absolute differences in pitch between each element in the standard test array and

its corresponding element in the template pitch array for each of the four tones. The tone comparison that produces the lowest absolute total pitch difference is then returned as the calculated tone of the test syllable.

For regression analysis, it is desirable to divide the syllable pitch track into cepstral regions (the number of which can be specified on the command line, and defaults to 4), apply the command-line specified regression to each of the regions individually, compare the results individually, and return the tone with the highest cumulative correlation. For example, while a best fit linear regression might characterize the relatively static first (high) tone over an entire syllable, it would not well characterize the third (dipping) tone over the entire syllable. Using linear or polynomial regression analysis, we may then compare the parameters of the resulting best fit lines to the corresponding parameters from the tone templates to determine the tone. Note that the intercept of linear best-fit lines and the first parameter of polynomial best-fit lines are not used (and could in fact lead to less accurate results), since the important feature of lexical tone is shape and not absolute pitch position.

Best-fit linear regression analysis (in `regression.c`) calculates the equation of the best fit line ($y = a + bx$) over the cepstral region using the following equations (Press 661-666):

$$b = \frac{(n \sum_{i=0}^{n-1} x_i y_i) - (\sum_{i=0}^{n-1} x_i)(\sum_{i=0}^{n-1} y_i)}{(n \sum_{i=0}^{n-1} x_i^2) - (\sum_{i=0}^{n-1} x_i)^2} \quad a = \frac{(\sum_{i=0}^{n-1} y_i) - (b \sum_{i=0}^{n-1} x_i)}{n}$$

where n is the number of samples in the cepstral region, the x_i are the indices with respect to the region, and the y_i are the pitch values at the indices. The calculated best-fit slope (b) for each cepstral region in the test syllable is then compared with the calculated best-fit slope of the corresponding cepstral region in each of the four template tones, and the tone of the template syllable with the least total absolute difference in slope is returned as the detected tone.

Best-fit polynomial regression analysis is slightly more involved (Press 681-688). We wish to find the degree $m - 1$ polynomial $f(x) = a_1 + a_2x + \dots + a_mx^{m-1}$ that best fits our n data points (x_i, y_i) for $0 \leq i \leq n - 1$. To do this, we minimize the sum of the differences of the squares by finding the point where the gradient vanishes in the following function:

$$S = \sum_{i=0}^{n-1} (y_i - f(x_i))^2 = \sum_{i=0}^{n-1} \left(y_i - \sum_{j=1}^m a_j x_i^j \right)^2$$

There is only one such point, and since the function is positive definite quadratic form, it will be a minimum: specifically, it is the point at which the derivative with respect to each a_k vanishes, and

$$\frac{\partial S}{\partial a_k} = - \sum_{i=0}^{n-1} 2x_i^k \left(y_i - \sum_{j=1}^m a_j x_i^j \right).$$

We know this vanishes when

$$\sum_{i=0}^{n-1} \left(y_i - \sum_{j=1}^m a_j x_i^j \right) x_i^k = 0,$$

rearranging the terms of which yields

$$\sum_{j=1}^m \left(\sum_{i=0}^{n-1} x_i^k x_i^j \right) a_j = \sum_{i=0}^{n-1} y_i x_i^k. \quad (1)$$

The a_k 's which satisfy this for all $1 \leq k \leq m$ are thus the correct parameters for our best-fit polynomial. We can use linear algebraic methods to solve (1) in the following way. Define a matrix A such that

$$A = \begin{bmatrix} x_0^0 & x_0^1 & x_0^2 & \cdots & x_0^m \\ x_1^0 & x_1^1 & x_1^2 & \cdots & x_1^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n-1}^0 & x_{n-1}^1 & x_{n-1}^2 & \cdots & x_{n-1}^m \end{bmatrix} = \begin{bmatrix} 1 & x_0^1 & x_0^2 & \cdots & x_0^m \\ 1 & x_1^1 & x_1^2 & \cdots & x_1^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n-1}^1 & x_{n-1}^2 & \cdots & x_{n-1}^m \end{bmatrix}.$$

If we left-multiply A by its transpose, we have

$$A^T A = \begin{bmatrix} \left(\sum_{i=0}^{n-1} x_i^0 x_i^0 \right) & \left(\sum_{i=0}^{n-1} x_i^0 x_i^1 \right) & \cdots & \left(\sum_{i=0}^{n-1} x_i^0 x_i^m \right) \\ \left(\sum_{i=0}^{n-1} x_i^1 x_i^0 \right) & \left(\sum_{i=0}^{n-1} x_i^1 x_i^1 \right) & \cdots & \left(\sum_{i=0}^{n-1} x_i^1 x_i^m \right) \\ \vdots & \vdots & \ddots & \vdots \\ \left(\sum_{i=0}^{n-1} x_i^m x_i^0 \right) & \left(\sum_{i=0}^{n-1} x_i^m x_i^1 \right) & \cdots & \left(\sum_{i=0}^{n-1} x_i^m x_i^m \right) \end{bmatrix},$$

which is a square m by m matrix whose entries are the coefficients we need for the left side of (1) above. If we right-multiply $A^T A$ by the vector

$$a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix},$$

we have a vector the entries of which correspond to the left side of (1). For the right side, we simply use the vector $A^T y$, where

$$y = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n-1} \end{bmatrix}.$$

Thus, linear algebraically, the equations represented in (1) correspond to the matrix equation

$$(A^T A)a = A^T y,$$

which we can solve using Gaussian elimination (implemented in `gaussian.c`). Then, rather than compare the resulting best-fit parameters directly, as we were able to do with linear regression analysis, we use a pseudo- L^1 norm by evaluating the best fit polynomial (ignoring the constant term) at each point in the cepstral region in the template and test tones, and summing their absolute differences. The tone of the template with the lowest absolute difference between its best-fit polynomials and those of the test syllable is thus returned as the detected tone of the test syllable.

All code is available at <http://dunster1.student.harvard.edu/~gilsinan/thesis/>.

III. Results

1. Speaker-dependent Analysis

Tabulation of the speaker-dependent analysis results of running the tone recognizer with the test input syllables is as follows. Each of the two volunteer speakers (Zhongjue Chen, male, and Tsiyun Cherry Fu, female) provided the series of four tones for different syllables. For example, for a given syllable, “bao,” the speaker pronounced each of the four tones of “bao” in succession. This process was then repeated for several different syllables. Instances of each *tone*, across syllables, for each speaker were then averaged to provide template tones for each speaker’s voice, leaving one syllable out for each template as the “unseen” or “held out” data with which to test the method. Thus, for Chen, we tested a total of 20 syllables, 5 of each tone, against 20 corresponding templates constructed from all instances of each tone save the syllable being tested. The tone recognizer was run once with each test syllable set against the proper template, with each analysis method. In the tables below, the template tones 1 through 4 are represented along the top row, while the test tones 1 through 4 are represented along the first column. The numbers contained in the cells represent the score for the test syllable against the template syllable using the specified regression. The lowest score is thus the estimate of which of the four tones the test syllable best matches. Correctly identified syllables therefore follow the diagonal of the table from top left to bottom right, unshaded and thick-bordered. Correctly rejected syllables are unshaded, and incorrectly identified syllables are shaded. These charts provide insight into which tones were mis-recognized, and what these tones were mis-recognized as.

1.a. Speaker-dependent: Mean squared distance analysis, no cepstral regions.

(Score is total absolute difference in pitch between test and template syllables.)

“bao”, female	1	2	3	4
1	2254.751050	4534.716314	6385.530920	7701.547346
2	1900.856734	1338.100614	7091.799935	8545.371165
3	2765.515039	4221.848916	2948.785389	4802.393345
4	8010.845650	8836.086379	3917.792194	5272.212888

“chen”, female	1	2	3	4
1	2027.250399	3283.914244	3438.798056	6147.491430
2	2997.961822	1172.716156	5256.263672	7781.316135
3	8528.973251	9570.145739	4020.042417	3923.481448
4	8413.716768	9465.414908	3839.646066	3371.706551

“er”, female	1	2	3	4
1	642.041935	3055.899274	5446.017781	6846.573616
2	4869.422793	2700.783966	7865.343235	9163.191902
3	4330.740393	5846.190834	1228.477636	2722.074673
4	8819.176526	10504.527785	3867.288685	3137.040087

“hai”, female	1	2	3	4
1	1385.249874	2014.034764	5422.782068	6925.917443
2	3365.002631	651.875735	7037.864862	8470.374892
3	4109.378650	5284.846298	1561.253513	3567.512796
4	8009.622328	9434.426695	2792.582804	2493.950573

“lei”, female	1	2	3	4
1	990.742965	3646.786312	4863.628351	6827.189892
2	2510.306896	1293.357129	7130.614514	9129.292213
3	5693.157660	6811.361634	1227.403221	2261.675945
4	7210.808866	8610.386717	3145.154800	1702.010168

“ma”, female	1	2	3	4
1	1623.155634	3192.433831	5397.462238	8187.772103
2	3302.843963	904.765061	6217.812700	9109.229187
3	5809.735793	7531.920584	3798.383186	4696.612246
4	8414.956832	11082.622603	9067.155561	9635.499153

“er”, male	1	2	3	4
1	508.789649	1843.734402	4712.694530	6330.209029
2	1382.583994	803.171669	4914.611602	7151.964322
3	3281.009550	3862.412270	7532.090062	4745.874713
4	3674.144150	5645.377581	5688.108690	2283.306245

“hai”, male	1	2	3	4
1	812.882596	1444.510821	4337.489275	5477.686430
2	2596.908567	841.380343	4986.427372	6964.811578
3	3528.442749	5324.706414	5862.172549	1601.751450
4	7460.731634	9260.412433	9579.742816	2709.042828

“jing”, male	1	2	3	4
1	849.711811	2177.898493	4950.731280	4836.849798
2	2793.136642	1262.510415	5234.113420	8134.227730
3	3525.036529	3046.236955	7540.521068	6210.855467
4	6267.616527	7452.366974	7181.728045	1702.821953

“lei”, male	1	2	3	4
1	459.644068	2226.599286	2825.242100	5969.125545
2	1003.025615	1374.644526	2747.811929	6381.088546
3	24507.811920	24858.764587	27002.067407	26122.359486
4	4614.622399	6729.705597	4700.491999	2096.894540

“yu”, male	1	2	3	4
1	528.645559	2271.111903	4554.549830	5041.695930
2	2070.977546	909.852077	4354.891204	7496.417701
3	2815.225164	3276.869789	6840.849938	5423.324967
4	6005.663117	7388.743009	7176.170859	1604.857835

1.b. Speaker-dependent: linear regression analysis, 4 cepstral regions.
(Score is total difference in slope.)

“bao”, female	1	2	3	4
1	2.351934	4.313471	4.854193	3.920347
2	1.543013	1.409652	3.255397	3.559651
3	1.973458	1.643816	1.569352	2.977864
4	5.630229	5.209699	4.010981	5.459123

“chen”, female	1	2	3	4
1	1.692137	1.631368	1.266215	2.322982
2	2.763794	0.554653	2.379945	3.761785
3	3.675063	2.890357	1.496525	1.789085
4	3.982473	2.486260	2.098407	1.661000

“er”, female	1	2	3	4
1	0.443763	2.471123	3.048369	2.645064
2	4.782970	2.736478	4.536187	5.557282
3	1.670706	2.825957	1.263285	1.255757
4	3.796136	4.433438	2.385686	1.539587

“hai”, female	1	2	3	4
1	1.111393	1.928292	1.929891	2.102268
2	2.816625	0.538071	2.793503	3.666820
3	3.080716	1.417316	1.645620	2.518938
4	4.032438	3.601826	1.082360	2.103175

“lei”, female	1	2	3	4
1	1.086573	3.816635	2.639781	3.297500
2	1.501083	1.330039	2.816755	3.085168
3	3.288577	2.654535	0.925455	1.515384
4	2.155402	4.135914	2.787425	1.469703

“ma”, female	1	2	3	4
1	0.528878	2.397059	2.323960	3.603731
2	3.007527	0.618719	2.458389	3.195340
3	4.326929	4.318292	3.204729	3.761123
4	6.643480	8.249394	7.236050	7.722671

“er”, male	1	2	3	4
1	0.404096	1.052572	4.714935	2.076592
2	0.826333	0.388674	4.175457	2.353635
3	1.429973	1.913259	6.013839	1.921550
4	0.876928	2.022636	5.230474	0.795568

“hai”, male	1	2	3	4
1	0.143167	1.026000	4.226590	1.191862
2	1.729138	0.866521	4.826231	3.064168
3	1.146457	1.462737	5.044766	1.136903
4	3.170446	3.486726	5.239205	1.941834

“jing”, male	1	2	3	4
1	0.431383	0.951379	4.242299	1.531241
2	1.418506	0.569724	3.898194	2.683681
3	2.753264	1.873980	4.593522	3.260684
4	1.690767	2.378206	5.390013	0.811515

“lei”, male	1	2	3	4
1	0.299817	1.289122	1.662243	1.897403
2	0.524076	0.719742	1.072987	1.949992
3	19.204444	19.340196	20.141073	20.300391
4	1.288030	2.494549	2.602660	0.755122

“yu”, male	1	2	3	4
1	0.224637	1.059486	4.170683	1.634731
2	1.155628	0.642338	3.813735	2.500072
3	2.113565	1.398281	4.525841	2.607878
4	1.807910	2.796152	6.094800	1.518995

1.c. Speaker-dependent: quadratic regression analysis, 4 cepstral regions.
(Score is total pseudo-L¹ absolute norm difference.)

“bao”, female	1	2	3	4
1	67987.061303	98514.627046	85132.090020	81295.361677
2	1521.551900	30722.731954	17340.194955	13503.466506
3	7741.993217	22785.572637	9403.035618	5566.307269
4	122388.165598	91860.60014	105243.137013	109079.865

“chen”, female	1	2	3	4
1	10104.613886	30864.248951	19809.056671	38695.662921
2	28493.789100	12475.073784	1419.881422	20306.487180
3	24947.219887	16021.642976	4966.450644	23853.056814
4	20732.019329	20236.843629	9181.651230	28068.257394

“er”, female	1	2	3	4
1	2904.960936	20487.326382	27467.501024	41552.340164
2	110661.624074	87269.336937	80289.162289	66204.32354
3	12842.641140	10549.646256	17529.820868	31614.659797
4	32110.983991	8718.696523	2278.215417	12346.316891

“hai”, female	1	2	3	4
1	29759.1947	12828.073780	2327.725656	20028.603143
2	38071.531992	4515.736540	6774.729031	11716.265828
3	33979.718395	8607.550208	2682.915377	15808.079592
4	31213.649087	11373.619400	595.954789	18574.148748

“lei”, female	1	2	3	4
1	22489.554844	61855.217511	41930.478946	66867.693967
2	3241.237477	36124.425432	16199.686715	41136.901743
3	44577.909861	5212.246860	25136.985717	3411.682781
4	10579.788398	28785.874502	8861.135779	33798.351086

“ma”, female	1	2	3	4
1	39799.260511	6677.068856	7579.848658	9913.739888
2	42727.941598	871.936934	10508.529571	6985.058871
3	37000.010727	6599.867748	4780.598720	12712.989679
4	45007.736691	20127.735683	17045.579616	21529.377234

“er”, male	1	2	3	4
1	6985.119415	10506.299241	1769.441174	5317.117314
2	1600.181077	5121.361115	7154.379514	626.086213
3	11122.695547	14643.875659	2368.134953	9454.693479
4	2589.981241	6111.161248	6164.579346	921.979144

“hai”, male	1	2	3	4
1	2904.278155	1214.388521	11612.881600	4375.808866
2	9339.413122	6069.217055	18048.016474	10810.943822
3	2900.837988	6171.034025	5807.765503	1801.022688
4	6137.959189	2867.763149	14846.562677	7609.489898

“jing”, male	1	2	3	4
1	7885.643315	2146.101686	16092.119688	6623.942714
2	3095.659142	2643.882499	11302.135570	1833.958547
3	2430.950568	5846.686884	8099.331239	1722.734234
4	630.455439	6369.997089	7576.020996	1892.156048

“lei”, male	1	2	3	4
1	615.195795	4283.935705	4286.356019	2385.863479
2	745.801744	4074.002682	4496.289071	2595.796491
3	25022.684798	29842.489103	21272.197360	23172.689897
4	5858.715392	1323.466258	9609.202697	7708.710133

“yu”, male	1	2	3	4
1	4110.886075	5808.456540	6032.317691	5105.447413
2	8338.190445	6640.620045	18481.394160	7343.629088
3	3081.209121	4778.779522	7061.994635	4075.770404
4	10845.236242	12542.806652	2368.244171	11839.797646

In all, 6 syllables of each tone from the female speaker (“bao,” “chen,” “er,” “hai,” “lei,” and “ma”), and 5 syllables of each tone from the male speaker (“er,” “hai,” “jing,” “lei,” and “yu”), for a total of 24 female syllables and 20 male syllables, were analyzed with the tone recognizer.

The recognizer achieved the following success rates:

	Mean Squared Distance Analysis		Best Fit Linear Regression, 4 Cepstral Regions		Best Fit Quadratic Regression, 4 Cepstral Regions	
	Success Rate	False Positive Rate	Success Rate	False Positive Rate	Success Rate	False Positive Rate
Female Tone 1	100%	5%	83.3%	5%	66.6%	10%
Female Tone 2	100%	0%	100%	5%	33.3%	15%
Female Tone 3	66.6%	10%	66.6%	15%	50%	35%
Female Tone 4	66.6%	5%	50%	5%	0%	15%
Female Average	83.3%	5%	75%	7.5%	50%	18.75%
Male Tone 1	100%	25%	100%	18.75%	40%	18.75%
Male Tone 2	80%	6.25%	80%	12.5%	40%	25%
Male Tone 3	0%	0%	0%	0%	10%	18.75%
Male Tone 4	100%	6.25%	100%	6.25%	10%	25%
Male Average	70%	9.375%	70%	9.375%	25%	9.375%
Total Average	76.65%	7.19%	75.25%	8.44%	37.5%	14%

Several trends are worthy of further discussion. The success rate for the female speaker was, in general, higher than the success rate for the male speaker, regardless of analysis method. This can be explained in part by the presence of more data with which to construct templates for the female speaker (5 syllables, versus 4 for the male), reducing noise and statistic error. Unsurprisingly, the easiest tone to recognize is the first (high) tone, due to the lack of variation between pronunciations. Whereas for the second (rising) and fourth (falling) tones there is a degree of freedom in the extent of the rise and fall, in the first tone there is no such freedom: the tone is consistently at a constant pitch. However, the first tone had the highest false positive rate of the male tones, suggesting that the other tones can be articulated in a subtle manner ripe for statistic mis-analysis.

The third (dipping) tone presented something of a problem for the tone recognizer, especially in male speech. Visual and aural inspection of third tone articulation reveals that it is indeed characterized by a variety of executions: the third tone can “dip,” as per its name, but it can also simply fall without rising, in the manner of the fourth tone, though less dramatically. In addition, the “dip” portion of the syllable can be anywhere within the syllable, not simply in the

middle. These variations present difficulties for analysis and recognizance methods based on simple averaging templates. For example, let us imagine that there are two possible methods of articulating a third tone, one that “dips” and one that “falls subtly.” Rather than taking these two different instances as distinct manners of executing the tone, the averaging process will meld the differences, resulting in a template that may be statistically reasonably close to both, but does not accurately represent either. This can be seen in the male and female mean-squared distance analysis results, where the third tone is mis-recognized most often as a first tone, and in the male linear best-fit analysis, where the third tone is recognized as anything but a third tone. Also note that in linear regression analysis, the female fourth tone was mis-recognized twice as a third tone, suggesting a similarity of execution (the falling without the rising) there.

Quadratic regression analysis failed spectacularly for all tones in both male and female speaker-dependent syllables compared to the other two analysis methods, though it did produce slightly better results for recognizance of the male third tone. This suggests that polynomial regression “overfits” the data points and suffers due to the statistic noise inherent in the test and template syllables.

2. Speaker-independent Analysis

For the speaker-independent analysis testing, we used the samples of the numbers zero through ten in Mandarin provided on the Oregon Graduate Institute Multi Language Telephone Speech Corpus. The numbers zero through ten contain instances of all four Mandarin lexical tones:

Syllable	Tone	Meaning
líng	2	Zero
yī	1	One
èr	4	Two
sān	1	Three
sì	4	Four
wǔ	3	Five
liù	4	Six
qī	1	Seven
bā	1	Eight
jǐu	3	Nine
shí	2	Ten

Tone templates were constructed from twenty-five instances of each tone culled from the first twenty-five sample files in the numbers corpus. Five independent sample files were then analyzed using these templates, for a total of fifty-five syllables analyzed. Results tables for speaker-independent analysis follow the same format as those for speaker-dependent analysis above, but with the OGI sample index number in the upper left identification cell and the spoken numeral with English transliteration in parentheses and lexical tone in brackets along the left-hand column.

2.a. Speaker-independent: mean-squared distance analysis

ma061num	1	2	3	4
líng (zero) [2]	2767.089714	894.831129	3159.514792	5680.549098
yī (one) [1]	386.169489	2231.272123	957.060704	2716.937450
èr (two) [4]	1039.681744	2980.945436	877.327788	2007.212485
sān (three) [1]	781.900512	2285.178201	292.957539	2654.352288
sì (four) [4]	3806.695810	5468.649114	3100.191016	1694.245049
wǔ (five) [3]	2437.364846	3838.673404	1738.634027	2327.350739
liù (six) [4]	3414.407900	5358.572536	3268.923046	1962.496117
qī (seven) [1]	191.079767	2081.953925	730.835474	2855.693993
bā (eight) [1]	661.984103	2104.923254	655.724958	2847.118698
jǐu (nine) [3]	1749.807549	3521.248721	1126.624136	1422.693847
shí (ten) [2]	3296.274909	1356.694736	3686.109484	6220.163499

ma063num	1	2	3	4
líng (zero) [2]	1872.903107	394.768061	2245.794394	4773.925752
yī (one) [1]	230.640504	2111.748050	809.724000	2815.421373
èr (two) [4]	2519.675387	4333.345951	1937.091350	1000.857121
sān (three) [1]	464.507629	2254.561432	319.272529	2641.137741
sì (four) [4]	2030.490787	3730.357114	1338.138769	1387.039423
wǔ (five) [3]	747.218711	2504.434963	287.141673	2399.245046
liù (six) [4]	1476.798395	3417.293462	1289.987784	1530.712739
qī (seven) [1]	647.441256	2539.668065	833.829057	2451.282693
bā (eight) [1]	635.436822	2472.584614	683.452123	2440.230680
jǐu (nine) [3]	729.870906	2605.235067	591.643828	2292.112117
shí (ten) [2]	836.327046	1233.976085	1263.488377	3682.882263

ma065num	1	2	3	4
líng (zero) [2]	2611.301912	1776.304438	2761.607984	5227.702066
yī (one) [1]	311.239861	2101.257204	709.284262	2812.359449
èr (two) [4]	5086.214458	7017.401426	4652.690539	2123.470700
sān (three) [1]	656.962583	1936.255009	526.766148	2965.296863
sì (four) [4]	5570.165707	7492.220426	5118.130553	2618.282814
wǔ (five) [3]	5209.158398	4548.602784	4947.296853	7049.836376
liù (six) [4]	8566.455573	10510.620154	8196.425219	5668.091493
qī (seven) [1]	254.021608	2121.491639	759.522223	2819.721962
bā (eight) [1]	351.609586	2061.936322	691.605177	2840.688826
jǐu (nine) [3]	7962.134566	9906.299181	7563.136491	5027.136503
shí (ten) [2]	3113.953462	1888.346771	3616.225241	6038.696813

ma067num	1	2	3	4
líng (zero) [2]	3196.563206	1540.986788	3762.940423	6160.016729
yī (one) [1]	498.846663	1787.158716	1112.732468	3253.340401
èr (two) [4]	7093.648309	8995.210459	6607.500743	4131.480892
sān (three) [1]	1602.857665	1434.626858	1417.294476	3731.951550
sì (four) [4]	9956.273462	11900.384496	9603.269972	7075.805633
wǔ (five) [3]	1498.348113	2335.283049	1131.304479	3016.631840
liù (six) [4]	4132.011843	6034.257877	4177.833975	2773.312238
qī (seven) [1]	157.009082	2056.257418	604.398734	2851.350753
bā (eight) [1]	553.102299	1689.499735	772.958497	3206.391527
jǐu (nine) [3]	5088.413665	6149.524934	4400.864842	4145.691150
shí (ten) [2]	2888.031979	1090.370731	3429.764386	5862.627022

ma073num	1	2	3	4
líng (zero) [2]	3385.921003	2243.406569	3464.458135	5866.188253
yī (one) [1]	328.309799	1630.569938	920.841766	3272.766455
èr (two) [4]	14140.18217	16035.71582	13642.28244	11164.85392
sān (three) [1]	557.996919	1617.286536	794.784723	3290.682618
sì (four) [4]	3557.81709	4071.538704	3088.421666	3593.368761
wǔ (five) [3]	2286.136877	2085.772884	1892.605679	3847.207596
liù (six) [4]	3278.596667	4642.356656	2709.402136	2013.333989
qī (seven) [1]	1627.963517	3234.2571	1322.254238	2260.423174
bā (eight) [1]	2735.785866	2885.233586	3013.394347	4589.550745
jǐu (nine) [3]	1773.811856	3466.675842	1114.46	1683.850524
shí (ten) [2]	1168.625382	1156.372857	1858.799706	4043.43795

2.b. Speaker independent: best-fit linear analysis, 4 cepstral regions

ma061num	1	2	3	4
líng (zero) [2]	1.321805	0.686019	1.310533	2.189179
yī (one) [1]	0.278255	0.921019	0.646672	0.69549
èr (two) [4]	0.332694	1.00591	0.556086	0.6514
sān (three) [1]	0.342581	0.790208	0.253808	0.937852
sì (four) [4]	1.479099	2.152315	1.20143	0.976346
wǔ (five) [3]	1.262116	1.935332	0.984447	1.583524
liù (six) [4]	1.495334	1.981	1.86375	1.199161
qī (seven) [1]	0.156838	0.792624	0.511962	0.832649
bā (eight) [1]	0.577667	1.004669	0.495249	0.99556
jǐu (nine) [3]	0.591775	1.264991	0.408692	0.578044
shí (ten) [2]	0.91417	0.311305	1.191838	1.807898

ma063num	1	2	3	4
líng (zero) [2]	0.477623	0.386052	0.741999	1.358059
yī (one) [1]	0.204771	0.759795	0.573187	0.899806
èr (two) [4]	1.008834	1.68205	0.731166	0.614633
sān (three) [1]	0.279899	0.875555	0.153385	0.976752
sì (four) [4]	0.819439	1.492655	0.541771	0.658691
wǔ (five) [3]	0.224407	0.826152	0.202788	0.815543
liù (six) [4]	0.495498	1.168714	0.539993	0.444966
qī (seven) [1]	0.237058	0.910274	0.553127	0.65667
bā (eight) [1]	0.331023	0.96681	0.513036	0.918518
jǐu (nine) [3]	0.316047	0.951833	0.422467	0.789408
shí (ten) [2]	0.534071	0.607439	0.434329	1.401444

ma065num	1	2	3	4
líng (zero) [2]	1.713079	1.266183	1.406107	2.172663
yī (one) [1]	0.348713	0.608315	0.686513	1.131827
èr (two) [4]	1.689517	2.362734	1.411849	0.795789
sān (three) [1]	0.426542	0.965606	0.344123	1.242264
sì (four) [4]	2.511753	3.184969	2.234084	1.618024
wǔ (five) [3]	2.152082	2.314409	1.843798	1.697378
liù (six) [4]	3.66109	4.296876	3.760832	2.793717
qī (seven) [1]	0.232195	0.786765	0.445939	0.97638
bā (eight) [1]	0.314552	0.60784	0.637224	1.115176
jǐu (nine) [3]	2.767358	3.440575	2.48969	2.201547
shí (ten) [2]	1.833244	1.377618	2.110912	2.726972

ma067num	1	2	3	4
líng (zero) [2]	0.948091	0.668924	1.22576	1.84182
yī (one) [1]	0.332413	0.938197	0.657535	1.032002
èr (two) [4]	2.743609	3.379395	2.843351	1.977458
sān (three) [1]	1.309881	1.283318	1.270271	2.16286
sì (four) [4]	3.654734	4.290521	3.754476	2.787361
wǔ (five) [3]	1.222644	1.286637	0.938916	1.079793
liù (six) [4]	1.92151	2.368406	2.289926	1.625337
qī (seven) [1]	0.126408	0.695503	0.343693	0.881698
bā (eight) [1]	0.28531	0.608174	0.342711	1.139391
jǐu (nine) [3]	0.610171	0.954234	0.510429	1.477544
shí (ten) [2]	1.068403	0.54374	1.346072	1.962132

ma073num	1	2	3	4
líng (zero) [2]	1.900137	1.453241	1.736968	2.704083
yī (one) [1]	0.226254	0.464677	0.486207	1.102267
èr (two) [4]	2.550609	3.223825	2.27294	2.043611
sān (three) [1]	0.451588	0.805969	0.396747	1.062424
sì (four) [4]	3.588141	3.141245	3.219724	3.884313
wǔ (five) [3]	1.535954	1.089058	1.167538	2.034285
liù (six) [4]	2.156103	2.281	2.225229	1.727754
qī (seven) [1]	0.581927	0.744254	0.573005	1.175816
bā (eight) [1]	2.761053	3.159627	2.93277	3.07919
jǐu (nine) [3]	0.563956	1.230104	0.422416	0.768928
shí (ten) [2]	0.638308	0.994093	0.657838	1.336841

2.c. Speaker-independent: quadratic regression analysis, 4 cepstral regions

ma061num	1	2	3	4
líng (zero) [2]	2295.019129	2283.996376	754.180958	3641.905421
yī (one) [1]	1592.631783	1581.609011	770.857448	2939.518055
èr (two) [4]	519.88654	530.909302	2060.72472	826.999743
sān (three) [1]	591.01025	602.033009	2131.848438	755.87603
sì (four) [4]	4529.205352	4518.182618	2988.367198	5876.091603
wǔ (five) [3]	3889.296307	3960.955752	2844.409891	4100.131627
liù (six) [4]	1376.009553	1364.986787	1810.014022	2722.895831
qī (seven) [1]	1818.717423	1829.740191	3359.555621	471.831153
bā (eight) [1]	1024.23448	1035.257242	2565.072666	917.371635
jǐu (nine) [3]	1466.301565	1455.278797	201.061254	2813.187845
shí (ten) [2]	627.369006	616.346242	968.084421	1974.255299

ma063num	1	2	3	4
líng (zero) [2]	145.847964	134.825203	1394.990216	1492.734241
yī (one) [1]	1461.36713	1472.389886	3002.205311	330.833078
èr (two) [4]	1803.352228	1875.011688	758.465804	2947.787688
sān (three) [1]	1863.334281	1852.31152	322.496098	3210.220593
sì (four) [4]	1286.735035	1358.394497	1118.784449	1768.940007
wǔ (five) [3]	2930.102024	2941.124788	4470.94021	1583.21575
liù (six) [4]	3662.770374	3673.793114	5203.608543	2315.884083
qī (seven) [1]	1679.3948	1690.417561	3220.232942	332.508518
bā (eight) [1]	6854.341013	6865.363804	8395.179189	5507.454751
jǐu (nine) [3]	2658.113794	2669.136554	4198.95197	1311.227515
shí (ten) [2]	5878.986015	5867.963233	4338.147836	7225.87226

ma065num	1	2	3	4
líng (zero) [2]	2917.133407	2906.110628	1376.295209	4264.019666
yī (one) [1]	3874.282674	3885.30543	5415.120839	2527.396397
èr (two) [4]	2102.777739	2091.754944	561.939537	3449.663995
sān (three) [1]	1396.712871	1468.372344	1126.387779	1761.33667
sì (four) [4]	12517.25703	12506.23432	10976.41886	13864.14335
wǔ (five) [3]	2611.952618	2600.929859	1071.114443	3958.838906
liù (six) [4]	10790.06631	10801.08915	12330.90455	9443.180099
qī (seven) [1]	491.248133	419.58867	1697.352007	1190.372458
bā (eight) [1]	2508.357821	2519.380584	4049.195988	1161.471539
jǐu (nine) [3]	2329.384804	2401.044258	1284.498391	2540.220137
shí (ten) [2]	10685.5871	10674.56434	9144.748896	12032.47335

ma067num	1	2	3	4
líng (zero) [2]	9224.230653	9213.208021	7683.392478	10571.1169
yī (one) [1]	1180.851333	1169.828573	1683.485693	2527.737629
èr (two) [4]	7719.666829	7730.68959	9260.505067	6372.780555
sān (three) [1]	13501.53713	13490.51438	11960.69892	14848.42343
sì (four) [4]	947.147127	958.16989	2487.985303	399.739153
wǔ (five) [3]	1810.124091	1821.146872	3350.962291	495.24729
liù (six) [4]	915.425608	904.402849	1871.902697	2262.311891
qī (seven) [1]	847.173109	836.150348	973.61917	2194.059383
bā (eight) [1]	2271.36738	2282.390139	3812.205554	924.481096
jǐu (nine) [3]	2802.839383	2813.862141	4343.677528	1455.953084
shí (ten) [2]	956.750931	945.728172	1273.037914	2303.637194

ma073num	1	2	3	4
líng (zero) [2]	7174.939522	7163.916792	5634.101381	8521.825839
yī (one) [1]	84.805656	95.828417	1625.643846	1262.080626
èr (two) [4]	4630.23973	4641.262493	6171.077873	3283.353431
sān (three) [1]	822.089936	878.480473	718.748246	2168.976213
sì (four) [4]	10510.11316	10499.09041	8969.274828	11856.99952
wǔ (five) [3]	7731.587253	7742.610014	9272.425363	6384.700931
liù (six) [4]	4586.072703	4575.049972	3045.234532	5932.959023
qī (seven) [1]	20354.53776	20365.56033	21895.37605	19007.65148
bā (eight) [1]	43444.41334	43455.43605	44985.2514	42097.52685
jǐu (nine) [3]	569.087974	640.747438	991.2918	1896.432659
shí (ten) [2]	601.080736	672.740193	1378.525964	1509.198501

	Mean Squared Distance Analysis		Best Fit Linear Regression, 4 Cepstral Regions		Best Fit Quadratic Regression, 4 Cepstral Regions	
	Success Rate	False Positive Rate	Success Rate	False Positive Rate	Success Rate	False Positive Rate
Tone 1	70%	2.9%	65%	8.6%	10%	8.5%
Tone 2	90%	2.2%	80%	4.4%	40%	11.1%
Tone 3	70%	22.2%	60%	20%	40%	37.7%
Tone 4	73.3%	5%	80%	5%	33.3%	37.5%
Total Average	75.8%	8%	71.25%	9.5%	30.8%	23.7%

Speaker-independent analysis succeeded slightly less frequently, on average, than speaker-dependent analysis, but achieved more consistent success rates across tones, and more robust success in recognizing the third tone in particular. This suggests that speaker-independent (“untrained”) lexical tone recognition in Mandarin can achieve a degree of success comparable with that of speaker-dependent systems, given suitable modeling techniques. The tone templates for speaker-independent analysis were constructed from a considerably broader corpus of speech

samples than those for speaker-dependent analysis, helping to eliminate statistic noise in the templates. However, quadratic regression still performed poorly, as an analogous reduction in statistic noise was not present in the test syllables.

Inspection of third tone utterances reveals even more variety than was present in the speaker-dependent corpus: some instances of the third tone do not have a falling section at all. Rather, the tone rises slightly from a kind of implied dip at the beginning of the syllable. This compounds the problem inherent in using straight averaging templates to characterize the third tone in arbitrary speaker-independent Mandarin.

IV. Conclusion

Unisyllabic lexical tone recognition in Mandarin Chinese lends itself well to statistical analysis based on fundamental frequency track. For comparison, we can posit two baseline methods of analysis against which to view the results of the method used in this study. The simplest is no attempt at tone recognition at all, or, equivalently (for systems attempting transcription), random tone selection. This will achieve, on average, a 25% success rate, since there are four lexical tones from which to choose. The mean-squared distance analysis and linear regression analysis described in this paper thus outperform a random methodology by quite a wide margin, on average, in both speaker-dependent and speaker-independent circumstances. Quadratic regression analysis tends to barely outperform a random methodology.

As a slightly less silly baseline, we can imagine a system which characterizes the tones based solely on the tone descriptions, or the ideal execution shape of the tones. Such a system might first perform a linear regression analysis on a sample tone divided into two cepstral regions. If the slope of the first region is below some negative threshold, and the slope of the second region is above an analogous positive threshold, the system will recognize the overall fit

as characteristic of the third (dipping) tone. Otherwise, the system will reanalyze the syllable in entirety, as one region, with a linear regression. If the resulting slope is between the positive and negative thresholds, it will characterize the sample as a first (high) tone. If the best-fit line has a slope greater than the positive threshold, the system will characterize the sample as a second (rising) tone, and if it is below the negative threshold, the system will see a fourth (falling) tone. Implementing such a system (in `tone.c`, accessible from the command line as the “-1” regression), we analyze all of the test tones used previously, with a positive slope threshold of 0.2 and a negative slope threshold of -0.2, and see the following results (speaker-independent, since no templates are used):

Baseline Shape Analysis		
	Success Rate	False Positive Rate
Tone 1	93.5%	25.0%
Tone 2	47.6%	3.8%
Tone 3	23.8%	9.0%
Tone 4	73.0%	12.3%
Total Average	59.5%	12.525%

Thus our baseline shape analysis technique achieves better results on average than a purely random method, as might be expected, but falls short of the consistent level of accuracy possible with more sophisticated statistic analysis based on templates. Note that the high success rate and corresponding high false positive rate of the first tone in our baseline shape analysis may be partially caused by non-ideal slope thresholds — the algorithm is catching non-first tone syllables within the first tone slope region. Adjusting these parameters to narrow the region in which a syllable is recognized as first tone causes fewer false recognitions of the first tone, but results in a correspondingly lower success rate. Note also the perennial difficulty in properly recognizing the third tone.

Many current implementations of English voice recognition software, such as Dragon Naturally Speaking from Dragon Systems or ViaVoice from IBM, rely on person-specific “training” of the software by spoken reading of known textual material. Applied to Mandarin, this technique can also be used to form the pitch track templates for each tone on an individual basis, or by dialect, as demonstrated in this study. In particular, we can note that the two individuals participating in the speaker-dependent analysis speak with slightly different intonations on occasion, most noticeably in the third tone. Whereas Zhongjue Chen, from Shanghai, produces a significant rise near the end of the third tone, Tsiyun Cherry Fu, speaking with an intonation more reminiscent of Beijing, lacks a such a rise at the end of the third tone. With person-specific templates, however, these slight variations in tonal execution are preserved for the analysis of test tones from the same speaker.

Also by virtue of templates, the methods used in this study for tone recognition in Mandarin are easily applicable to other languages which feature lexical tone, such as Yue Chinese (Cantonese), which has seven tones and 66 million speakers worldwide, and Thai, with five tones and 20 million speakers (Grimes).

Supplementary techniques may improve the accuracy of a template-based statistic analysis system similar to the one in this study. “Dynamic Time Warping” (Fu, 5) addresses the possibility of a speaker spending more time in a particular part of a syllable than another. For example, in articulating the third tone, a speaker might spend an inconsistent amount of time in the rising portion, or the falling portion, in different instances of the tone. The system presented in this paper normalizes all syllables to a standard time length, but does no further time-based processing or analysis. Thus, such inconsistencies in tone articulation, while trivial to human speakers, would not be well handled by our analysis methods. Similarly, differences in duration

of the syllables themselves might provide additional clues regarding lexical tone. The third tone, for instance, is generally considerably longer in duration than the fourth tone, since the third tone syllable must host a relatively complex contour compared to that of the fourth tone syllable. Once again, our system ignores these duration-based distinctions. Dynamic time warping techniques, however, deal with them by employing dynamic programming to non-linearly align and fit sections of test syllables with corresponding sections of template syllables.

Other tone recognition systems (Fu, 9) vary in method and accuracy. A vector quantization system employing codebook matching was implemented for continuous speech by Heng-Jie Ma, and achieved 82% accuracy. Hidden Markov Model (HMM)-based systems are popular for recognition of lexical tone, and include a technique by Xi-Xian Chen et al. involving estimation of pitch periods by three-level center clipping autocorrelation, and corresponding creation of an HMM for each tone, achieving an experimental recognition accuracy of 96% for isolated syllables. Wu-Ji Yang et al. combined HMM and vector quantization frameworks to create a system which achieved 93.75% accuracy for isolated syllables. Other techniques include neural networks, Multi-layer Perceptron (MLP), and an application of Fuzzy Sets theory.

Accurate lexical tone recognition in Mandarin Chinese is an indispensable part of the development of a comprehensive voice recognition system, and success in this area will greatly help to improve and simplify the computer interface experience for Chinese speakers who wish to use traditional characters for communication. Such a voice recognition system will eliminate the need for the slow, cumbersome, and difficult to learn keyboard-based Mandarin character entry paradigms which plague current users, by smoothly transcribing speech into traditional ideographs.

V. Acknowledgements

The Author would like to thank the following people, who have been instrumental in the writing of this Thesis: Professor Michael S. Brandstein, for help with the processing techniques, and for giving much needed direction; Dr. Thomas J. Brennan, for help with the mathematics of the analysis, specifically the polynomial regression; Professor Stuart Shieber, for conceptual help and tips on analysis methodology; and Professor Lisa Lavoie for extensive revision suggestions and help with the phonetics and phonology. The responsibility for any factual or judgmental errors that remain, however, is naturally the Author's alone.

VI. Bibliography

- _____. "Chinese Speech Recognition Research."
<<http://www.research.ibm.com/beijing/projects/speech/sr.html>>. IBM Research.
- Fu, Stephen W.K., et. al. "A Survey on Chinese Speech Recognition."
<<http://citeseer.nj.nec.com/fu96survey.html>>. 1996.
- Grimes, Barbara F. (Editor). *Ethnologue: Languages of the World, 14th Edition*.
<<http://www.sil.org/ethnologue/>>. 2000.
- Katamba, Francis. *An Introduction to Phonology*.
London: Logman Group UK Limited, 1989.
- Press, William, et. al. *Numerical Recipes in C: The Art of Scientific Computing, Second Edition*.
Cambridge: Cambridge University Press, 1992.
- Rabiner, Lawrence R. / Schafer, Ronald W. *Digital Processing of Speech Signals*.
Englewood Cliffs, N.J.: Prentice-Hall, 1978.
- Rice, John A. *Mathematical Statistics and Data Analysis*.
Belmont, California: Duxberry Press, 1995.
- Yeh Teh-ming. "The Variation of Tone Sandhi in Mandarin Chinese."
<<http://udv239-3.ruk.cuni.cz/conferences/tone/sylaby/Yeh.htm>>. 1999.