

1 **A SYSTEM AND METHOD FOR DETERMINING THE TONE OF A SYLLABLE**
2 **OF MANDARIN CHINESE SPEECH**

3
4 **CROSS-REFERENCE TO RELATED APPLICATIONS**

5 The present invention also relates to pending U.S. Patent Application Serial No.:
6 , filed _____, 1994, invented by Hsiao-Wuen Hon and Bao-Sheng Yuan, entitled "A
7 System And Method For Recognizing A Tonal Language," which is incorporated
8 herein by reference. The present invention also relates to pending U.S. Patent
9 Application Serial No.: _____, filed _____, 1994, invented by Hsiao-Wuen
10 Hon, Yen-Lu Chow, and Kai-Fu Lee entitled "Continuous Mandarin Chinese Speech
11 Recognition System Having An Integrated Tone Classifier," which is incorporated
12 herein by reference.

13

14 **BACKGROUND OF THE INVENTION**

15 1. **Field of the Invention.**

16 The present invention relates generally to speech recognition systems. In
17 particular, the present invention relates to a system and method for determining the
18 tone of a syllable of speech. Still more particularly, the present invention relates to a
19 system and method for identifying the tone of a syllable of Mandarin Chinese speech.

20 2. **Description of the Background Art.**

21 Mandarin Chinese is a tonal syllabic language where a syllable and tone
22 combine to define the meaning of an utterance. In Mandarin Chinese, there are 4
23 lexical tones (high and level, rising, falling-rising, and falling) and 1 neutral tone.
24 Exemplary wave forms for electrical signals representing the tones of Mandarin
25 Chinese are shown in Figures 1A, 1B, 1C, 1D, and 1E. The high and level tone, the
26 rising tone, the falling-rising tone, the falling tone, and the neutral tone will each be
27 referred to generically as a language tone in this application. The tone is the behavior
28 of the fundamental frequency (pitch) of the audio signal. The tone of a syllable affects
29 the meaning of the syllable. Syllables with the same phonetic structure but different

1 tones usually have significantly different meanings and correspondence to different
2 written characters. Thus, to recognize accurately an audio signal of Mandarin Chinese,
3 a speech recognition system must recognize both the syllable and the tone of the
4 syllable.

5 Once a speech recognition system has identified a syllable, the speech
6 recognition system must identify the tone of the syllable. Frequency analysis of the
7 tone of the syllable is often time consuming and cumbersome. This is particularly
8 problematic for speech recognitions systems performing real-time speech recognition.
9 These systems generally must quickly and accurately identify the tone of a syllable in
10 real time. Furthermore, there is a growing trend to incorporate speech recognition
11 systems into small, low cost consumer goods. The speech recognition system
12 incorporated into such goods must be sufficiently inexpensive that it does not overly
13 inflate the cost of the device.

14 Many systems and methods for determining the tone of a syllable of Mandarin
15 Chinese speech have been used in the prior art. These prior art systems and methods,
16 such as vector quantization, hidden Markov models, or neural networks, are very
17 computationally intensive and therefore, maybe unsuitable for real time recognition of
18 Mandarin Chinese speech. These prior art systems, which rely on short-term
19 frequency analysis, also do not provide good frequency resolution and provide less
20 than acceptable accuracy.

21 There is a need for a system and method for quickly, accurately, and
22 inexpensively identifying the tone of a syllable of Mandarin Chinese speech.

23

24 SUMMARY OF THE INVENTION

25 The present invention overcomes the deficiencies and limitations of the prior art
26 with a system and method for analyzing the tone of a syllable of Mandarin Chinese
27 speech. The novel tone recognition system of the present inventor comprises a
28 coefficient determinator, a coefficient modeler, and a comparator. The coefficient
29 determinator receives a digitized signal of a syllable from the A/D converter. The

1 coefficient determinator estimates the pitch of the syllable and describes the pitch by
2 using a least squares method to fit a second order polynomial to the pitch. The
3 coefficient determinator can quickly and easily describe the pitch using a second order
4 polynomial. The second order polynomial is preferred because of its low
5 computational requirements. The output of the coefficient determinator is coupled to
6 an input of the comparator. Through this coupling, the coefficient determinator
7 provides a signal representing the polynomial to the comparator. A second input of
8 the comparator is coupled to an output of the coefficient modeler. Through this
9 coupling the comparator receives models of the language tones of the tonal language.
10 A model of a language tone comprises expected coefficients for a second order
11 polynomial that describes the language tone, an expected duration of syllables having
12 the language tone, and a co-variance of the coefficients and duration. The comparator
13 compares the polynomial, which describes the syllable to the models, determines the
14 model that most closely matches the syllable, and generates a signal indicating the
15 matching language tone.

16 The present invention also includes a system and method for training a
17 coefficient modeler. The system for training comprises a model generator, a coefficient
18 A training memory, a coefficient B training memory, a coefficient C training memory,
19 and a duration T training memory. For a model, the model generator receives a large
20 number of uttered syllables all known to have the same language tone. The model
21 generator estimates the pitch of each syllable, fits a second order polynomial to each
22 syllable, and stores the resulting coefficients and duration in the appropriate memories.
23 The model generator then determines the arithmetic mean and co-variance for the
24 coefficients and duration and generates a signal for the coefficient modeler to accept
25 the model.

26

27

BRIEF DESCRIPTION OF THE DRAWINGS

Figures 1A, 1B, 1C, 1D and 1E are graphical representations of electrical signals having the behavior of the 4 lexical tones and one neutral tone of Mandarin Chinese speech;

Figure 2 is a block diagram of an exemplary embodiment of a system for training a coefficient modeler;

Figure 3 is a block diagram of a first embodiment of a speech recognition system that includes a tone recognition system according to the present invention;

Figure 4 is a block diagram of a second embodiment of a system for recognizing syllables of Mandarin Chinese speech that includes the tone recognition system of the present invention;

Figure 5 is a block diagram of the speech recognition system of the second embodiment of a system for recognizing syllables of Mandarin Chinese speech;

Figures 6A and 6B are flowcharts showing the preferred method for training the tone recognition system of the present invention to identify the tone of a syllable of Mandarin Chinese speech; and

Figure 7 is a flowchart showing the preferred method for identifying the tone of a syllable of Mandarin Chinese speech using the speech recognition system of the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

Referring now to Figures 1A, 1B, 1C, 1D, and 1E, graphical representations of electrical signals that represent the 4 lexical tones and 1 neutral tone of Mandarin Chinese speech are shown. The signals shown in Figures 1A, 1B, 1C, 1D, and 1E originate as audio signals generated by a human speaker. A speech recognition system receives the audio signals through a microphone. The microphone converts an audio signal to an electrical signal. Each syllable of Mandarin Chinese speech has a tone. The tone is the behavior of the pitch contour, the fundamental frequency contour, of the audio signal of the syllable. The tone is preserved when the audio signal is converted into an electrical signal. Thus, the tone of a syllable may refer to the behavior of the

1 fundamental frequency of the audio signal or the electrical signal into which the audio
2 signal is converted.

3 Figures 1A, 1B, 1C, 1D, and 1E show the tones of Mandarin Chinese speech as
4 fundamental frequencies plotted as a function of time. Figure 1A shows a first tone
5 signal 40 having a high and level tone. Figure 1B shows a second tone signal 42 having
6 a rising tone. A third tone signal 44, shown in Figure 1C, has a falling-rising behavior.
7 Figure 1D shows a fourth tone signal 46 having a behavior of the last lexical tone of
8 Mandarin Chinese speech, the falling tone. Finally, Figure 1E shows a fifth tone signal
9 having the neutral tone.

10 Referring now to Figure 2, a block diagram of a first embodiment of a system 10
11 for recognizing a syllable of speech of a tonal language is shown. The system 10
12 comprises a microphone 76, and analog-to-digital converter 78, a segmenter 79, a
13 syllable recognition system 80, and a tone recognition system 82.

14 The microphone 76 is a conventional microphone for converting analog audio
15 signals into analog electrical signals. The microphone 76 has an input for receiving
16 audio signals and an output for outputting electrical signals. The output is coupled to
17 a line 88. The analog-to-digital converter ("A/D converter") 78 is a conventional
18 analog-to-digital converter for
19 converting analog, electrical signals to digital, electrical signals. The A/D converter 78
20 has an input, which is coupled to line 88, for receiving an analog, electrical signal and
21 an output for outputting a digital, electrical signal. The output of the A/D converter
22 78 is coupled to a line 90. The segmenter 79 is a conventional segmenter for parsing a
23 continuous input into single syllables of Mandarin Chinese speech. The segmenter 79
24 generates output signals of single syllables. The segmenter 79 has an input coupled to
25 line 90 and an output coupled to a line 91.

26 The syllable recognition system 80 is a system for determining the phonetic
27 structure of a syllable of a tonal language and for combining the identified phonetic
28 structure with an identification of the language tone of the syllable to recognize the
29 syllable. The syllable recognition system 80 has a first input, coupled to line 91, for

1 receiving a syllable from the segmenter 79, a second input, coupled to a line 96, for
2 receiving a signal that identifies the language tone of the syllable from the tone
3 recognition system 82, and an output, coupled to a line 98, for outputting a signal
4 representing the recognized syllable. The syllable recognition system 80 uses a hidden
5 Markov model and state analysis to identify the phonetic structure of the syllable. The
6 syllable recognition system 80 combines the identification of the phonetic structure
7 with an identification of the language tone of the syllable to generate a signal indicative
8 of the complete meaning of the syllable.

9 The syllable recognition system 80 is preferably an Apple PlainTalk Chinese
10 Syllable Recognition System (the "PlainTalk System") from Apple Computer, Inc. of
11 Cupertino, California modified to utilize long-term tonal analysis. The PlainTalk
12 System has a first system for determining the phonetic structure of a syllable, a second
13 system for performing short-term tonal analysis on the syllable to identify the language
14 tone of the syllable, and a third system for combining the output of the first system and
15 the second system into a recognition. The output of the third system forms an output
16 of the PlainTalk System. The PlainTalk System may be modified to utilize long-term
17 tonal analysis by removing the second system and coupling line 96 to an input to the
18 third system so that the third system combines the output of the first system, for
19 determining the phonetic structure of the syllable, and the output of the tone
20 recognition system 82 into an output of the syllable recognition system 80. An
21 alternative modification to the PlainTalk System is to modify the third system to
22 receive inputs from the first system, the second system, and the tone recognition
23 system 82 and to combine the three inputs into a recognition which is output on line 98.
24 The syllable recognition system 80 may alternately be a syllable based voice typewriter
25 from Star Incorporated of the People's Republic of China similarly modified as
26 described in connection with the preferred PlainTalk System.

27 As shown in Figure 2, the tone recognition system 82 comprises a coefficient
28 determinator 84, a comparator 86, and a coefficient modeler 38. The coefficient
29 determinator 84 has an input coupled to line 91 to receive a syllable from the

1 segmenter 79 and an output coupled to a line 92. The coefficient modeler 38 is
2 constructed by the trainer 34, shown in Figure 3, and has a first input and an output,
3 both of which are coupled to line 94. The comparator 86 has a first input, coupled to
4 line 92 to receive the output of the coefficient determinator 84, a second input, coupled
5 to line 94 to receive models from the coefficient modeler 38, a first output, and a second
6 output, coupled to line 94, through which the comparator 86 signals for the coefficient
7 modeler 38 to transmit the models. The output of the tone recognition system 82 is
8 formed by the first output of the comparator 86 and is coupled to line 96 so that the
9 identification of the language tone is transmitted to the second input of the syllable
10 recognition system 80.

11 The coefficient determinator 84 receives a syllable from the segmenter 79 on line
12 91. The coefficient determinator 84 estimates the fundamental frequency, the pitch, of
13 the syllable using fourier analysis. Alternatively, the coefficient determinator 84 may
14 use a low pass filter or other means to determine the pitch of the input syllable. The
15 coefficient determinator 84 adjusts the coefficients of a second order polynomial to
16 match closely the pitch contour. The coefficient determinator 84 uses a least squares
17 curve fitting method to adjust the coefficients. Once the coefficients best describe the
18 pitch contour, the coefficient determinator 84 generates a signal to transfer the
19 coefficients along with the duration of the input signal, grouped together in a 4
20 dimensional vector, to the comparator 86. By using a second order polynomial, the
21 coefficient determinator 84 can quickly and accurately describe the tone of the syllable
22 and thus, facilitate a quick and accurate identification of the language tone of the
23 syllable.

24 The coefficient modeler 38 has a first input and a second input that is coupled to
25 a line 7, shown in Figure 3, and an output. The first input and the output are coupled
26 to a line 94. The coefficient modeler 38 comprises a tone 1 model memory 58, a tone 2
27 model memory 60, a tone 3 model memory 62, a tone 4 model memory 64, and a tone 5
28 model memory 66. The memories 58, 60, 62, 64, 66 are shown in Figure 3. Each of the
29 tone model memories 58, 60, 62, 64, 66 stores a model of a tone of Mandarin Chinese

speech. For example, the tone 1 model memory 58 may store a model of the high and level tone, the tone 2 model memory 60 may store a model of the rising tone, the tone 3 model memory 62 may store a model of the falling-rising tone, the tone 4 model memory 64 may store a model of the falling tone, and the tone 5 model memory 66 may store a model of the neutral tone. Each model comprises an A coefficient, a B coefficient and a C coefficient of a polynomial of the form:

$$f(t)=At^2+Bt+C$$

Where t is time. Each model also includes a time duration T. The A coefficient, B coefficient, C coefficient, and duration T are the expected coefficients and duration for the language tone. The A, B, and C coefficients and duration T are grouped together into a 4 dimensional vector. The language tones are preferably modeled using a second order polynomial because of the simplicity and low computational overhead of the second order polynomial. As can be seen from Figures 1A, 1B, 1C, 1D, and 1E, a second order polynomial is adequate to model accurately the pitch contour of each language tone. A model includes an expected duration because the expected coefficients alone do not provide sufficient information to determine accurately the language tone of a syllable. The coefficients with the duration are sufficient to determine accurately the language tone of a syllable. Thus, the present invention overcomes the limitations of the prior art by using a second order equation to model quickly and accurately to determine the language tone of an utterance of Mandarin Chinese speech. While the present invention has been described with respect to five tones, those skilled in the art will appreciate that the number of training data memories 22, 24, 26, 28, 30, shown in Figure 3, and the number of tone models memories 58, 60, 62, 64, 66 may be increased or decreased according to the number of tones in the tonal language being evaluated.

Each model also includes a co-variance matrix for the A coefficient, the B coefficient, the C coefficient and the duration T. The co-variance matrix is a 4x4 matrix

1 which describes the dependence of each of the A, B, and C coefficients and the duration
 2 T on each other. Alternately, the model may include a data structure for storing the
 3 variance of the A, B, and C coefficients and duration T. In this alternate embodiment, it
 4 is assumed that the A, B, and C coefficients and duration T are independent. The
 5 variance is, however, a diagonal co-variance matrix. That is, the diagonal entries of the
 6 co-variance matrix describe the variances and all other entries, the entries off the
 7 diagonal, are 0. Thus, the variance is a sub-set of the co-variance and is therefore,
 8 included within the co-variance. Those skilled in the art will recognize the variance
 9 and its relation to the co-variance. If a higher order polynomial is used to describe the
 10 language tones of a tonal language other than Mandarin Chinese, the co-variance
 11 matrix is enlarged to accommodate the additional coefficients. For example, if a fifth
 12 order polynomial were used, the co-variance matrix would be a 6x6 matrix. The
 13 present invention can also model the tone of utterances of tonal languages other than
 14 Mandarin Chinese. A higher order polynomial may be needed to model accurately the
 15 language tones of a tonal language having language tones that are more complex than
 16 the language tones of Mandarin Chinese.

17 The comparator 86 receives the coefficients and duration, which describe the
 18 pitch of the input syllable, from the coefficient determinator 84 and receives the models
 19 of the language tones from the coefficient modeler 38. The comparator 86 compares
 20 the coefficients and duration to the models to determine the model that the coefficients
 21 and duration most closely resemble. The comparator 86 uses a multivariate normal
 22 density, also known as a Gaussian classifier, to compare the input to the models. The
 23 multivariate normal density is given by the equation:

24

$$25 \quad p(x) = (2\pi)^{-d/2} | \Sigma |^{-1/2} \exp\{ -1/2 (x-\mu)^t \Sigma^{-1} (x-\mu) \}.$$

26

27 Where $p(x)$ is the multivariate normal density; x is an input vector of the A, B, and C
 28 coefficients and duration T of the second order polynomial that the coefficient
 29 determinator 84 determined for the input signal; d is the dimension of the input vector,

1 in this case 4; $|_ |$ is the determinant of the co-variance matrix of the model; $^{-1}$ is the
2 inverse of the co-variance matrix of the model; μ is a vector of the arithmetic means of
3 the A, B, and C coefficients and duration T of the model; and $(x-\mu)^t$ is the transpose of
4 the difference vector of x and μ . The comparator 86 determines the multivariate
5 normal density using the mean vector, μ , of each model. The comparator 86
6 temporarily stores the multivariate normal density determined for each model. The
7 comparator 86 recognizes the model having the greatest multivariate normal density as
8 the model of the language tone of the syllable. The comparator 86 generates a signal
9 indicating this language tone and impresses the signal on line 96 to transmit it to the
10 second input of the syllable recognition system 80.

11 Referring now to Figure 3, a block diagram of a trainer 34 is shown. Figure 3
12 also shows training data memory 32, and the coefficient modeler 38. The trainer 34
13 uses data stored in the training data memory 32 to train the coefficient modeler 38. The
14 training data memory 38 is a conventional random access memory or read only
15 memory for storing data.

16 The training data memory 32 comprises a tone 1 training data memory 22, a
17 tone 2 training data memory 24, a tone 3 training data memory 26, a tone 4 training
18 data memory 28, and a tone 5 training data memory 30. Each of the tone 1-5 training
19 data memories 22, 24, 26, 28, 30 is associated with a language tone of Mandarin
20 Chinese and stores utterances that have the associated tone. For example, the tone 1
21 training data memory 22 may be associated with the high and level tone, the tone 2
22 training data memory 24 with the rising tone, the tone 3 training data memory 26 with
23 the falling-rising tone, the tone 4 training data memory 28 with the falling tone, and the
24 tone 5 training data memory 30 with the neutral tone. The tone 1 training data
25 memory 22 stores at least 20 utterances of Mandarin Chinese speech having the
26 language tone that is associated with the tone 1 training data memory 22. Each
27 utterance is a digitized, single syllable. Similarly, the tone 2 - 5 training data memories
28 24, 26, 28, 30 store at least 20 digitized, single syllables having the language tone with
29 which each is associated. While 20 is the minimum number of utterances, preferably

1 each tone 1 - 5 training memories 22, 24, 26, 28, 30 stores several hundred utterances.
2 The tone 1 - 5 training data memory 22, 24, 26, 28, 30 are each coupled to line 68 and
3 impress signals of the digitized utterances on line 68 to output the digitized utterances
4 from the training data memory 32.

5 The trainer 34 has a first input that is coupled to line 68 and a first output that is
6 formed by line 74. The trainer 34 comprises a model generator 36, an A coefficient
7 memory 50, a B coefficient memory 52, a C coefficient memory 54, and a T duration
8 memory 56. The model generator 36 first estimates the fundamental frequency, the
9 pitch, of a syllable using fourier analysis. Alternately, the model generator 36 may use
10 a low pass filter or other means to determine the pitch of a syllable. The model
11 generator 36 adjusts the coefficients of a second order polynomial so that the
12 polynomial best describes the pitch contour of an utterance. The model generator 36
13 has a first input, a second input, a first output, and a second output. The first input is
14 coupled to line 68 to receive data from the tone 1-5 memories 22, 24, 26, 28, 30. The
15 first output is coupled to a line 70 which is also coupled to an input of each of
16 memories 50, 52, 54, 56. The second input is coupled to a line 72 which is also coupled
17 to an output of each of memories 50, 52, 54, 56. The second output is coupled to line 74
18 which forms an output of the trainer 34. The model generator 36 uses a least squares
19 curve fitting method for determining the coefficients of a polynomial that accurately
20 describes a tone. For example, the model generator 36 receives a digitized utterance
21 from the tone 1 model data 22. The model generator 36 uses fourier analysis, a low
22 pass filter, or other techniques to determine the behavior of the pitch of the utterance.
23 The model generator 36 then uses the least squares curve fitting method to adjust the
24 coefficients of a second order polynomial so that the polynomial best describes the
25 pitch contour of the utterance.

26 The A coefficient memory 50, the B coefficient memory 52, the C coefficient
27 memory 54, and duration T memory 56 are all memories for storing data. Each of the
28 memories 50, 52, 54, 56 have an input coupled to the first output of the model
29 generator 36 by a line 70, and an output coupled to the second input of the model

1 generator 34 by line 72. After the model generator 36 has determined the coefficients
2 of the polynomial that best describes a tone and has determined the duration of the
3 utterance, the model generator 36 generates a signal and impresses the signal on line 70
4 to store the value of the A coefficient in the A coefficient memory 50, the value of the B
5 coefficient in the B coefficient memory 52, the value of the C coefficient in the C
6 coefficient memory 54, and duration of the utterance in the duration T memory 56.

7 The model generator 34 preferably determines the coefficients of the
8 polynomials for utterances in the training data 32, which have the same language tone,
9 together. Once the model generator 34 has determined the coefficients and durations
10 for all utterances that have the same language tone, the model generator 36 generates
11 the model for that language tone.

12 In alternate embodiment, the training data need not be segregated according to
13 language tone. Each utterance would have associated with it an indicator of its
14 language tone. The trainer would have an A coefficient training memory, a B
15 coefficient training memory, a C coefficient training memory, and a duration T training
16 memory for each language tone of the language. In the case of Mandarin Chinese,
17 there would be 5 A coefficient memories, 5 B coefficient B memories, 5 C coefficient
18 memories, and 5 duration T memories. In this alternate embodiment, the model
19 generator would record the coefficients and duration in the set of memories indicated
20 by the tone indicator.

21 The coefficient modeler 38 comprises a tone 1 model memory 58, a tone 2 model
22 memory 60, a tone 3 model memory 62, a tone 4 model memory 64, and a tone 5 model
23 memory 66. The coefficient modeler 38 has an input formed by line 74, and each model
24 memory 58, 60, 62, 64, 66 has an input that is coupled to line 74. Furthermore, each
25 model memory 58, 60, 62, 64, 66 has an output coupled to line 94; line 94 forms an
26 output of the coefficient modeler 38. Each model memory 58, 60, 62, 64, 66 stores a
27 model of a language tone; that is, there is a model memory 58, 60, 62, 64, 66 for each
28 language tone. If the present invention is being used to recognize a syllable of a tonal

1 language other than Mandarin Chinese, there are model memories for each language
2 tone of the tonal language.

3 Once the model generator 36 has determined the coefficients and duration for
4 each utterance in one of the memories 22, 24, 26, 28, 30, the model generator 36
5 determines the arithmetic mean and co-variance matrix of the coefficients and
6 durations stored in the memories 50, 52, 54, 56. The average coefficient, average
7 duration and co-variance matrix of the coefficients and duration constitute a model of a
8 language tone. The model generator 36 determines the model and generates a signal
9 on line 74 to transfer the model to the coefficient modeler 38 where the model is stored
10 in the appropriate model memory 58, 60, 62, 64, 66. The tone model memories 58, 60,
11 62, 64, 66 each have an output, which is coupled to line 94, for outputting models of the
12 tones of Mandarin Chinese speech.

13 In the preferred embodiment, the model generator 34 receives all the utterances
14 stored in the tone 1 training data memory 22. The model generator 34 fits a second
15 order polynomial to an utterance and generates signals on line 70 to store the A, B, and
16 C coefficients and duration T in memories 50, 52, 54, 56. The model generator 34
17 repeats this process for each utterance stored in the tone 1 training data memory 22.
18 The model generator 34 then receives the A, B, and C coefficients and duration T from
19 the memories 50, 52, 54, 56 on line 72 at the second input. The model generator 34
20 determines the arithmetic means and co-variance matrix to generate a model of the
21 first language tone and then generates a signal to transfer the model to the tone 1
22 model memory 58. After transferring the model, the model generator 34 asserts a
23 signal on line 70 to the memories 50, 52, 54, 56 to reset the memories 50, 52, 54, 56, and
24 the model generator 34 performs the described functions on the data stored in the tone
25 2 training data memory 24.

26 Referring now to Figure 4, a block diagram of a second embodiment of the
27 system 10 for analyzing the tone of a syllable of Mandarin Chinese speech, constructed
28 in accordance with the present invention, is shown. The system 10 preferably
29 comprises an input device 12, an output device 14, a processor 16, a memory means 18,

1 and a speech recognition system 20. The input device 12, output device 14, processor
2 16, memory means 18, and speech recognition system 20 are coupled in a von Neuman
3 architecture via a bus 22 such as in a personal computer. The processor 16 is preferably
4 a microprocessor such as a Motorola 68040; the output device 14 is preferably a video
5 monitor; and the input device 12 is preferably a keyboard and mouse type controller
6 and a microphone for inputting audio signals. The input device 12 includes an A/D
7 converter for digitizing analog signals from the microphone. In an exemplary
8 embodiment, the system 10 is a Macintosh Quadra 840AV computer system from
9 Apple Computer, Inc. of Cupertino, California. Those skilled in the art will realize that
10 the system 10 could also be implement on an IBM personal computer or any other
11 computer system.

12 The memory means 18 is constructed with random access memory and read
13 only memory. The memory means 18 stores data and program instruction steps for the
14 system 10. The memory means 18 may be a conventional dynamic random access
15 memory and a conventional disk drive. The memory means 18 includes a training data
16 memory 32. The training data memory 32 is coupled through the bus 22 to transmit
17 and receive signals from the speech recognition system 20.

18 The speech recognition system 20 is coupled to the bus 22 to receive a digitized
19 input signal from the input device 12. The speech recognition system 20 parses the
20 input signal to obtain a syllable of Mandarin Chinese speech and generates an output
21 of the recognized syllable. The speech recognition system 20 determines both the
22 phonetic structure and language tone of the syllable.

23 Referring now to Figure 5, a block diagram of the speech recognition system 20
24 is shown. The speech recognition system 20 comprises a segmenter 79, a syllable
25 recognition system 80, a tone recognition system 82, and a trainer 34. The recognition
26 system 82 comprises a coefficient modeler 38, a coefficient determinator 84, and a
27 comparator 86. In this second embodiment, the segmenter 79, syllable recognition
28 system 80, tone recognition system 82, and trainer 34 are memories storing sets of
29 program instruction steps that when executed by the processor 16 perform the

1 functions described above with reference to Figure 2 and Figure 3. For convenience,
2 the segmenter 79, a syllable recognition system 80, a tone recognition system 82, a
3 trainer 34, coefficient modeler 38, a coefficient determinator 84, and a comparator 86
4 have been given the same numerals in Figure 5 as in Figures 2 and 3. Those skilled in
5 the art will recognize that while described in Figures 2 and 3 as separate devices with
6 their own processors, the present invention may be implemented with each device
7 being a set of program instruction steps sharing a single processor 16.

8 Each of the syllable recognition system 80, tone recognition 82, and trainer 34
9 are coupled to the bus 22 to transmit and receive signals to and from the input device
10 12, the output device 14, the processor 16, and memory means 18 and to transmit and
11 receive signals from each other as shown in Figures 2 and 3 and described above. The
12 trainer 34 is coupled to the bus 22 to receive data from the training data memory 32 in
13 the memory means 18. The trainer 34 is also coupled to the bus 22 to transmit models
14 of the language tones to the coefficient modeler 38. Through the bus 22, the segmenter
15 79 receives a digital input signal from the input device 12 and transmits a syllable to
16 the syllable recognition system 80 and to the coefficient determinator 84. As has been
17 stated, the syllable recognition system 80 receives a syllable from the segmenter 79
18 through the bus 22. The syllable recognition system 80 also receives a signal indicating
19 the language tone of the syllable from the comparator 86 through the bus 22. The
20 syllable recognition 80 signals the recognition of the syllable to the output device, or to
21 a look up table or other memory device, through the bus 22. The bus 22 couples the
22 coefficient determinator 84 to the comparator 86 so that the coefficient determinator
23 may transmit the coefficients and duration that describe the pitch contour of a syllable
24 to the comparator 86. Finally, the comparator 86 receives models of the language tones
25 from the coefficient modeler 38 over the bus 22.

26 Referring now to Figures 6A and 6B, a flow chart is shown of a preferred
27 method for training the coefficient modeler 38 for the language tones of a syllable of
28 Mandarin Chinese speech. The method begins in step 700 where the trainer 34 receives
29 an utterance or signal representing a syllable of Mandarin Chinese speech having a

1 known tone. The method preferably processes the utterances having the same
2 language tone together in one group. For example, the model generator 36 processes
3 the utterances in the tone 1 training data 22 together in one group. In step 702, the
4 coefficient determinator 84 uses fourier analysis to estimate the pitch contour of the
5 input signal. The coefficient determinator 84 may alternately use a low pass filter or
6 other techniques to estimate the pitch contour of the input. Next, in step 704, the
7 model generator 36 fits a second order polynomial to the estimated pitch contour. The
8 model generator 36 uses a least square-error curve fitting method for fitting the second
9 order polynomial to the tone. Those skilled in the art will recognize the least square
10 method and will realize that there are alternative methods for fitting a second order
11 polynomial to the pitch. If the model generator 36 is being used to train the coefficient
12 modeler 38 to model a tonal language other than Mandarin Chinese, a polynomial
13 having an order higher than the second order may be required to model accurately the
14 language tones.

15 In step 706, the model generator 36 selects the A coefficient of the polynomial
16 and stores it in the coefficient A training memory 50. Next in step 708, the model
17 generator 36 selects the B coefficient and stores it in the B coefficient training memory
18 52. Similarly, in step 710, the model generator 36 selects the C coefficient and stores it
19 in the C coefficient training memory 54. Finally, in step 712, the model generator 36
20 determines the duration of the utterance and stores it in the duration T training
21 memory 56.

22 In step 714, the model generator 36 determines if there is another utterance to be
23 processed that has the current language tone in the training memory 32. If there is
24 another utterance to be processed, the method returns to step 700 where the model
25 generator 36 receives the next utterance from the training memory 32.

26 If in step 714 there are no more utterances having the current language tone to
27 be processed, the model generator 36 uses the data stored in the coefficient A training
28 memory 50, coefficient B training memory 52, coefficient C training memory, and
29 duration T training memory 56 to generate a model of the language tone. In step 716,

1 the model generator 36 receives the A coefficients stored in the A coefficient training
2 memory 50 and determines the arithmetic mean of the A coefficients. Next in step 718,
3 the model generator 36 receives the B coefficients stored in the B coefficient training
4 memory 52 and determines the arithmetic mean of the B coefficients. The model
5 generator 36 then, in step 720, receives the C coefficients from the coefficient C training
6 memory 56 and determines the arithmetic mean of the C coefficients. Finally, in step
7 722, the model generator 36 receives the durations from the duration T training
8 memory 56 and determines the arithmetic mean of the durations. The model generator
9 36, in step 724, uses the arithmetic means and the data in the A coefficient training
10 memory 50, B coefficient training memory 52, C coefficient training memory 54, and
11 duration T training memory 56 to determine a co-variance matrix for random variables
12 A, B, C, and T. The model generator 36 gathers the arithmetic means together into a 4
13 dimensional vector and joins the vector to the co-variance matrix to form a model of
14 the selected language tone. The model generator 36 may alternately determine the
15 variance of each of the A coefficients, B coefficients, C coefficients, and duration T in
16 step 716, 718, 720, and 722, respectively and not determine the co-variance matrix in
17 step 724. The variance, however, is a sub-set of the co-variance, and determining the
18 variance alone yields a less accurate model than determining the co-variance matrix.
19 The determining variance alone, however, saves significant computation time. It is
20 preferred to determine the co-variance.

21 The model generator 36 signals, in step 726, the coefficient modeler 38 to receive
22 a model and then transfers the model to the coefficient modeler 38 for storage. The
23 model corresponds to the training data. For example, if the model generator 36 has
24 just completed processing the utterances of the tone 1 training data 22, then the model
25 generator 36 signals the coefficient modeler 38 to receive the tone 1 model 58.

26 In step 728, the model generator 36 determines if there is more tone training
27 data to be processed. If there is more tone training data to be processed, the model
28 generator 36 resets, in step 730, the coefficient A training memory 50, coefficient B
29 training memory 52, coefficient C training memory 56, and duration T training

1 memory 58. The model generator 36 then selects, in step 732, the next tone training
2 data to process and the method returns to step 700. If, in step 728, there are no more
3 sets of tone training data to be processed, the method ends.

4 Referring now to Figure 7, a flow chart of a preferred method for identifying the
5 tone of a syllable of Mandarin Chinese speech is shown. The method begins in step 800
6 where the tone recognition system 82 receives an input of a digitized syllable of
7 Mandarin Chinese speech. Upon receipt, the coefficient determinator 84 estimates the
8 pitch of the syllable, in step 802. In step 804, the coefficient determinator 84 fits a
9 second order polynomial to the estimated pitch. The coefficient determinator 84
10 preferably uses a least squares curve fitting method to fit a second order polynomial to
11 the pitch. The coefficient determinator 84 alternatively may use many other methods
12 to fit a polynomial to the tone. If the coefficient determinator 84 is being used to
13 recognize tones of a tonal language other than Mandarin Chinese, a higher order
14 polynomial may be used to model accurately the pitch. The coefficient determinator 84
15 then transmits the polynomial to the comparator 86. The comparator 86 determines
16 and temporarily stores the multivariate normal density for the tone 1 model 58 using a
17 Gaussian classifier as has been described above with respect to Figure 3. Next, in step
18 808, the comparator 86 determines the multivariate normal density for the tone 2
19 model 60 and stores the result. In step 810, the comparator 86 determines the
20 multivariate normal density for the tone 3 model 62 and stores the result. The
21 comparator 86 then, in step 812, determines the multivariate normal density for the
22 tone 4 model 64 and stores the result. Finally, in step 814, the comparator 86
23 determines the multivariate normal density for the tone 5 model 66 and stores the
24 result. In step 816, the comparator 86 selects the tone model 58 that received the
25 highest multivariate normal density as the model that best describes the input. The
26 comparator 86 selects the language tone of the selected tone model as the language
27 tone of the syllable. In step 816, the comparator 86 generates a signal to the syllable
28 recognition system 80 that indicates the language tone of the syllable.

1 While the present invention has been described with reference to certain
2 preferred embodiments, those skilled in the art will recognize that various
3 modifications may be provided. For example, the present invention may be used to
4 recognize any tonal language, such as Cantonese, and is not limited to Mandarin
5 Chinese. The present invention may also be used with handwriting recognition
6 systems and other data recognition systems. These and other variations upon and
7 modifications to the preferred embodiments are provided for by the present invention,
8 which is limited only by the following claims.



1 WHAT IS CLAIMED IS:

2 1. A system for recognizing a language tone of an input signal that includes
3 a syllable of a tonal language, the system comprising:

4 a coefficient determinator, having inputs and outputs, for determining the
5 coefficients of a polynomial that describes the pitch contour of the
6 syllable, an input coupled to receive a signal of the syllable;

7 a coefficient modeler, having inputs and outputs, for storing models of language
8 tones of the tonal language and for outputting, from an output, a signal
9 of the models; and

10 a comparator, having a first input coupled to an output of the coefficient
11 determinator for receiving the coefficients, a second input coupled to an
12 output of the coefficient modeler for receiving the models, and an output,
13 for comparing the coefficients determined by the coefficient determinator
14 to the models of the language tones and for generating a signal specifying
15 the language tone having a model that most closely matches the syllable
16 input to the coefficient determinator.

17
18 2. The system of claim 1, wherein the coefficient determinator also
19 determines the duration of the syllable.

20
21 3. The system of claim 1, wherein the coefficient modeler stores models of
22 the language tones of Mandarin Chinese.

23
24 4. The system of claim 1, wherein a model of a language tone comprises
25 expected coefficients of a polynomial that describes the fundamental frequency of
26 syllables having the tone of the tonal language.

27
28 5. The system of claim 4, wherein the model further comprises an expected
29 duration of syllables having the language tone.



1
2 6. The system of claim 5, wherein the model further comprises a co-variance
3 matrix for the expected coefficients and expected duration.

4
5 7. The system of claim 5, wherein the comparator uses a multivariate
6 normal density to compare the coefficients and duration, received from the coefficient
7 determinator, to the models received from the coefficient modeler.

8
9 8. The system of claim 4, wherein the expected coefficients of the model are
10 arithmetic means of coefficients of polynomials that describe the pitch contour of more
11 than one syllable having the language tone and the expected duration is an arithmetic
12 mean of durations of more than one syllable having the language tone.

13
14 9. The system of claim 4, wherein the polynomial is a second order
15 polynomial.

16
17 10. A system for training a coefficient modeler to model a language tone of a
18 tonal language, the system comprising:

19 a model generator, having a first input coupled to receive more than one
20 syllable each having a known language tone of the tonal language, a
21 second input, a first output and a second output, the model generator for
22 determining the coefficients of a polynomial that describes the pitch
23 contour of a syllable and for determining an arithmetic mean of the
24 coefficients of polynomials that describe more than one syllable having
25 the same language tone; and

26 a coefficient C training memory having an input coupled to the first output of
27 the model generation and an output coupled to the second input of the
28 model generator, for storing constant coefficients of polynomials that
29 describe the pitch contour of syllables having the same language tone.



11. The system of claim 10, further comprising:

a coefficient A training memory, having an input coupled to the first output of the model generator and an output coupled to the second input of the model generator, for storing second order coefficients of polynomials that describe the pitch contour of syllables having the same language tone; and

a coefficient B training memory, having an input coupled to the first output of the model generator and an output coupled to the second input of the model generator, for storing first order coefficients of polynomials that describe the pitch contour of syllables having the same language tone.

12. The system of claim 10, further comprising a duration T training memory, having an input coupled to the first output of the model generator and an output coupled to the second input of the model generator, for storing durations of syllables having the same language tone.

13. The system of claim 11, wherein the model generator also determines a co-variance matrix for the coefficients of the polynomials that describe more than one input syllable having the same language tone.

14. A method for recognizing a language tone of a syllable of a tonal language, the method comprising the steps of:
receiving an input of a syllable of the tonal language;
fitting a polynomial to the tone of the syllable;
comparing coefficients of the polynomial to expected coefficients of language tones of the tonal language;
determining the language tone whose expected coefficients most closely match the coefficients of the polynomial; and



1 generating a signal indicating the recognized language tone whose expected
2 coefficients most closely match the coefficients of the polynomial.

3
4 15. The method of claim 14, further comprising the steps of:
5 comparing a duration of the syllable to expected durations of syllables having
6 the language tones of the tonal language; and
7 determining the language tone whose expected duration most closely matches
8 the duration of the syllable.

9
10 16. The method of claim 15, wherein an expected coefficient is an arithmetic
11 mean of coefficients of the same order of more than one polynomial and the expected
12 duration is an arithmetic mean of durations of more than one syllable.

13
14 17. The method of claim 16 further comprising the step of comparing the
15 coefficients of the polynomial to a co-variance matrix for the expected coefficients of
16 the language tones.

17
18 18. The method of claim 17 where, in the steps of comparing coefficients of
19 the polynomial and duration of the syllable to expected coefficients and duration, a
20 multivariate normal density is used.

21
22 19. The method of claim 18, wherein the step of determining the language
23 tone includes a step of selecting the language tone having a model that yields the
24 highest multivariate normal density, when the model is compared to the syllable, as
25 the language tone of the syllable.

26
27 20. A method for training a coefficient modeler to model a language tone of a
28 tonal language, the method comprising the steps of:
29 receiving more than one syllable having the language tone;



1 fitting a polynomial to the tone of each syllable;
2 determining an arithmetic mean of the values of each coefficient of the
3 polynomials; and
4 generating signals to transmit the arithmetic means of the coefficients to the
5 coefficient modeler.

6
7 21. The method of claim 20 further comprising the steps of:
8 determining the duration of each syllable; and
9 determining an arithmetic mean of the durations.

10
11 22. The method of claim 20 further comprising the steps of:
12 determining a co-variance matrix for the values of the coefficients of the
13 polynomials and durations; and
14 generating a signal to transmit the co-variance matrix to the coefficient modeler.

15
16 23. A system for recognizing a language tone of a syllable of a tonal
17 language, the system comprising:
18 means for receiving an input of a syllable of the tonal language;
19 means for fitting a polynomial to the tone of the syllable;
20 means for comparing coefficients of the polynomial to expected coefficients of
21 language tones of the tonal language;
22 means for determining the language tone whose expected coefficients most
23 closely match the coefficients of the polynomial; and
24 means for generating a signal indicating the recognized language tone whose
25 expected coefficients most closely match the coefficients of the
26 polynomial.

27
28 24. The system of claim 23, further comprising:



1 means for comparing a duration of the syllable to expected durations of the
2 language tones of the tonal language; and
3 means for determining the language tone whose duration most closely matches
4 the duration of the syllable.

5

6 25. The system of claim 24 including means for comparing the coefficients of
7 the polynomial to a co-variance matrix for the expected coefficients of the language
8 tone.

9

10 26. The system of claim 25, wherein the determining means includes means
11 for selecting the language tone having a model that yields the highest multivariate
12 normal density, when the model is compared to the syllable, as the language tone of
13 the syllable.

14

15 27. A system for training a coefficient modeler to model language tones of a
16 tonal language, the system comprising the steps of:

17 means for receiving more than one utterance having the language tone;
18 means for fitting a polynomial to the tone of each utterance;
19 means for determining an arithmetic mean of the values of each coefficient of
20 the polynomials; and
21 means for generating signals to transmit the arithmetic means of the coefficients
22 to the coefficient modeler.

23

24 28. The system of claim 27, further comprising:
25 means for determining a co-variance matrix for the values of the coefficients of
26 the polynomials and durations; and
27 means for generating a signal to transmit the co-variance matrix to the
28 coefficient modeler.